BBN Report-8622

Raytheon
BBN Technologies

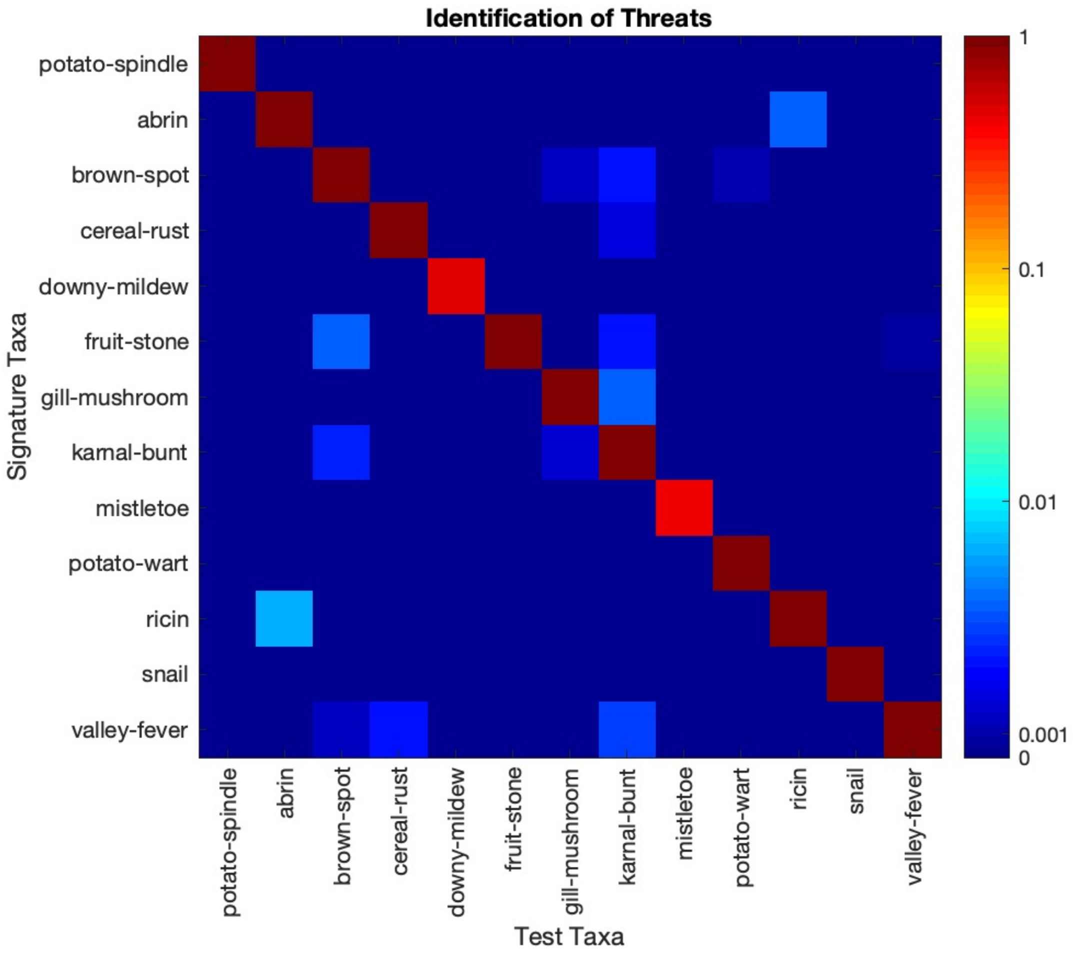
IDT
INTEGRATED DNA TECHNOLOGIES

Rapid Tests for Virus Genes that Suppress the Host Antiviral Defenses: FAST-NA Screening Technology Tasks

Development and Transition of FAST-NA Screening Technology

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

December 2021



The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Intelligence Advanced Research Projects Activity or the U.S. Government.

Raytheon BBN Technologies,
Integrated DNA Technologies

Rapid Tests for Virus Genes that Suppress the Host Antiviral Defenses: FAST-NA Screening Technology Tasks

BBN Report-8622
December 2021

Development and Transition of FAST-NA Screening Technology

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

The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Intelligence Advanced Research Projects Activity or the U.S. Government.

Sponsored by:
Intelligence Advanced Research Projects Activity
Contract No. W911NF-17-2-0092

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

Abstract: The Framework for Autogenerated Signature Technology for Nucleic Acids (FAST-NA) is a signature extraction method and accompanying software for detection of small nucleic acid sequences of concern in orders submitted for DNA synthesis. Building on prior success in developing FAST-NA for detection of viral pathogens, in this project we extended the application of FAST-NA to bacterial, eukaryotic, and viroid threats. We have also demonstrated that the FAST-NA method can be effective beyond its original intended application area, for both pathogen screening in next generation sequencing data and for region of interest detection in novel pathogens.

Following the conclusion of government-supported development, FAST-NA has been further developed to produce a commercially licensed software product that is now being deployed within the DNA synthesis industry.

Disclaimer: The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressly or implied, of the Intelligence Advanced Research Projects Activity or the U.S. Government.

Table of Contents

Figures and Tables

Figures

Tables

1 Task Objectives

The project "Development and Transition of FAST-NA Screening Technology" aimed to extend the Framework for Autogenerated Signature Technology for Nucleic Acids (FAST-NA), previously developed by BBN for detection of small nucleic acid sequences of concern in samples submitted for synthesis [1]. FAST-NA is based on the FAST signature extraction method that was originally developed by BBN for the detection of malware in network traffic [2, 3]. In the prior project developing FAST-NA, we extended FAST for use with nucleic acids, demonstrated efficacy and scalability for detection of viral pathogens, and demonstrated proof of principle for applicability to bacterial and eukaryotic threats [1].

The concept of operations for the application of FAST-NA to nucleic acid screening 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 threat level, justifying judgments using the metadata associated with matching signatures.

The goals for this project were to improve the efficacy, scope, and scalability of FAST-NA, particularly regarding its application to bacterial and eukaryotic threats and to investigate potential transition of the prototype system

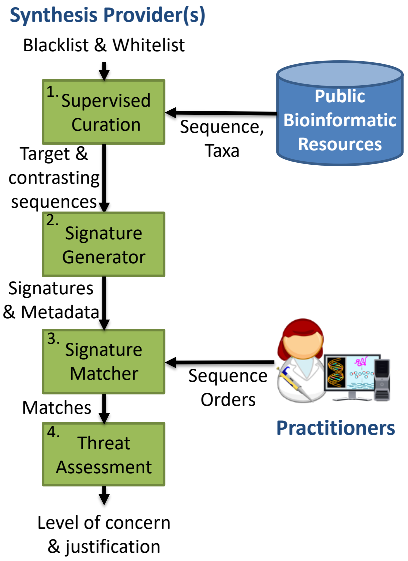


Figure 1. FAST-NA signature-based screening CONOPs.

for use with industrial screening systems. The tasks executed in pursuit of this objective were:

1. Improve scalability of FAST-NA software
2. Improve efficacy of FAST-NA viral screening
3. Extend FAST-NA screening to bacterial, eukaryotic, and viroid threats
4. Evaluate transition requirements for use in industrial screening systems.

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

2 Technical Problems

Through this investigation, we aimed to answer four core questions regarding the development and transition of FAST-NA pathogen screening:

- Can FAST-NA be scaled up to handle the much larger datasets for bacterial and eukaryotic threats?
- What adjustments are needed in order to enable effective signature generation for bacterial, eukaryotic, and viroid threats?
- What further developments are likely to be necessary in order to enable deployment of FAST-NA in industrial settings?
- Are there other application areas where FAST-NA is likely to be of use?

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

- Scaling improvements for FAST-NA (Section 4.1),
- Threat signature development (Section 4.2),
- Analysis of industrial needs vs. current FAST-NA implementation (Section 4.3), and
- Pilot efforts in other application areas (Section 4.4).

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 review the methods used in implementation of the FAST-NA screening software, curation of training and test data, and our experimental pipeline.

3.1 FAST-NA Screening Software

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 [4] used for pruning 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. 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

Both threat and contrast data 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 have found clustering to be effective generally at the Order or Family level. Contrast data is then collected from closely related non-threats, generally one to two taxonomic levels higher than the threat data, comprising all sequences from

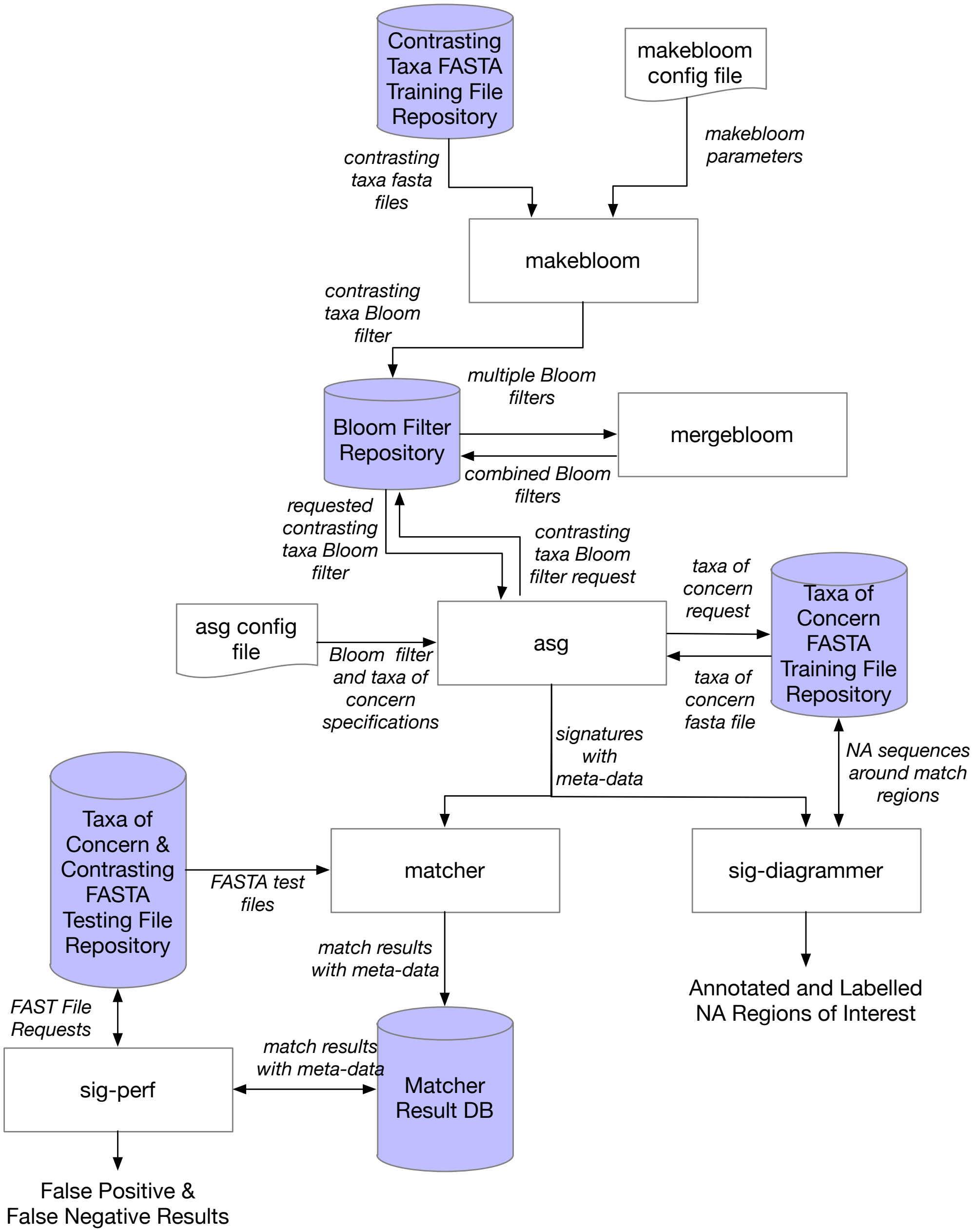

Contrasting Taxa FASTA Training File Repository
makebloom config file
makebloom parameters
contrasting taxa fasta files
makebloom
contrasting taxa Bloom filter
multiple Bloom filters
Bloom Filter Repository
mergebloom
combined Bloom filters
requested contrasting taxa Bloom filter
contrasting taxa Bloom filter request
taxa of concern request
Taxa of Concern FASTA Training File Repository
asg config file
asg
Bloom filter and taxa of concern specifications
taxa of concern fasta file
signatures with meta-data
NA sequences around match regions
Taxa of Concern & Contrasting FASTA Testing File Repository
FASTA test files
matcher
sig-diagrammer
match results with meta-data
Annotated and Labelled NA Regions of Interest
FAST File Requests
match results with meta-data
sig-perf
Matcher Result DB
False Positive & False Negative Results

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

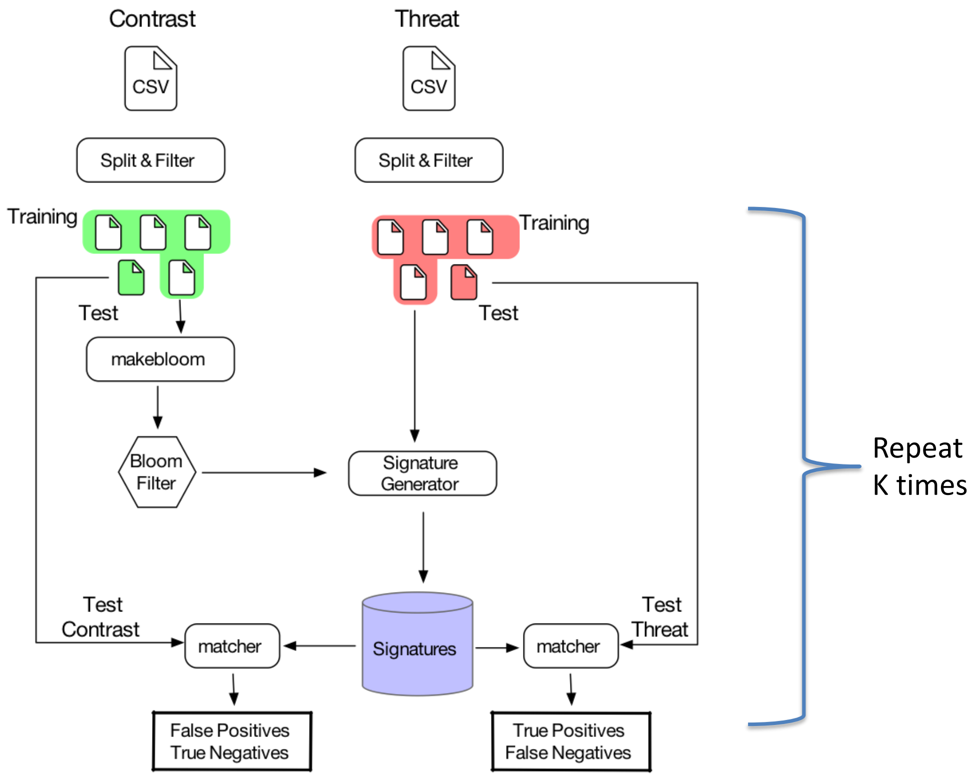


Figure 3. Architecture of k-fold cross-validation in experimental pipeline.

NCBI in the taxon that can be positively identified as not belonging to a threat taxa in the cluster.

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 3. 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 FAST-NA programs, 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 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) the training data is amino acid sequences, 2) nucleic acid 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.

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 1 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

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

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.

4 Technical Results

Here we report on the results from execution of the tasks defined above.

4.1 Scaling Improvements for FAST-NA

Scale is one of the key challenges for operation of FAST-NA against bacterial and eukaryotic threats. The genomes of most viruses are on the order of 10-20 kilobases, with the largest threats being in the smallpox cluster in the 100-300 kilobase range. Bacteria, on the other hand, are in the megabase range, and fungi in the tens of megabases, raising the expected scale by two orders of magnitude. Viroid and toxin threats, on the other hand, pose no scaling issues, since both are smaller than viruses. As a consequence, the total volume of genetic data for bacterial and eukaryotic threats is much larger and the number of signatures expected to be required for threat detection much larger as well.

In the curation of data for FAST-NA supported by this project, we worked with a total of approximately 3 gigabytes of viral training data, from which 25 million signatures were produced. When bacterial and eukaryotic threats were added, the total size swelled to approximately 500 gigabytes of training data and 1.4 billion signatures.

These posed significant issues that needed to be overcome for successful operation of FAST-NA, since running training and testing for the largest viral sequence collections consumed approximately 100 GB of RAM and approximately a day of time.

Via profiling and analysis, we identified a number of key bottlenecks and opportunities for scaling, which were addressed by refactoring the various affected FAST-NA software components. In particular, we made the following key improvements:

- **Shift to Incremental I/O:** During training and testing, FAST-NA initially accumulated all results in memory, then wrote them out when the training or testing epoch was completed. There was no need to do this, however, and shifting to writing results as soon as they were generated dramatically reduced the amount of memory required.
- **Compact signature format:** Signature files were becoming extremely large. Switching from their original JSON format to a compact CSV for-

mat reduced the size greatly. This also allowed incremental read of signature collections, which reduced both memory and loading time.

- **Accelerated signature generation:** The actual generation of signatures was found to have an inefficiency in its implementation that was causing a bottleneck. Optimization addressing this bottleneck provided an approximately 100x speedup in signature generation. Parallelization allowed yet further speedup.
- **Optimization of hash functions:** A significant speedup in hashing speed, affecting both Bloom filter and signature generation, was produced by a more precise calculation of the number of hashes required, rather than the conservative overestimate previously used.
- **Reduce sparsity of Bloom filter:** The rate of sequence conservation was much higher than initially predicted, meaning that our estimates setting Bloom filter size were far higher than the actual requirements, resulting in Bloom filters being far sparser than necessary. More accurate estimates reduced size requirements by an order of magnitude or more.
- **Optimization of matcher:** Specific optimizations of the matcher state machine allowed additional acceleration in its speed of operation.

Together, these optimizations allowed operation to scale to successfully work with bacterial and eukaryotic threats, with running training and testing for the largest taxa requiring approximately 800 GB of RAM and approximately one week of time.

4.2 Threat Signature Development

The majority of FAST-NA development effort supported by this project was put toward development of signatures for recognition of threats. This comprised both improvement of the prior work on viral threat recognition and the first systematic development of signatures for detection of threats from other kingdoms.

This process was executed with a joint workflow that simultaneously engages in signature tuning, validation of signature efficacy for screening, and training data curation, illustrated in Figure 4. This process is a slightly enhanced and extended version of the process developed in the preceding effort and documented in its final report [1]. The process is executed as follows:

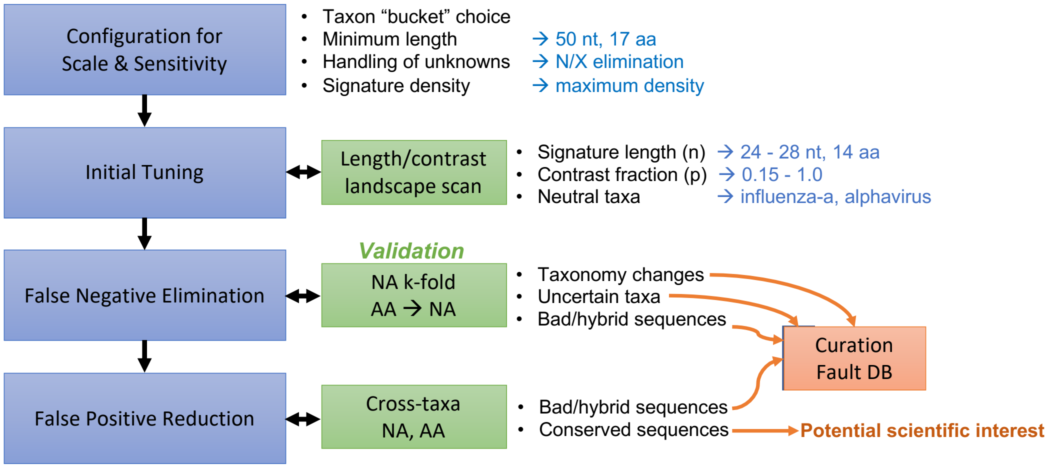

Configuration for Scale & Sensitivity
Taxon "bucket" choice
Minimum length → 50 nt, 17 aa
Handling of unknowns → N/X elimination
Signature density → maximum density
Initial Tuning
Length/contrast landscape scan
Signature length (n) → 24 - 28 nt, 14 aa
Contrast fraction (p) → 0.15 - 1.0
Neutral taxa → influenza-a, alphavirus
False Negative Elimination
Validation
NA k-fold
AA → NA
Taxonomy changes
Uncertain taxa
Bad/hybrid sequences
Curation Fault DB
False Positive Reduction
Cross-taxa NA, AA
Bad/hybrid sequences
Conserved sequences
Potential scientific interest

Figure 4. Joint process for tuning, screening validation, and curation used for FAST-NA signature development.

- FAST-NA is first configured for the threat space to be addressed. Threat taxa are first organized into clusters, generally at the Order or Family level of taxonomic rank. This is because it is a difficult and often ill-defined process to attempt to separate signatures from closely related threats, especially for organisms like bacteria where closely related strains frequently exchange genetic material. Other key parameters are determined at this stage as well, such as the minimum target length for detection (in our case, 50 nucleotides and the corresponding 17 amino acids), which unknowns to prohibit from signatures (in our case, all of them), and whether all k-mers or only a subset should be considered as potential signatures (in our case, all of them).

- Next, an initial tuning process evaluates the appropriate signature length to use for each taxa by evaluating how false positives and false negatives change with increasing signature length. Generally false positives drop as signature length increases, but when signatures are too long the false negatives rise as well. Accordingly length is set to the minimum level that has significant false positive rejection without introducing significant false negatives. Likewise, false negatives can also be tuned by adjustment of the fraction of contrast material used and identification of non-threat taxa that are to be considered neutral rather than contrasting because they are too close to the threat taxa to be reliably distinguished. For example, Influenza A is designated as a neutral taxa for Influenza A H1N1 because subtype classification considers only the Hemagglutinin (HA) and Neuraminidase (NA) proteins, but all the rest of the viral segments recombine with other Influenza A subtypes as well.

- Training and testing is then conducted incrementally both within a single threat cluster using k-fold validation and across clusters using cross-taxa

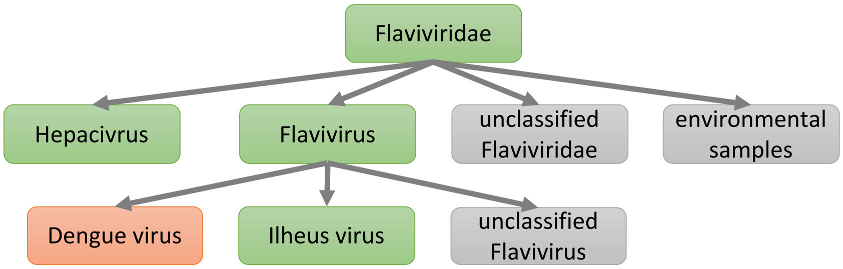


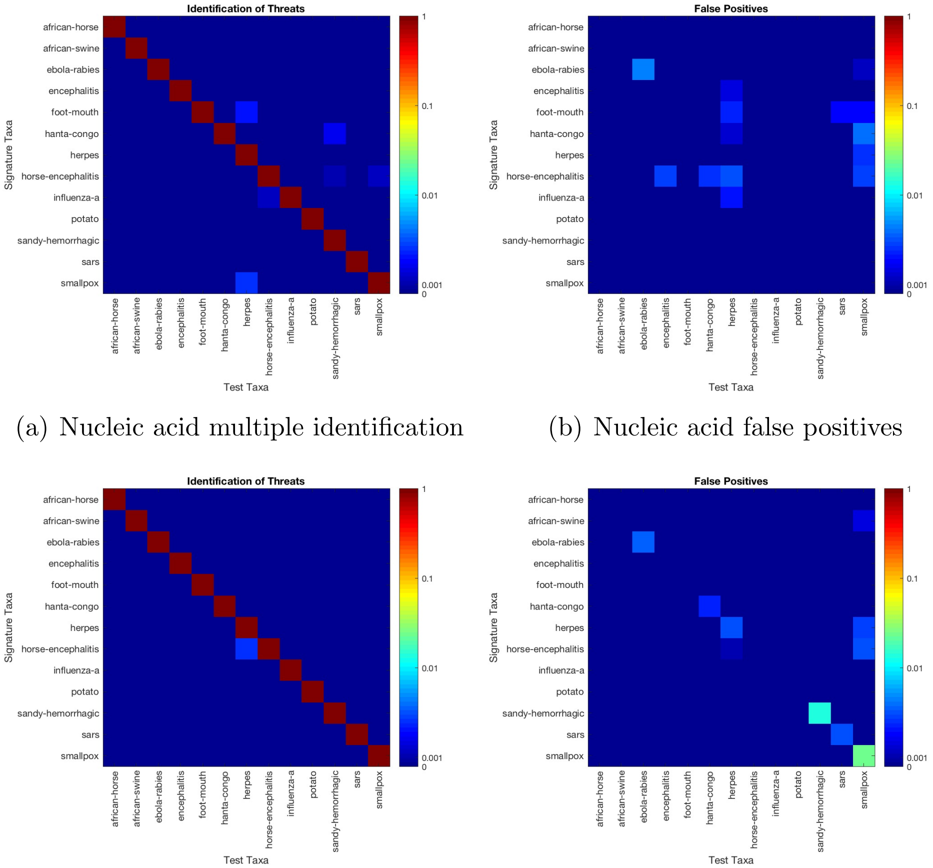
Figure 5. Example of pseudotaxa (grey) that need to be removed from FAST-NA training data, since their contents cannot be clearly classified as either threat (red) or contrast (green).

testing, reducing false positives and false negatives until the target performance is reached. Bioinformatic analysis of the incremental results from each cycle identifies issues in curation that then lead to adjustments in the curation of the training taxa, including identification of taxonomy issues and identification of incorrectly categorized sequences and synthetic hybrid sequences. Important systematic adjustments included identification of pseudotaxa that needed to be excluded (Figure 5), identification of certain highly conserved materials such as cellular rRNA and tRNA for automatic exclusion from signatures, and removal of common primer sequences.

This process is intensive both computationally and in terms of bioinformatic analysis. During the course of development supported under this project, 142 training and testing iterations batches were conducted, lasting between one day and two weeks per iteration, with iterations on different taxa often running in parallel.

Signatures for detection of viral threats, already near target performance levels by the end of the prior effort, were further refined. Figure 6 shows the signature performance achieved for viral threats by the end of supported development on this project. Nucleic acid signatures improved from an average per taxa rate of 0.67% multiple identification of threats to 0.11% multiple identification and from 1.08% false positives to 0.32% false positives. Protein signatures also improved, from an average per taxa rate of 0.20% multiple identification of threats to 0.033% multiple identification and from 0.91% false positives to 0.48% false positives. This put the viral rates well under the target average rate of 1% multiple identification and 1% false positives. False negatives remained at zero, as desired.

Eukaryotic threats are divided into two general classes: fungal threats, most of which are agricultural pathogens (e.g., cereal rust, brown spot) and organ-



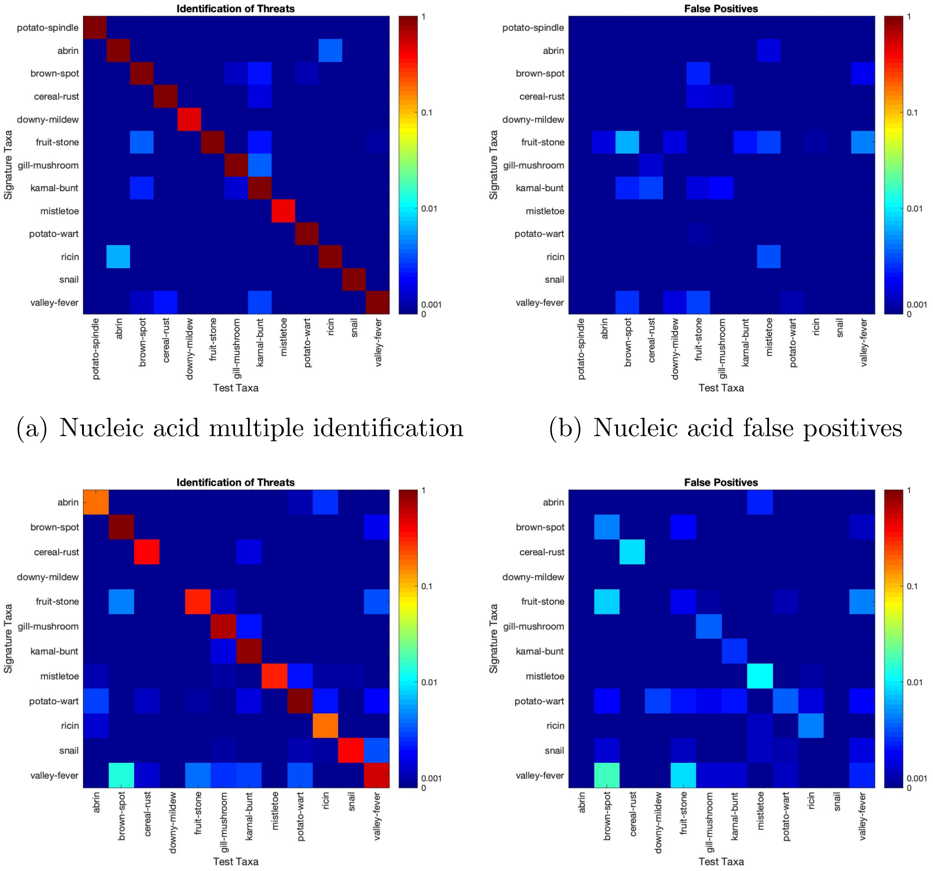
(a) Nucleic acid multiple identification

(b) Nucleic acid false positives

(c) Protein multiple identification

(d) Protein false positives

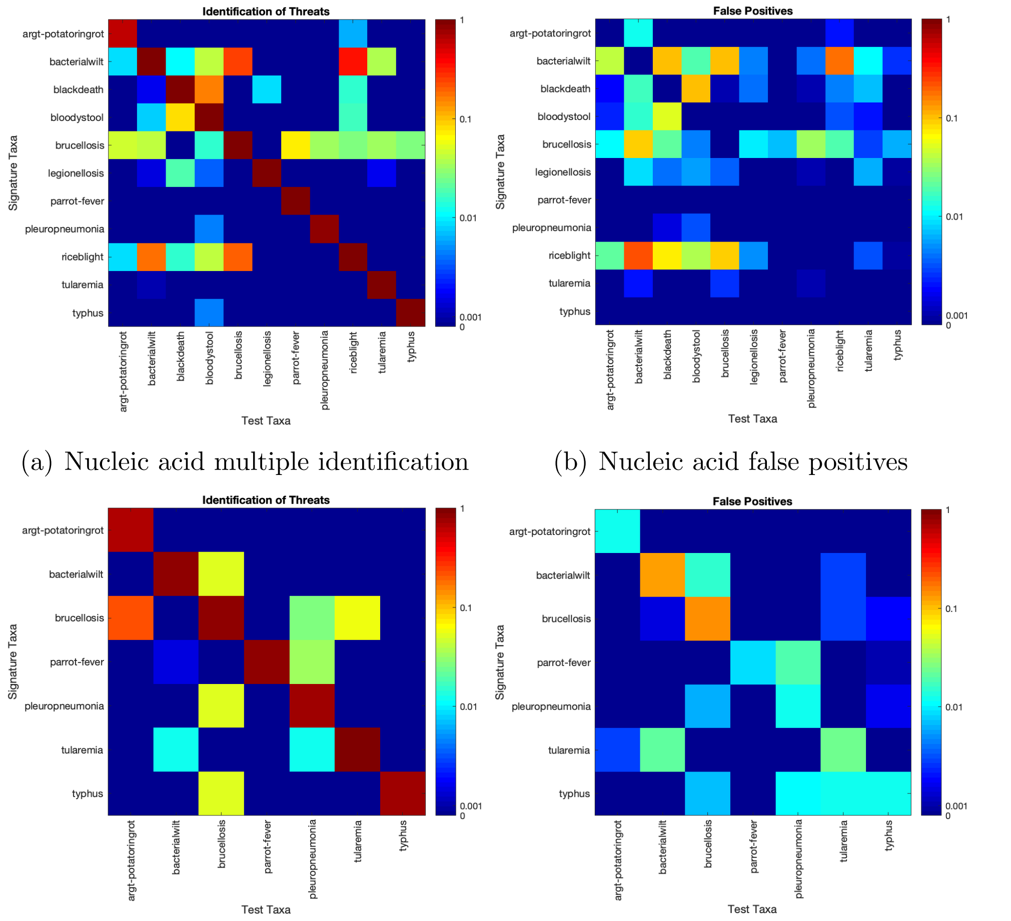
Figure 6. FAST-NA signature performance in detection of viral threats.



(a) Nucleic acid multiple identification (b) Nucleic acid false positives

(c) Protein multiple identification (d) Protein false positives

Figure 7. FAST-NA signature performance in detection of eukaryotic and viroid threats.



(a) Nucleic acid multiple identification

(b) Nucleic acid false positives

(c) Protein multiple identification

(d) Protein false positives

Figure 8. FAST-NA signature performance in detection of bacterial threats.

isms that produce controlled toxins (e.g., snail conotoxins, ricin). The only controlled viroid threat taxa, the potato spindle tuber viroid, is also an agricultural pathogen, so its signature development was bundled together with the eukaryotic threats for evaluation of potential cross-taxa false positives. Viroids are such a distinct category, however, that no significant false positives were ever detected either for viroid signatures detecting other taxa or for eukaryotic threat signatures detecting viroids. Figure 7 shows the final results achieved with eukaryotes and viroids during this development supported by this project. Nucleic acid signatures achieved an average per taxa rate of 0.36% multiple identification of threats and 0.52% false positives. Amino acid signatures achieved an average per taxa rate of 0.69% multiple identification of threats and 1.15% false positives. False negatives were low as well but not yet entirely eliminated, with a rate of 0.02%. In sum, this put the eukaryote and viroid rates close to the target rates, but needing additional tuning after project-funded development concluded.

Bacterial threats proved the most difficult to tune, both due to the tendency of bacteria to exchange genetic material and due to amount of poorly categorized and unannotated shotgun sequencing data in many taxa. Certain taxa pairs were also marked as expected confusions due to the level of known genetic exchange between threat clusters, such as the *Shigella* (cluster "bloody-stool") and pathogenic *E. coli* (cluster "shiga-ehec"). As a result of these challenges, by the conclusion of the development supported by the project, tuning had not even begun in earnest for the largest six threat taxa with nucleic acid signatures and for more than half of the taxa for protein signatures. For those taxa where tuning was conducted, progress was steady, but not completed by the time that project-funded development completed.

Figure 8 shows the final results achieved with bacteria during the development supported by this project. For the eleven taxa in progress with nucleic acid signatures, the average per taxa rates were 11.3% multiple identification of threats and 10.9% false positives. For the seven taxa in progress with amino acid signatures, the average per taxa rates were 7.6% multiple identification of threats and 31.9% false positives. Although these rates were high, note that they are mostly driven by a small number of particularly problematic taxonomic pairings, and thus readily subject to continued tuning processes. With regards to false negatives, for the seven taxa that had both nucleic acid and protein signatures developed, false negatives were low but non-zero, with a rate of 0.01%. In sum, these results indicated that the bacterial threats were on track for achieving the target rates with this tuning process, but needed significant additional tuning after project-funded development concluded.

4.3 Analysis of Industrial Needs vs. Current FAST-NA Implementation

The final state of FAST-NA development supported with funding from this project was not sufficient for deployment. Additional transition development was needed in order to allow FAST-NA to meet the needs for deployment as a DNA synthesis screening mechanism. We assessed the key required transition work to be:

- Complete development of bacterial signatures, including nucleic acid signatures for large taxa and protein signatures for all taxa.
- Development of signatures for controlled toxins.

- Refinement of signatures via cross-kingdom testing and testing against common model organisms and artificial sequences.
- Reduction of matcher memory requirements to allow simultaneous matching against all taxa with less than 16GB of RAM.
- Refactoring of matching workflow from batch-focused start/stop workflow to a continuous workflow with memory-resident operations.
- Wrapping of matching capabilities with an appropriate API for usage within a typical DNA screening CONOPS.
- Documentation, containerization, and hardening of software for deployment.

The relevant government representatives declined to fund this transition and recommended instead that FAST-NA be further developed without government support as a commercial software product. Accordingly, we have followed this recommendation and invested to address all of the listed requirements, thereby developing a FAST-NA product that is available to qualified organizations for use as commercially licensed software.

4.4 Pilot efforts in other application areas

In addition to the primary development of FAST-NA as a tool for DNA synthesis screening, we also conducted pilot experiments in the application of FAST-NA to other biosecurity areas.

4.4.1 Pathogen Detection in Sequencer Data

We had hypothesized that FAST-NA could be applied to analyze not just DNA synthesis orders but DNA sequencer output. In both cases, the core function of applying signatures to scan nucleic acid sequence information is the same. There are significant differences in the CONOPS between analysis of sequencer output, however, that require adjustment of how FAST-NA is applied and interpreted:

- **Sequencing offers many opportunities for detection:** With DNA synthesis screening, a decision needs to be made based on a single sequence, and that sequence may be very short. With DNA sequencing, however, if a pathogen is present then many different portions of its sequence should appear in many different fragments of the sequencer ouput.

With sequencing, there are thus many opportunities for detection and these should include many different highly diagnostic fragments of the genome.

- **Sequencing errors can cause false detections:** DNA synthesis orders are precise in the sequence that is being requested. Sequencing is an error prone operation, however, including many different types of error that can lead to false detection events.

This hypothesis was tested in collaboration with a group at Army CBC who were working to develop fieldable pathogen detection capabilities based on next generation sequencing (NGS) methods. In early experimentation in this collaboration, we determined that amino acid signatures produced a generally more reliable detection signal than nucleic acid signatures. We also determined that false positives could be significantly reduced by filtering bases via quality level indicated in the FASTQ file, changing all bases below some threshold to n to indicate an unknown nucleic acid. In experiments, the quality threshold was set either to 14 or to 7.

We then conducted three sets of blinded experiments, the results of which are reported in Figure 9. Of the twenty samples tested, correct determinations were made for nineteen. For all six no-threat samples, the determination was in every case no threat. For the fourteen samples with one or more threats, the correct threat category was detected in thirteen cases. The one sample where a threat was not detected was with Abrus precatorius, the plant that produces the abrin toxin. In this case, the threat was still detected via nucleic acid signatures, but those results were not used due to the prior finding of amino acid signatures being generally more reliable.

Another significant finding in these experiments was that FAST-NA signatures can indeed be effective for detecting novel pathogens. The FAST-NA signatures used for these tests were based on a July, 2019 snapshot of NCBI data, prior to the emergence of the SARS-CoV-2 virus. Despite the lack of training against SARS-CoV-2, however, the two test samples of SARS-CoV-2 were successfully detected as being part of the SARS threat cluster, based on their similarity to a closely related bat coronavirus.

We thus conclude that FAST-NA can be effectively adapted to pathogen detection in sequencer data. Adaptation to this CONOPS would require some additional efforts in refining and automating the interpretation of matches to ensure reliable detection and an appropriate tradeoff between false negatives and false positives for this application.

Sample description	Threat	Correct?
Bacillus thuringiensis kurstaki	No	Yes
Yersinia pestis	Yes	Yes
Bacillus thuringiensis kurstaki and VEEV	Yes	Yes
Bacillus anthracis	Yes	Yes
Bacillus anthracis in simulated natural background	Yes	Yes
Real world environmental aerosol sample (no biothreats)	No	Yes
Yersinia pestis in simulated natural background	Yes	Yes
E. coli	No	Yes
Clinical SARS-CoV-2 positive human sample	Yes	Yes
MSA2002 mock community (mixture of 20 bacterial species)	Yes	Yes
SARS-CoV-2 RNA sample	Yes	Yes
Canine coronavirus	No	Yes
Botrytis cinerea and Pyricularia oryzae	Yes	Yes
Thermothielavioides terrestris and Abrus precatorius	Yes	No
Thermothielavioides terrestris and Bipolaris oryzae	Yes	Yes
Saccharomyces cerevisiae	No	Yes
Coccidioides immitis	Yes	Yes
Coccidioides posadasii and Coccidioides immitis	Yes	Yes
Zymo Mock Community	Yes	Yes
Aerosol environmental background	No	Yes

Figure 9. Results of application of FAST-NA to blinded sequencing data.

4.4.2 Analysis of SARS-CoV-2 as a Novel Pathogen

We had hypothesized that FAST-NA could be used for analysis of novel pathogens in two different configurations. First, FAST-NA can be used to evaluate similarities between novel pathogens and existing pathogens by applying existing signatures to determine which portions match with a novel pathogen. Complementarily, FAST-NA can be used to evaluate significant differences between novel pathogens and existing pathogens by using the existing pathogens as contrast data and building a signature from a novel pathogen.

The emergence of the novel SARS-CoV-2 virus provided a test case for both of these scenarios. As noted above in the section on pathogen detection in sequencer data, the overlap between SARS-CoV-2 and the prior SARS-related threat cluster allowed detection of SARS-CoV-2 samples without any training against the new pathogen due to its similarities with the other coronaviruses closely related to the original SARS virus.

We also applied FAST-NA in the second configuration to evaluate the new virus for potential regions of significance. Immediately after the publication of the SARS-CoV-2 genome in January of 2020, we ran an analysis of the genome using FAST-NA. Specifically, we applied FAST-NA to identify all of the unique 10-mer sequences in all of the amino acid sequences for SARS-CoV-2 then available from NCBI: 63 amino acid sequences available in NCBI, comprising a total of 49379 amino acids. For contrasting sequences, we used a July, 2019 snapshot of all protein sequences in family *Coronaviridae* available from NCBI, a total of 50574 sequences comprising a total of approximately 40 million residues. The resulting collection of unique 10-mer amino acid sequences were then concatenated where overlapping within the same parent sequence and trimmed to remove non-unique flanking portions.

All told, this process identified 61 multi-amino-acid regions as significant unique sequences for SARS-CoV-2, comprising a total of 1669 amino acids (3.4% unique and non-repeated), spread across 8 non-duplicative sequences. In addition, we also identified 45 single amino-acid polymorphisms. Figure 10 summarizes the distribution of unique sequence regions across these 8 open reading frame (ORF) sequences. Two of these have notably high amounts of unique content: the large 1ab polyprotein QHO60603.1 has much unique material, though the fraction is not large, while the surface glycoprotein QHO60594.1 has both a large amount and large fraction of unique material. Further examination showed that the unique material in these

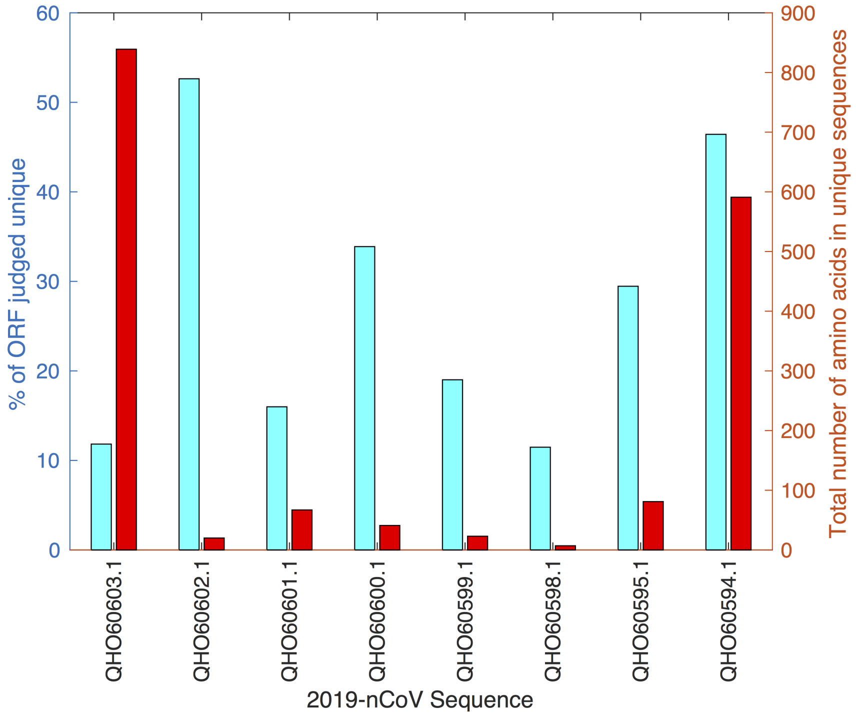


Figure 10. Summary statistics of distinguishing amino acid sequences identified for SARS-CoV-2, showing the fraction of each ORF judged to be part of unique sequences and the total number of amino acids in unique sequences in the ORF. The large 1ab polyprotein QHO60603.1 has much unique material, though the fraction is not large, while the surface glycoprotein QHO60594.1 has both a large amount and large fraction of unique material.

two ORFs is strongly clustered. Taking a cluster as any sequence of at least three unique regions with no more than 50 amino acids separating them, we found that QHO60603.1 had two clusters, one spanning from residues 916–1294, the other from 6417–6715, containing 47% of the unique material in the sequence. The QHO60594.1 sequence, meanwhile, has a single large cluster, spanning from residues 9 to 883 and comprising all of the unique material in the sequence.

In summary, analysis of the amino acid sequences of SARS-CoV-2 identifies three large highly unique regions of the genome that distinguish it from all other *Coronaviridae*, plus several dozen other smaller regions of uniqueness. These results were published as a preprint on February 2nd, 2020 [5]. A few weeks later the first structural analysis of the SARS-CoV-2 spike protein was published [6], in which the critical regions of the spike protein were identified as one of the three regions of uniqueness that we identified. Although the authors of that paper were unaware of our results and working independently in parallel, their analysis was an important confirmation that FAST-NA may be able to play a useful role in analysis of novel pathogens.

5 Summary and Discussion

5.1 Progress Against Waypoints

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

- Scaling improvements for FAST-NA: **complete**
- Threat signature development: **complete**
- Analysis of industrial needs vs. current FAST-NA implementation: **complete**
- Pilot efforts in other application areas: **complete**

5.2 Important Findings and Conclusions

Our findings in this report are as follows:

- FAST-NA can be scaled for use with bacterial and eukaryotic threat taxa with large volumes of data.
- FAST-NA can reduce false positives in screening for viral, bacterial, eukaryotic, and viroid threats, without introducing false negatives.
- Effective industrial deployment of FAST-NA requires dramatic reduction in memory requirements, plus wrapping and hardening of software for deployment within an industrial screening context.
- FAST-NA has potential for application in a number of other areas of biosecurity concern.

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 all controlled biological threats. These capabilities are being deployed as a commercial software offering, in order to support the maintenance and development of the FAST-NA software and signature collection.

Deployment as commercial software, however, does limit the breadth of deployment of this screening technology, due to the requirement that its support be funded by user licenses. Deployment, at least in the near term, must thus necessarily focus on only the relatively small number of large-scale organizations who are already de facto required to invest significant resources in pathogen sequence screening. **If the government supported maintenance and development of FAST-NA as a bioinformatic resource for the public good, then FAST-NA could potentially be rapidly deployed to many more users.** This would allow biosecurity screening to be deployed cheaply and rapidly by a much larger group of organizations that are either small or not traditionally in the biosecurity space, including small DNA synthesis companies, biotechnology companies other than DNA synthesis companies, government laboratories, universities, plasmid repositories, etc.

We have also demonstrated FAST-NA is likely to be effective for other applications, such as screening for pathogens in next generation sequencing (NGS) data and analysis of novel pathogens. Other potential areas of value for biosecurity applications include oligo screening, gain-of-function detection, combinatorial library screening, and analysis of network traffic for biothreat planning. We recommend further investigation of such possibilities.

5.5 Commercial/Proprietary/Third-Party Material in Deliverables

None.

References

[1] Applicability of FAST-NA to nucleic acid screening. Final Report on IARPA Contract No. 2018-17110300002, February 2019.

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

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

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

[5] Jacob Beal, Thomas Mitchell, Daniel Wyschogrod, Jeff Manthey, and Adam Clore. Highly distinguished amino acid sequences of 2019-nCoV (Wuhan coronavirus). *bioRxiv*, 2020.

[6] Daniel Wrapp, Nianshuang Wang, Kizzmekia S Corbett, Jory A Goldsmith, Ching-Lin Hsieh, Olubukola Abiona, Barney S Graham, and Jason S McLellan. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367(6483):1260–1263, 2020.