Clarifying the Problem of DNA Screening

Max Langenkamp[1]

Abstract

Screening DNA synthesis orders is one crucial step in preventing biological catastrophe. The current lack of consensus over the approach to screening DNA is an obstacle to adoption and standardization. This article aims to clarify the debate around DNA screening by first framing the problem, proposing a relevant threat model, and then discussing the desired properties of a screening system. The intended audience includes policymakers and technologists interested in the problem of DNA screening.

I. Introduction

The dizzying rate of progress in synthetic biology brings both promise and peril. In addition to advances like prime editing[2] and self replicating synthetic cells, the costs of synthesizing and sequencing DNA have decreased significantly. This is very promising for the discovery of new drugs, but has also increased the risk of biological catastrophes. Estimates for the frequency of attempted bioterror attacks range from between 0.35 to 3.5 per year[3]. This is low but concerning given the increased capabilities for harm that new technologies such as benchtop DNA synthesizers may pose. It is getting easier for an adversary to find a dangerous sequence, synthesize it, insert it into the appropriate vectors and spread it[4].

---

[1] Max Langenkamp is an independent researcher investigating biosecurity. He received his BSc and MEng from MIT in 2022, where he did research on the SecureDNA system. Questions can be sent to maxlangenkamp@me.com

[2] Anzalone, Andrew V., et al. "Search-and-replace genome editing without double-strand breaks or donor DNA." *Nature* 576.7785 (2019): 149-157.

[3] See Millett, Piers, and Andrew Snyder-Beattie. "Existential risk and cost-effective biosecurity." *Health security* 15.4 (2017): 373-383. supplemental material at https://www.liebertpub.com/doi/suppl/10.1089/hs.2017.0028

[4] It is important not to trivialize the significant barriers posed by inserting genes into vectors. For helpful discussion, see pages 24-26 of NTI's report on benchtop synthesizers: Carter, Sarah et al. "Benchtop DNA Synthesis Devices: Capabilities, Biosecurity Implications, and Governance." *Nuclear Threat Initiative* (2023)

If an individual wants to synthesize DNA right now, all they need to do is log onto the portal of a DNA synthesis company like Twist or IDT, and enter the DNA sequence that they want to receive. Orders can be up to 1.8kb for double stranded DNA[5], cost around 7-9 cents per base pair, and will arrive in 6 to 9 business days via mail.

Current DNA synthesis oversight is done on a voluntary basis using guidelines that are vague and do not protect against adversaries with basic knowledge of biology and computer security. The guidelines in question are the Human and Health Services 2010 Screening Framework Guidance, and the International Gene Synthesis Consortium (IGSC) Harmonized Screening Protocol. **There are no laws governing DNA screening in the United States, nor any other countries[6]**.

Eight approaches for DNA screening are currently commercially or publicly available: Aclid, BLISS, Raytheon's FastNA, SecureDNA, SeqScreen, Batelle's ThreatSeq, and NTI's Common Mechanism[7]. Each approach, however, makes distinct assumptions about the problem, and so makes different design tradeoffs. Hitherto there has been little to no explicit discussion of the assumptions and tradeoffs of various approaches to screening. The aim of this paper is to contribute to such a discussion by framing the problem of DNA screening and discussing the properties desired.

By the end of this paper, the reader should:

1) Understand the functional approach towards DNA screening
2) Articulate a simple threat model relevant for DNA screening
3) Develop a sense of the key tradeoffs involved in the design of DNA screening systems

II. What sequences to screen?

Imagine, for a moment, that you are an employee at a DNA synthesis provider. After being given a DNA sequence *x,* you are asked whether you think the company should manufacture the DNA sequence. What would you do?

---

[5] For context, the genome of most viruses is at least 5 kbp. Smallpox is 190 kbp.

[6] Note that there are certain minimal requirements in the United States around customers of gene synthesis companies; gene synthesis providers are legally obligated not to do business with customers that are on a prohibited person or entity list. These lists include the Department of Treasury Office of Foreign Assets Control (OFAC) list of Specially Designated Nationals and Blocked Persons (SDN List), the Department of State list of individuals engaged in proliferation activities, and the Department of Commerce Denied Persons List (DPL). See 2010 HHS Screening Framework Guidance for more information

[7] For more information, see Gretton, Dana. “Platforms for Biological Control”. [Master’s thesis, Massachusetts Institute of Technology] (2022)

The Federal Select Agent Program (FSAP) list and the Commerce Control List (CCL) provide a natural language list that includes items like “Chikungunya virus” and “Avian influenza”. On its own, however, this cannot tell you whether *x* should be screened.

NIH has a database that is helpful here: a short search will reveal that “Chikungunya virus” has 11,604 base pairs, and can be found in the publicly listed Genbank dataset. But how do we compare *x* to Chikungunya?

We could exactly compare *x* to Chikungunya and everything else on the FSAP and CCL. But if even a single base pair differs, this procedure would yield a false negative. *x* might match 11,603 out of the 11,604 base pairs of Chikungunya virus, but the final inconsequential change from an ‘A’ to a ‘T’. We would then, for all intents and purposes, be providing someone with viral DNA. “For all intents and purposes” here is clearly not captured by exactly matching the DNA. Can we do better by making a fuzzy matching algorithm?

Homology screening captures small sequence alterations but can be easily evaded

We could choose random small windows (say, of 100 bp) from *x* and Chikungunya virus, and do many attempts at exact matching. Similarly, we could use a popular sequence alignment algorithm like BLAST to compare the two. BLAST uses a heuristic to find short matches between two sequences and produces a similarity score based on alignment length, number of matches, and other factors. This class of approaches using sequence similarity at the base pair level is referred to as ‘homology-based screening’. It is also what is currently proposed by the HHS 2010 Screening Framework Guidance, and the IGSC Harmonized Screening Protocol.

However, as a examination of a codon table reveals, there are many sequences that can map to an amino acid. If *x* were simply a re-coded version of Chikungunya virus, where for instance some Phenylalanines had been swapped from UUU to UUC, a homology-based approach would not catch the underlying sequence.

	U		C		A		G		
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys	U
	UUC		UCC		UAC		UGC		C
	UUA	Leu	UCA		UAA	Stop	UGA	Stop	A
	UUG		UCG		UAG	Stop	UGG	Trp	G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	Gin	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	Lys	AGA	Arg	A
	AUG	Met	ACG		AAG		AGG		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	Glu	GGA		A
	GUG		GCG		GAG		GGG		G

Source

It gets more difficult. In addition to recoding, we might worry that parts of *x* are altered without undermining its pathogenicity (for example, Chikungunya with non coding DNA removed).

Functional Screening

In addition to seeing DNA sequences as long strands of nucleotides, as the homology view roughly does, **we can also begin to map groupings of nucleotides to specific biological functions**. For instance, within the genome of SARS-Cov-2, Nsp14 refers to a subsequence coding for an enzyme that prevents host cell protein production. It is around 1,500 base pairs long, out of the 29,000 base pairs coding for SARS-Cov-2[8].

In a recent paper, Godbold et al. proposed using such a functional view for biological agents. They annotate the function of 2,750 sequences, which they term Functions of Sequences of Concern (FunSoCs)[9], and propose to use such a list for screening DNA orders. They provide many examples of functions that might be considered harmful to human hosts: enzymes that degrade tissue, adhere to host cells, suppress host immune signaling, and so on.

[8] According to Tahir, Mohammed. "Coronavirus genomic nsp14-ExoN, structure, role, mechanism, and potential application as a drug target." *Journal of Medical Virology* 93.7 (2021): 4258-4264., Nsp14 is 60 kDa, which corresponds to ~500 amino acids, or 1,500 bp.

[9] Godbold, Gene D., et al. "Categorizing sequences of concern by function to better assess mechanisms of microbial pathogenesis." *Infection and Immunity* 90.5 (2022): e00334-21.

ProteinNames	Organism	UniProt_Accession	MHP_Cytoskeleton_Dynamics	PathGO_Term	Part_Of_Protein_Complex?
Aerolysin	Aeromonas hydrophila	P09167	Aerolysin is a substrate of the T2SS and is the primary effector of barrier impairment induced by Aeromonas hydrophila infection. Aerolysin treatment decreases transepithelial resistance with tight junction deformation and cytoskeletal changes [PMID21917902].	PATHGO:0000072 (mediates binding to cell surface glycoprotein in another ...	

Example of annotated FunSOC[10].

Under the functional view of DNA screening, the screening problem is straightforward:

For a given order x, does x contain a subsequence that has been identified as presenting a dangerous function?[11]

Rather than using the entire sequence in our example order *x*, we can instead search to see if *x* contains a subsequence like Nsp14 that encodes a function that enables harm in humans[12]. Loosely, we could think of this as similar to the federal approach to restricting the lower receiver in assault rifles, rather than banning specific rifle models.

Note that this is a big step towards addressing what counts as a dangerous sequence: a dangerous sequence is one that contains a subsequence that biologists have labeled as coding for a protein that causes harm to its human hosts.

---

[10] Taken from supplemental file 2 from Godbold, Gene D., et al. "Categorizing sequences of concern by function to better assess mechanisms of microbial pathogenesis." *Infection and Immunity* 90.5 (2022): e00334-21.

[11] A more full definition: “for a given DNA query x, is f(x) in dangerous database D, where D includes all currently known functional subsequences of concern, and f is a one-to-many function whose domain is all sequences greater than some subsequence length n and whose range are the possible subsequences contained, with added transformations for obscuration resistance.”

[12] In practice, we can get much smaller than 1500 bp. Screening windows for SecureDNA, for instance, are on the order of 50bp

If properly characterized, the functional subsequence should precisely capture those sequences we are concerned about, and essentially none of those that we do not. At least two of the existing approaches to screening — SecureDNA and SeqScreen — currently employ screening based on a functional view. However, perhaps the most broadly used screening standard, the International Gene Synthesis Consortium's Harmonized Screening Protocol, proposes using a homology based-approach.

Though the functional approach is more principally sound than a homology-based approach, there are certain challenges with the functional approach. One difficulty, however, lies in characterizing the space of functional equivalents; examples such as Godbold et al's database are allow us to screen exact matches to known functional subsequences, but what about alterations to the functional subsequences? As biological design tools improve[13], this is a concern that will be important to address.

III. Threat model

I'll be focusing on the threat model I consider as most plausibly leading to biological catastrophe.

This involves ordering dangerous DNA from an existing provider, who will remotely synthesize the DNA. The adversaries may then insert the DNA into a vector and spread it in a public place. Later on, I plan to extend this threat model to newly emerging hardware.

System Overview: This refers to a synthetic biology service that provides custom DNA sequences to customers. The service conducts biosecurity screening on requested DNA sequences to prevent the synthesis of dangerous biological agents. The order process is conducted through an online portal, and the company maintains a database of known hazardous sequences.

Adversaries[14]:
While accidental misuse of DNA synthesis is important to minimize, the scenarios significantly more likely to cause widespread catastrophe involve deliberate attempts to cause a pandemic[15]. In this scenario, the adversary's primary aim is to synthesize an agent that is on a restricted list such as the CCL or FSAL.

Access: The adversaries have access to an ordering portal for DNA.

---

[13] Sandbrink, Jonas B. "Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools." *arXiv preprint arXiv:2306.13952* (2023).
[14] For more detailed discussion of potential adversaries, see Esvelt, Kevin. "Delay, detect, defend" Geneva Center for Security Policy, 2022
[15] The DNA for a virus on its own is not harmful to humans and requires several steps in order to pose a threat. In the case of DNA synthesis, accidental harms are highly unlikely.

Adversary Capabilities

- Ligation: Adversary can order from different providers and do simple ligation of sequences
- Sequence obfuscation: an adversary can modify the screened DNA while preserving function

Note that there may be cybersecurity vulnerabilities within, for instance, the DNA ordering portal, or the manufacturer network. Such attack surfaces are out of the scope of this article. For now, I will consider only the actual screening of the DNA.

IV. Properties of a Screening System

This section provides an overview of the various properties that are desirable in a DNA screening system.

Property	**Description**	**Reasoning**
Order privacy	Individual orders kept private from adversaries that may have access to incoming network traffic	First, companies with sensitive IP will be far more willing to use screening systems that guarantee private orders. Second, adversaries may be able to gain information about database contents if all orders can be monitored
High throughput	System can be easily scaled to handle over 10 million queries per second.	By 2029, queries are expected to exceed 10 million per second[16]. Any centralized system needs to take very high throughput needs into consideration.
Function-based	Orders screened on the basis of subsequences recognized to be critical to the function of the dangerous organism as a whole	A function-based view of screening more accurately captures sequences that may be considered dangerous. This approach implies the ability to recognize emerging pathogens, and to combat adversaries attempting to ligate sequences.

[16] From SecureDNA's *Cryptographic Aspects of DNA Screening*:

Obfuscation resistance	Orders that have inconsequential substitution of base pairs, or that exploit the degeneracy of the codon table are still recognized	A number of obvious attacks stem from simple changes to DNA that do not affect the sequences
Database privacy	Contents of the hazardous database are kept secret	With access to a curated database of dangerous DNA sequences, skilled adversaries may be empowered to synthesize pathogens in other ways.
Low false alarm	False positives are very infrequent	A high rate of false positives require DNA providers to either manually review screening decisions (which is both slow and expensive) or deny orders unnecessarily (also expensive).
Accuracy	If a sequence is capable of causing widespread harm, it should be flagged.	A database that does not contain crucial sequences of concern, even if obfuscation resistant and function-based, will not prevent an adversary from ordering dangerous DNA
Rapid update	Additions to the database happen quickly as new hazards are uncovered	Newly released or publicized pathogen sequences
Low cost	Screening remains highly affordable (or free) as the number of nucleotides per order increases	Costs of synthesis per nucleotide will continue to significantly drop. Screening costs, however, stay relatively fixed[17] and will need to be very affordable to remain a viable option in the future.
Human interpretability	Humans can understand the reason the sequence was flagged (i.e. which dangerous sequence the order was identified as)	This option makes it more palatable to private companies, especially in the event of false positives

[17] See James Diggans and Emily Leproust. Next steps for access to safe, secure dna synthesis. Frontiers in bioengineering and biotechnology, 7:86, 2019.

The importance of high throughput, order privacy, low false alarm rate, and low cost has been broadly recognized by all the contemporary approaches to screening. However, certain properties remain in contention. Section II above discussed the functional versus the list-based view of screening. Here we'll briefly discuss a crucial property in tension: database privacy and human legibility.

Should the hazardous database be private?

Of all the existing approaches towards screening, only one (SecureDNA) is explicitly designed to keep the hazardous database private. It is worth considering, then, the importance of database privacy. While it does not appear to have been a topic of close deliberation by other systems, let us briefly consider the case for keeping the hazardous database public.

The case for a public hazardous database:

- The DNA of FSAP and CCL biological agents are already publicly available, so keeping the database public does not add harm
- A public database allows people, for instance employees at a DNA synthesis company, to see why a sequence has been flagged, which can mitigate the chance of false alarms, and perhaps crucially increase the likelihood of voluntary screening adoption[18]
- Facilitating trust and flexibility in the ecosystem; in the event that a sequence is flagged but the provider has reason to doubt the verity of the flag, a public database or else public reasoning would allow human flexibility

The case for a private hazardous database:

- An effective screening database will include more than the existing public lists, including functional variants of known pathogens and possible novel pathogens (e.g. those that have recently crossed over to humans)
- Having a comprehensive and accessible list of hazardous sequences will adversaries who wish to design novel pathogens
- Having a public database used for screening would assist adversaries in finding ways to evade screening (for example, by obfuscating sequences)

A public database makes it easier for adversaries to design sequences to evade screening. However, a private database may prevent adoption of widespread DNA screening. Gene synthesis companies may balk, for instance, at the prospect of being confronted by a pharmaceutical company demanding an explanation for a flagged order. If screening is not adopted widely, it will not be necessary for adversaries to cleverly evade screening; they will instead order sequences from a synthesis provider who does not screen.

[18] It is possible for a screening system to provide reasoning without publicly releasing all sequences. However, it

However, if biological catastrophic risk stems from intelligent and resourceful adversaries who aim to design pathogens, providing a public hazards database could provide a crucial resource for widespread harm.

An ideal outcome would clearly be to have widespread screening using a private database. The screening system could provide a helpful explanation for its decisions without revealing the underlying sequence. The tradeoffs are difficult and currently being navigated by the providers of the various screening solutions.

Finally, it is worth emphasizing that, until we begin screening and collecting data, we will not be able to estimate the frequency of flagged orders. If, for instance, flagged orders happen less than 10 times a year, the scenario of the disgruntled pharmaceutical company will likely be a nonissue. In general, we will need to adjust and adapt the screening strategy based on its deployment in the world.

The debate around the necessary properties of a secure screening system deserves far more space than we can give it here. The hope is that the aforementioned discussion has been helpful in orienting the reader towards this vital question.

Note on sensitive information:
To the best of my personal judgment, the information contained in this article does not present information that can be used for harm. If you believe otherwise, please let me know.

Thank you to Tessa Alexanian, Sam Curtis, Allan Costa, Sella Nevo, Nick Stares, Dana Gretton, and Braden Leach for feedback and support on this article.

Those wishing to cite this paper may cite it as:
Langenkamp, Max. "Clarifying the Problem of DNA Screening." (2023).

Appendix A: A Reading List

On functional screening

Godbold, Gene D., et al. "Improved understanding of biorisk for research involving microbial modification using annotated sequences of concern." *Frontiers in Bioengineering and Biotechnology* 11 (2023): 587.

On the various screening approaches

Aclid: Aclid, Inc. https://aclid.bio/ accessed May 2023.

BLISS: L. Simirenko, N. J. Hillson, S. Deutsch, and J. F. Cheng. Bliss: The black list sequence screening pipeline. In 2016 Synthetic Biology: Engineering, Evolution Design, 2016.

Fast-NA: Raytheon Intelligence and Space. https://web.archive.org/web/20230526214640/https://www.raytheonintelligenceandspace.com/what-we-do/advanced-tech/fast-na accessed May 2023.

SecureDNA:

- Baum, Carsten, et al. "Cryptographic Aspects of DNA Screening." (2020). Found also on securedna.org
- SecureDNA, “Random adversarial threshold search enables specific, secure, and automated DNA synthesis screening”

Seqscreen: Balaji, Advait, et al. "SeqScreen: accurate and sensitive functional screening of pathogenic sequences via ensemble learning." *Genome Biology* 23.1 (2022): 133.

ThreatSeq: Battelle. https://www.battelle.org/commercial-offerings/industry-solutions accessed August 2022.

NTI Common Mechanism: Nuclear Threat Initiative et al. Common mechanism to prevent illicit gene synthesis. Published March, 22, 2019.