Fast, accurate, and secure DNA synthesis screening with random adversarial thresholds

Dana Gretton[1], Erika A. DeBenedictis[1,2], Andrew B. Liu[3], Andrew C. Yao[4,*], and Kevin M. Esvelt[1,*]

[1]Media Lab, Massachusetts Institute of Technology
[2]Department of Bioengineering, Massachusetts Institute of Technology
[3]Program in Bioinformatics and Integrative Genomics, Harvard Medical School
[4]Institute for Interdisciplinary Information Sciences, Tsinghua University
*Corresponding authors. andrewcyao@mail.tsinghua.edu.cn; esvelt@mit.edu

Summary

Detecting matches to DNA sequences from possible bioweapons may enable secure and universal DNA synthesis screening.

Abstract

Global DNA synthesis is growing faster than Moore's Law, but not all DNA synthesis machines screen for sequences that could be used to build biological weapons of mass destruction. Current approaches are computationally intensive, cannot be automated due to high false positive rates, and unavoidably disclose which sequences are considered threats. Here we describe a novel architecture that identifies exact DNA and protein sequence matches within a database of randomly selected fragments from possible bioweapons and functional equivalents. Removing database entries matching unrelated sequences in GenBank can eliminate nonrandom false positives, enabling automation using algorithms permitting oblivious cryptography. Random adversarial threshold screening may offer a fast, automated, and secure method of safeguarding the international community from synthetic bioweapons of mass destruction.

Introduction

The growing power of modern biotechnology has magnified the importance of preventing access to potential bioweapons. Historical pandemics were responsible for many of the greatest catastrophes in human history, some of which resulted in tens of millions of deaths(*1*, *2*). The genomes of many responsible viruses are freely available to anyone with an internet connection, as are publications describing how they can be assembled from synthetic DNA and propagated (citations deliberately omitted). Inexpensive commercial *de novo* DNA synthesis and assembly services have made these agents accessible to a growing number of individuals with relevant technical skills. Worse, continued advances may enable the creation of bioweapons more devastating than natural pandemic agents(*3*).

A world in which many thousands of individuals can single-handedly construct and release autonomously spreading biological agents is a world unlikely to flourish.

Most individuals with the skills required to create bioweapons cannot synthesize DNA on their own and must order it from companies. Members of the International Gene Synthesis Consortium (IGSC), a trade industry group committed to biosecurity, screen commercial DNA synthesis orders above a certain length for sequences that match the Regulated Pathogen Database(*4*). Unfortunately, 20% of commercially synthesized DNA is generated by non-members who do not screen. Since the list of IGSC members is public, the extent to which current screening meaningfully restricts access to bioweapons is questionable.

The IGSC firms deserve praise for prioritizing safety because doing so is costly: current screening methods generate a nontrivial number of false positives that require evaluation by human experts. As the price of synthetic DNA continues to fall, the effective cost of screening is growing(*5*), discouraging other companies from participating.

Even if all current DNA synthesis providers screened, the anticipated arrival of benchtop synthesizers and assemblers enabling on-site gene synthesis is likely to open another window of vulnerability. As same-day production would be the sole competitive advantage of benchtop machines over commercial providers that send DNA by next-day airmail, these machines could not plausibly wait for humans to curate false positives.

Worse, advances stemming from legitimate biotechnology research may inadvertently enable the weaponization of biological agents currently thought to be innocuous. These agents may be easier to synthesize, edit, and propagate relative to more traditional pathogens. Unconstrained by the evolutionary pressures limiting the virulence of natural pathogens, engineered agents could be more destructive than any historical pandemic. Even if such a potential bioweapon were identified in advance, attempting to restrict access using current screening methods would highlight the existence of a credible weapon of mass destruction, thereby incentivizing rogue states and other well-resourced rogue actors to build the agent as a “dead-hand switch” with which to threaten the international community(*6*).

In summary, the growing number of individuals with relevant laboratory skills, the increasing relative cost of screening synthesized DNA, the anticipated arrival of benchtop DNA synthesizers that cannot employ current screening approaches, and the high probability of future discoveries enabling the weaponization of currently innocuous agents will collectively increase the risks posed by biological weapons of mass destruction.

All of these risks would be greatly reduced if DNA synthesis orders could be privately and inexpensively screened for hazards without generating false positive matches(*3*). Here we describe Random Adversarial Threshold (RAT) search, an approach designed to achieve scalable, universal, and fully automated DNA synthesis screening without risking disclosure of either synthesis orders or credible information on potential bioweapons (Fig. 1).

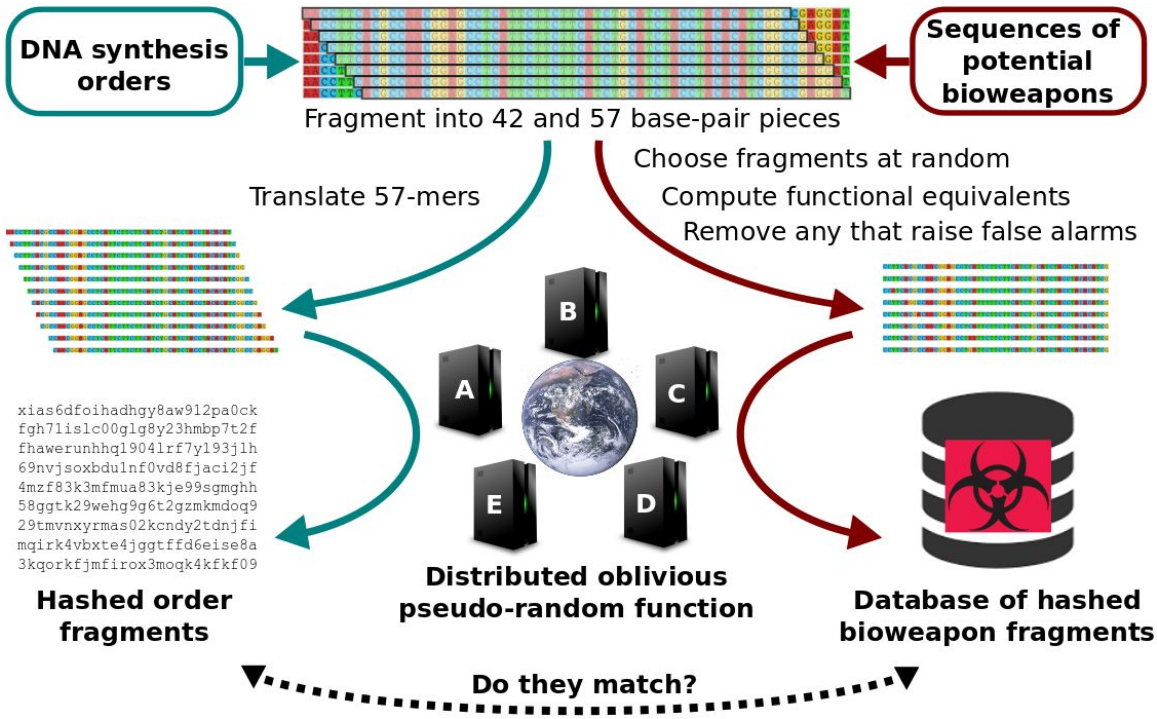


Figure 1 | Random adversarial threshold search compares sequence fragments from DNA synthesis orders to a hashed database containing randomly chosen fragments from potential bioweapons. Mutated functional equivalents of each fragment are computed and added to the database up to a randomly determined 'adversarial threshold'. Since adversaries don't know which fragments or how many variants are included, escaping RAT search requires them to build equivalent constructs with many mutations throughout the genome and hope that the resulting construct is still functional. Database fragments are short enough to preclude evasion by assembling smaller pieces, large enough to avoid random false positives, and pre-screened to ensure that they don't match unrelated sequences in GenBank. Distributed hashing using an oblivious pseudo-random function can protect the privacy of DNA synthesis orders and the database, including from theoretical quantum attacks.

Main text

Traditional DNA synthesis screening strategies rely on sequence alignment algorithms related to BLAST that identify query sequences matching "seed" regions within the genomes of known hazards, then extend the alignment in both directions while scoring for similarity(*7*). While excellent for identifying related sequences, the extension step is computationally intensive and often finds matches to unrelated genomes. Distinguishing these false positive results from actual threats requires curation by human experts, precluding full automation(*5*).

We hypothesized that a more accurate, efficient, and potentially cryptographically securable DNA synthesis screening method could be realized by searching for exact matches to a carefully curated database of sequence fragments. In Random Adversarial Threshold (RAT) screening, each DNA synthesis order is broken into short fragments of 42 base pairs and 19 amino acids for comparison to a database of equal-sized fragments derived from potential bioweapons (Fig. 1).

These sequences are large enough to minimize the number of random false positives, yet short enough that reliably assembling large constructs capable of autonomous spread from smaller pieces would be challenging(*8*). The most advanced assembly methods, and presumably under-development assembly machines, use much larger oligonucleotides(*9*). Crucially, non-random false positives can be avoided by pre-screening database additions to ensure that no entries match any unrelated sequences in GenBank.

DNA and translated peptide fragments corresponding to newly identified bioweapons that are not yet publicly known are randomly chosen for inclusion in the database using a distribution function biased towards highly conserved sequences. To prevent adversaries from evading screening by including mutations predicted to preserve function, a ranked-order list of plausibly-functional variants of each selected fragment is computed using one or more predictive models(*10–15*), then pruned to remove any entries matching sequences from unrelated genomes (Fig. 2). The top variants are included in the database out to a randomly determined "Adversarial Threshold". The higher the threshold, the lower the chance that an evasion attempt will escape RAT screening. As predictive methods improve, the database can be securely updated to keep pace by the authorized experts who added the original entries.

All sequences corresponding to a potential bioweapon are assigned a unique hazard group identification upon database entry. Any query from a DNA synthesizer that finds a RAT match in the database will reject the entire order and record the hazard group of the matching fragment. Too many hits in the same hazard group will lock the responsible synthesizer(s), report the incident to authorities, and prompt the global system to adaptively defend by adding many more previously computed fragments and variants from that hazard group to the database. Adaptive defense can prevent an adversary from interrogating the database or iteratively testing designs using one or multiple compromised synthesizers.

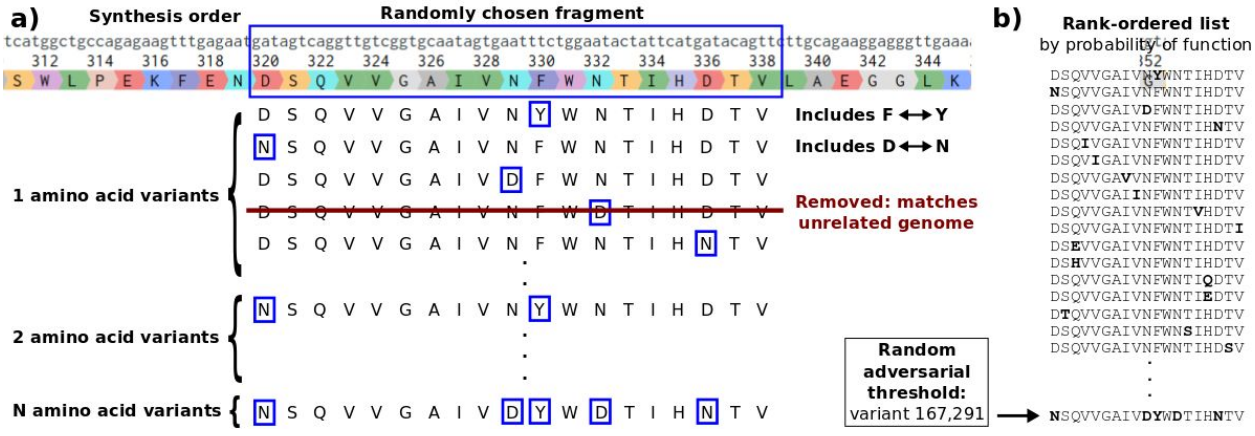


Figure 2 | The RAT database includes randomly chosen DNA fragments from noncoding regions and peptides from proteins. a) DNA variants are calculated by structural similarity; peptide variants are determined using predictive models to evaluate the likelihood that each mutated peptide remains functional. Peptides matching unrelated genomes in GenBank are removed. b) The random adversarial threshold is the number of variants included in the database from the computed rank-ordered list.

RAT screening offers five key benefits:

Speed: As an exact-match lookup, a RAT screen using fragments of length L can theoretically achieve O(L) efficiency, meaning that it swiftly returns the outcomes of all queries independent of the database size. This efficiency can be traded for additional security.

Accuracy: The random false positive rate can be set arbitrarily small by increasing the fragment length, although this must be balanced against the ease of assembling larger oligonucleotides. Choosing L = 42 base pairs and 19 amino acids define sets of over 10^{24} members, which should not generate any random false positives for a database of 10^9 entries even when 10^{15} base pairs are synthesized globally, as is anticipated in 2030. For comparison, a billion entries is more than enough to include every unique fragment present in the Reference Viral Database at the time of writing(*16*). More sophisticated estimates are detailed in the Supplementary Text.

Automation: With negligible false positives, RAT screening does not require human curation and can be fully automated. Ideally, the system would be implemented at a software level by current commercial providers and in-house synthesis cores and also built into all next-generation enzymatic synthesis machines with hardware locks.

Efficacy: Because an adversary seeking to evade a RAT search cannot know which fragments are protected, nor to what degree, they must include sufficient mutations to overcome the unknown adversarial threshold level for every possible fragment across the entire coding region of the biological agent. Such a level of diversification typically abolishes function(*17–22*), especially for autonomous agents such as viruses that encode multiple essential activities and must interact with cellular components. Moreover, the system can automatically increase the adversarial threshold upon detecting an attempt to build a specific agent by one or more synthesizers. A detailed security analysis is provided in the Supplementary Text.

Security: RAT search is compatible with distributed oblivious multiparty hashing, which can preserve the privacy of queries – and therefore of trade secrets – while also rendering the database uninterpretable. A parallel collaborative effort has detailed a scalable and secure cryptographic protocol designed to implement RAT screening, including quantum-resistant measures(*23*).

Suppose that an adversary attempts to build a mutated version of a protected component of a bioweapon that will evade screening. Assume that n genomic fragments corresponding to the hazard are included in the RAT database, with an adversarial threshold defined by v_i variants included for fragment i. Not knowing which fragments are present, the adversary must incorporate substitutions throughout the entire sequence to a level that they believe will exceed the adversarial threshold for $v_{i..n}$ by using an amino acid substitution prediction program to design and order a mutated version W (Fig. 3a).

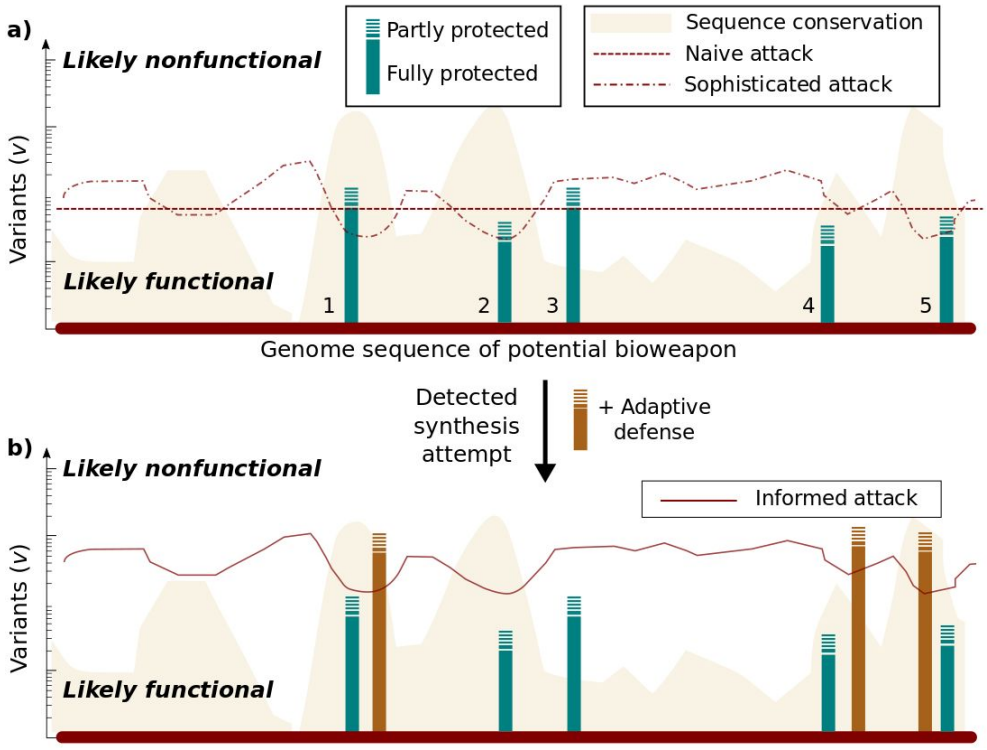


Figure 3 | Graphic representation of RAT screening against a particular bioweapon. a) Five fragments are included in the database, most but not all in conserved regions, along with a number of randomly determined variants that may be biased towards sequences more likely to be functional. A naive attack that simply introduces mutations at a constant rate is likely to be detected and also risks being nonfunctional; a sophisticated attacker may tune mutation load to the likelihood of obtaining a functional sequence across regions, but still risks discovery and creating a nonfunctional agent. b) If too many attacks on sequence in the hazard group are detected, the system adaptively adds more fragments and variants, precluding informed attacks based on probing or database interrogation.

For each fragment w_i, there are three possible outcomes:

1. w_i is present in the database, causing the entire order to be rejected without detailing which sequences were problematic and the incident to be logged; too many such matches will cause the system to increase n with higher thresholds for that agent (Fig. 3b)
2. w_i escapes detection, but contributes to making W nonfunctional
3. w_i escapes detection, and W remains functional or can be evolved to become functional

Success requires the adversary to achieve the third outcome for all w_i through w_n. In theory, choosing a single sufficiently high adversarial threshold v_i for a single fragment w_i and including all variants could span the set of functional variants: all possible w_i that are not present in the database would render W non-functional, providing perfect protection. Unfortunately, this cannot be experimentally verified in practice, and increasing v to include many combinatorial variants of a single fragment becomes costly in terms of database size, so we assume variants are included at random, albeit potentially biased towards ranked higher in the list.

Database size is limiting because more variants increase the likelihood of purely random false positive matches to synthetic sequences not found in GenBank. Together, the effective cap on the database size and the increased difficulty of evading a high adversarial threshold suggests that RAT screening will be most protective when only a handful of DNA and protein fragments are included in the database, each with as many variants as possible.

Discussion

Recent events suggest that the scientific and biosecurity communities struggle to refrain from highlighting newly discovered routes to potential weapons of mass destruction, with troubling implications(citations omitted). Suppose that a well-meaning scientist stumbles across a previously unknown bioweapon and naively attempts to warn the world. Current incentives encourage scientists to investigate the credibility of the claim, biosecurity analysts to describe the implications, and science journalists to publicize the possibility and resulting findings to the world. If subsequent experiments suggest that the threat is credible, well-resourced rogue actors would be directly incentivized to construct the bioweapon.

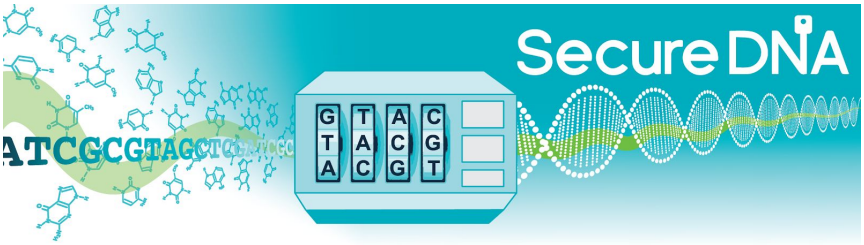
With a well-publicized RAT screening system in place, the same researcher could safely take action to restrict global access without creating information hazards(*24*, *25*) by securely conveying the information to an authorized expert. If a minimal number of experts concurred that the threat is serious, they would use their unique keys to add RAT-selected sequences from the potential bioweapon to the hashed database without requiring further disclosure (Fig. 2). Similar sequences would be chosen from a handful of 'decoys': related agents that might seem to pose a threat, but are not actually of concern. Decoys can ensure that anyone who finds a match to the database will learn only that it corresponds to a plausible-seeming agent, not that it is a credible bioweapon. This prevents adversaries from learning about potential weapons of mass destruction by checking whether known viruses are in the database through attempted synthesis using a variety of synthesizers.

Crucially, legitimate researchers studying restricted agents could obtain synthetic DNA without any delays from authorized commercial providers using 'whitelist'-compatible synthesizers. These machines would accept certificates from the relevant biosafety committee or other national authority verifying that the ordering laboratory has approval to work with the whitelisted organisms. The machine would consequently refrain from comparing synthetic DNA fragments that match the genomes of those particular organisms to the RAT database.

We harbor no illusions that RAT screening could prevent a well-resourced group from constructing any bioweapons of mass destruction that they are aware of, as such any such actors could build their own synthesizers. Rather, our proposal is intended to prevent the disclosure of new bioweapons and to raise the difficulty of obtaining such weapons for those individuals and small groups who possess only the necessary technical skills in biology.

A world in which less than a hundred groups can build autonomous bioweapons of mass destruction is far safer than a world in which many thousands of individuals can unleash such agents single-handedly.

RAT search permits private screening, is computationally efficient, and has a negligible false positive rate. Implementation appears diplomatically feasible and could eventually become universal if initiated before the market transitions to enzymatic DNA synthesis. While not perfectly protective, especially against well-resourced rogue actors, secure and universal DNA synthesis screening would substantially mitigate the potentially catastrophic risks posed by increasingly widespread access to autonomous bioweapons.



Author contributions

K.M.E. and A.Y. conceived the study and devised the random adversarial design with assistance from D.G. The system was tested by E.A.D. taking the role of an adversary, and A.L. conducted bioinformatics analyses of the Reference Viral Database. K.M.E. and A.Y. wrote the paper with contributions from all authors.

Acknowledgements

We are deeply grateful to our colleagues of the Secure DNA Project who developed the cryptographic approaches enabling secure screening: Carsten Baum, Hongri Cui, Ivan Damgard, Mingyu Gao, Omer Paneth, Ron Rivest, Vinod Vaikuntanathan, and Yu Yu. We are similarly grateful to Lan Xue, Meicen Sun, and Kenneth Oye, who advised us on policy, and to numerous members of the International Gene Synthesis Consortium who contributed to discussions of operational feasibility. Finally, we thank Emma Chory for designing the Secure DNA logo.

References

1. K. Harper, *The Fate of Rome: Climate, Disease, and the End of an Empire (The Princeton History of the Ancient World)* (Princeton University Press, 1st Edition., 2017).

2. N. P. A. S. Johnson, J. Mueller, Updating the accounts: global mortality of the 1918-1920 "Spanish" influenza pandemic. *Bull. Hist. Med.* **76**, 105–115 (2002).

3. K. M. Esvelt, Inoculating science against potential pandemics and information hazards. *PLoS Pathog*. **14**, e1007286 (2018).

4. International Gene Synthesis Consortium, Harmonized Screening Protocol V2 (2017), (available at https://genesynthesisconsortium.org/wp-content/uploads/IGSCHarmonizedProtocol11-21-17.pdf).

5. J. Diggans, E. Leproust, Next Steps for Access to Safe, Secure DNA Synthesis. *Front Bioeng Biotechnol.* **7**, 86 (2019).

6. D. Hoffman, *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy* (Anchor, 1 edition., 2010).

7. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

8. D. G. Gibson, Synthesis of DNA fragments in yeast by one-step assembly of overlapping oligonucleotides. *Nucleic Acids Res.* **37**, 6984–6990 (2009).

9. C. Plesa, A. M. Sidore, N. B. Lubock, D. Zhang, S. Kosuri, Multiplexed gene synthesis in emulsions for exploring protein functional landscapes. *Science*. **359**, 343–347 (2018).

10. Y. Bromberg, B. Rost, SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35**, 3823–3835 (2007).

11. Y. Choi, G. E. Sims, S. Murphy, J. R. Miller, A. P. Chan, Predicting the functional effect of amino acid substitutions and indels. *PLoS One*. **7**, e46688 (2012).

12. T. A. Hopf, J. B. Ingraham, F. J. Poelwijk, C. P. I. Schärfe, M. Springer, C. Sander, D. S. Marks, Mutation effects predicted from sequence co-variation. *Nat. Biotechnol.* **35**, 128–135 (2017).

13. V. E. Gray, R. J. Hause, J. Luebeck, J. Shendure, D. M. Fowler, Quantitative Missense Variant Effect Prediction Using Large-Scale Mutagenesis Data. *Cell Syst*. **6**, 116–124.e3 (2018).

14. A. J. Riesselman, J. B. Ingraham, D. S. Marks, Deep generative models of genetic variation capture the effects of mutations. *Nat. Methods*. **15**, 816–822 (2018).

15. M. Miller, D. Vitale, P. C. Kahn, B. Rost, Y. Bromberg, funtrp: identifying protein positions for variation driven functional tuning. *Nucleic Acids Res.* **47**, e142 (2019).

16. N. Goodacre, A. Aljanahi, S. Nandakumar, M. Mikailov, A. S. Khan, A Reference Viral Database (RVDB) To Enhance Bioinformatics Analysis of High-Throughput Sequencing for Novel Virus Detection. *mSphere*. **3** (2018), doi:10.1128/mSphereDirect.00069-18.

17. H. H. Guo, J. Choe, L. A. Loeb, Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 9205–9210 (2004).

18. V. E. Gray, K. R. Kukurba, S. Kumar, Performance of computational tools in evaluating the functional impact of laboratory-induced amino acid mutations. *Bioinformatics*. **28**, 2093–2096 (2012).

19. M. Miller, Y. Bromberg, L. Swint-Kruse, Computational predictors fail to identify amino acid substitution effects at rheostat positions. *Sci. Rep*. **7**, 41329 (2017).

20. V. E. Gray, R. J. Hause, D. M. Fowler, Analysis of Large-Scale Mutagenesis Data To Assess the Impact of Single Amino Acid Substitutions. *Genetics*. **207**, 53–61 (2017).

21. E. L. Jackson, S. J. Spielman, C. O. Wilke, Computational prediction of the tolerance to amino-acid deletion in green-fluorescent protein. *PLoS One*. **12**, e0164905 (2017).

22. V. O. Pokusaeva, D. R. Usmanova, E. V. Putintseva, L. Espinar, K. S. Sarkisyan, A. S. Mishin, N. S. Bogatyreva, D. N. Ivankov, A. V. Akopyan, S. Y. Avvakumov, I. S. Povolotskaya, G. J. Filion, L. B. Carey, F. A. Kondrashov, An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genet*. **15**, e1008079 (2019).

23. Baum C, Cui H, Damgard I, Esvelt KM, Gao M, Gretton D, Paneth O, Rivest R, Vaikuntanathan V, Wichs D, Yao A, Yu Y, Cryptographic Aspects of DNA Synthesis Screening. *Secure DNA Project* (2020), (available at www.securedna.org).

24. N. Bostrom, Information Hazards: A Typology of Potential Harms from Knowledge. *Review of Contemporary Philosophy*. **10**, 44–79 (2012).

25. G. Lewis, P. Millett, A. Sandberg, A. Snyder-Beattie, G. Gronvall, Information Hazards in Biotechnology. *Risk Anal*. **39**, 975–981 (2019).

Supplementary Text

Accuracy (false positive rate)

We consider the scenario that an innocent biologist is simply trying to synthesize a random DNA string *w* of length *z*. Let *p* be the probability that the request to synthesize *w* is rejected by the screening protocol, given a database of |D| elements. A simple probability calculation shows

$p <= |D|/4^L$ for DNA and $p <= |D|/20^L$ for peptides

What is the requisite size of the database |D|? We begin by making a few reasonable assumptions.

First, known hazards such as Ebola virus need not be counted. While they will be present in the database, synthesizers can pre-screen for such fragments and inform researchers if an entered sequence will cause the order to be rejected. If a previously unknown hazard becomes frequently discussed in a biosecurity context, it can similarly be designated as known.

Second, fragments matching unrelated innocuous genomes and commonly used biological tools will never be added to the database.

Third, since defensive systems that seek to save as many lives as possible should prioritize blunting the most damaging assaults, screening should focus on biological agents that can spread from host to host without human intervention.

Viruses are consequently the most concerning agents, so for rough estimation purposes we will focus on them. There are slightly less than 10^9 base-pairs of viral DNA in the Reference Viral Database at the time of writing. It's difficult to estimate how many viruses could plausibly be turned into bioweapons, and further complicated by the similarities of conserved sequences in the same family. That is, by picking random sequences biased towards conserved regions, many related viruses may be covered.

For simplicity, assume that we believe that 20 distinct virus families could be readily weaponized, with another 80 included as decoys. How many functional variants are required to protect against those 100 families? The answer depends on how many we wish to devote to guarding against each. If we choose five fragments, two DNA and three protein, and devote an average of 2 million functional variants to each of them, that totals ten million variants per family, for a total of $|D| = 10^9$ sequences in the database. Equivalently, we could choose ten times as many fragments with an average of 200,000 functional variants for each.

Next, we consider the random false positive rate we can tolerate, assuming that false positives are truly random, which may not be the case given the genetic code. For simplicity, let us stipulate that we want to see an average of zero random false positives per year until at least 2030. Assuming that 10^{12} base-pairs were synthesized in 2020, and the number will double

yearly, there will be 10^{15} base-pairs synthesized in 2030. To obtain an average of less than one false positive each year, the set of sequences defined by length L must be greater than 10^{24}. Solving, we obtain $L_D = 42$ base-pairs of DNA and $L_P = 19$ amino acids for peptides.

Can complex constructs be assembled from smaller sequences? Theoretically such methods exist, but in practice they have never been used to assemble anything even the size of the smallest virus, let alone something tha could be weaponized. Current assembly methods assume larger oligonucleotides and cannot readily be adapted to much smaller ones; different labor-intensive approaches that compensate for the thermodynamic disadvantage of smaller fragments. are required. Automated DNA assemblers should similarly be optimized for the assembly longer oligonucleotides, requiring an unusual technical skillset and a tremendous amount of patience for an individual or small group: it would likely be easier for a sufficiently well-resourced group to pursue an alternative approach.

Efficacy (effectiveness of screening against adversaries)

We consider the scenario when an adversary seeks to construct a particular harmful DNA of size n; call the target V. The adversary hopes to synthesize a DNA W such that W is functionally not too far away from V, e.g. the difference in fitness between V and W is < A. That is, the fitness must be high enough that the basic reproductive rate $R_0 > 1$; the exact value of A presumably depends upon the harmful DNA in question.

The most natural strategy for the adversary is to break W into w_1, w_2, ...,w_m (of length > L each, assuming L is small enough such that it is too difficult or laborious to assemble W from fragment of length < L) so that the fitness difference between V and W is < A (where W is the concatenated string $w_1w_2...w_m$). How likely is it that every w_i can escape screening successfully? To rephrase: Given a particular attack trying to synthesize a certain string w functionally equivalent to target a bio-weapon u, what is the probability that the database (which is randomly constructed in our recipe within limits of the ranked-order list computed) fails to reject w? Let q be that probability.

To simplify the discussion, assume for the moment that each w_i is exactly of length L. [It can be argued that this is almost the best the adversary can do.] Clearly, q is a complicated function of how w_i's are related as strings.

Scenario 1: Take the extreme case when u is of the form $u_1,u_2...u_m$ where all u_i are identical, e.g. W is a concatenated repeat of u_i. Suppose the adversary picks identical (but fixed) w_i for all i also. Let q_i be the probability that w_i passes screening under the RAT protocol.

Observation 1: Then $q = q_i$.

Observation 2: Note q_i depends on how the predicted functional variants of a fragment are chosen to be included in the database under RAT, which in turn may depend to some extent on the prediction algorithm used.

Analysis in Scenario 1: Let Z_i be the set of strings z (of length L) within a predicted function radius score A/m centered at u_i. Notice that Z_i are independent of i. Let w_i^{opt} be the string in Z_i maximizing the probability of NOT being randomly chosen by RAT; call this maximum probability $ß_i$. The adversary's best strategy is to pick $w_i = w_i^{opt}$.

Fact 2: $q = ß_i$.

Note that $ß_i$ could be 0, in which case the database always catches the attempt.

Scenario 2: Take the case when the fragments u_i are in "general positions", meaning they are distinct from one another. In this situation, with the randomness provided by RAT, we have:

Observation 1': $q = q_1 * q_2 * \ldots * q_m$.

Observation 2': q_i still depends on how variants are randomly chosen but q is less sensitive to this choice because devoting more variants to fewer fragments decreases q_i for the typical fragment, and may push $ß_i$ close to 0 for one or more. Randomization enables

Analysis in Scenario 2: Using the notation in the analysis in scenario 1 above. The only difference is that Z_i are different for distinct i. The same analysis leads to:

Fact 3: $q = ß_1 ß_2 \ldots ß_m \geq (q_{med})^{m/2}$

where q_{med} = (median of $ß_1, ß_2, \ldots, ß_m$).

Thus in Scenario 2, the chances for the adversary to succeed is exponentially smaller than in Scenario 1. Experimental testing is likely required to measure ß for a variety of fragments of differing degrees of conservation, which will help determine how many fragments are optimal, and whether more should be devoted to DNA versus proteins.