PlasmID: A Novel DNA-Based t-of-n Authentication Scheme

Joseph Kim
joekim02@mit.edu

Sruthi Parthasarathi
spar@mit.edu

James Xiu
jxiu@mit.edu

Abstract

DNA cryptography is a new field of cryptography where DNA is used to store information, instead of digital bits. Most previous DNA-based cryptographic schemes only use DNA as a trivial base-four encoding of information, and do not take advantage of the biological properties of DNA. In this paper, we present PlasmID, a novel DNA-based t-of-n authentication scheme that is purely biological and does not require the use of any computerized devices. The scheme involves sharing a set of DNA strands containing antibiotic resistance genes to each person, and authenticating groups by combining their DNA to form plasmids, which are then transformed into bacteria. Groups are authenticated if bacterial cells transformed with their joint solutions are able to resist an unknown set of antibiotics held by the authenticator. While PlasmID is cryptographically correct and secure, it involves a number of exponential dependencies which currently result in heavy costs and limitations in practice. However, we also hope this work paves the way for more cryptographic implementations that harness the full potential of biological primitives.

1 Introduction

DNA cryptography is a rapidly evolving field that involves using DNA to encode information. Using DNA brings many benefits for computation; for example, operations on DNA can be easily performed in parallel and with minimal power, and DNA allows for a huge amount of information density.

1.1 A History of DNA Cryptography

The idea of a "biological computer" was first proposed by Feynman in the 1950s. However, biotechnology at the time could not support the experiments required to put these ideas into practice. It was not until 1994 when Adleman first demonstrated the idea of DNA computation, using DNA and biotechnology to solve a small instance of the Hamiltonian Path Problem. This opened the field of DNA computation as researchers sought to explore its capabilities. Such examples include using DNA computation for SAT solving and solving the maximum clique problem in graphs [Niu+20].

DNA cryptography builds from this field, but aims to solve problems related to the hiding of messages as opposed to general computation. In 1995, Boneh et. al, was among the first to use DNA in a cryptographic setting, breaking DES with 56-bit keys over a period of 4 months [BDL96]. As a recent example, Sohal and Sharma presented a novel symmetric encryption scheme using DNA to secure cloud computation [SS22]. However, this field is still quite underdeveloped; Niu et. al [Niu+20] stated that most DNA cryptosystems rely on specific primer rules to encode data into DNA base pairs, which presents a large vulnerability in hiding the primers from an attacker.

1.2 Motivation

While many of the previous works highlight the massive potential of the field, they also only focus on using DNA as a medium of information as opposed to truly combining biological procedures with cryptography. They simply use DNA as an additional layer of encryption, encoding numbers as their base-four representations,

sequencing the DNA in order to perform computations digitally, and then synthesizing new DNA sequences representing the results. Using this approach, any existing cryptographic scheme can be trivially implemented using DNA. However, sequencing and synthesizing DNA has high time and monetary cost (on the order of days and tens of thousands of dollars), so it is difficult to justify the use of DNA in this manner. In order to more fully harness the power DNA holds, this paper aims to utilize qualities unique to DNA and relevant biological procedures in a t-of-n authentication scheme, avoiding trivial solutions where DNA is no different from a quaternary encoding.

1.3 Relevant Biological Procedures

As PlasmID relies on several biological lab techniques, we briefly outline the overarching ideas and end goals of each below.

1.3.1 DNA Binding and Replication

A highly useful property of DNA that is exploited by biological systems is its complementarity. The four bases (A, T, C, G) can be split into two pairs of nucleotides that only bind to each other (A and T, C and G). Therefore, given a DNA sequence for a single strand, the complementary strand is uniquely determined.

As the energy for binding the complementary sequence is so much lower than the energy for binding any other sequence, when complementary single strands are mixed, this reaction occurs naturally, and the two strands come together to form a single fragment of double-stranded DNA. We will henceforth refer to this process as mixing, and return to it when we discuss reconstruction of the authentication signature.

As for DNA replication, this complementarity property allows biological systems to synthesize the complementary strand given a single strand, an abundance of unbound base pairs, and enzymes that can "stitch" together the backbone of adjacent nucleotides. This process can therefore be used to convert any single-stranded DNA sequence into a double-stranded one, and will help clean up the product of the reconstruction phase.

1.3.2 Gel Electrophoresis

Carefully designed DNA fragments can form plasmids when mixed, which are circular pieces of DNA. Since plasmids are often what we wish to move into bacterial cells, one important step we can perform is to separate fully formed plasmids from excess linear DNA.

Gel electrophoresis allows for this by stratifying a sample as the componenents move through an agarose gel. Since DNA is negatively charged, applying an electric field across the gel causes the individual fragments to move from one end to the other, but the rate at which they do is dependent on physical factors such as size and polarity.

Since linear DNA is able to move through the pores of an agarose gel more quickly than circular DNA, running a gel electrophoresis can separate these two classes of fragments, and the portion of the gel containing the plasmids can be excised for further use and purification.

1.3.3 Bacterial Transformation

Once a plasmid of interest has been constructed and purified, bacterial transformation is used to introduce the foreign DNA into bacterial cells. First, the harvested cells are suspended in a cold buffer, to which the purified plasmid DNA is added. The mixture is kept on ice, but subject to brief heat shocks that increase the permeability of the cell membrane and facilitate DNA uptake. The cells can also be optionally supplemented with a recovery medium to aid with recuperation from transformation and expression of newly incorporated resistance genes, which are often included in plasmids to later help select only cells that have successfully acquired the recombinant DNA.

1.4 Bacterial Selection

As mentioned in the previous section, the last stage of a transformation procedure is selecting for the bacterial cells that incorporated the plasmid of interest. This is usually done by designing the recombinant plasmid to contain an antibiotic resistance gene that is not already found in the bacterial genome – therefore only the bacteria with the additional plasmid will be able to survive treatment with the corresponding antibiotic.

To actually perform this selection, the bacterial cells from the transformation are inoculated on agar plates containing the relevant antibiotics, and incubated at a higher temperature (typically around 37 degrees Celsius for E. coli). Cells with the recombinant plasmid then

grow and produce colonies, which form visible patches on the plate.

2 t-of-n Authentication

2.1 Overview

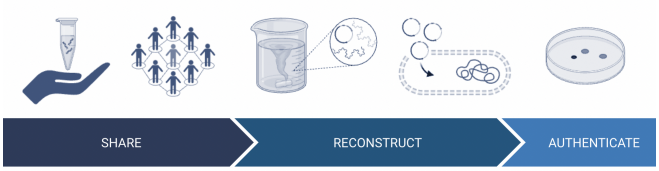

Figure 1: A schematic diagram of the proposed biological scheme for t-of-n authentication. The three stages, as shown, are distributing shares, mixing the shares to construct plasmids, and authenticating the signature by verifying that the plasmid contains all of the genes of interest.

TODO: parameters The overall scheme consists of three broad stages, as shown in the schematic diagram in Figure 1:

1. **Share**: Each participant is given a vial of single-stranded DNA fragments to use as their personal share.

2. **Reconstruct**: When a set of people come together, they mix their vials to facilitate binding of complementary sequences, from which a plasmid is formed if and only if at least t people are present. In addition, we show that such a plasmid is likely to contain at least one copy of every antibiotic resistance gene needed to survive in the authentication step.

3. **Authenticate**: The mixture is purified to retrieve copies of the plasmid (if they exist), and the copies are transformed into bacterial cells as a preprocessing step. The authenticator then grows the bacterial cells on plates containing a set of antibiotics that are unknown to the participants (henceforth referred to as the "secret"), and verifies the signature if there exist colonies that survive.

2.2 Share Generation

Figure 2: Each DNA fragment consists of 5 distinct regions that allow for individual identification, complementary binding, and a system for keeping track of the number of participants that have contributed to a growing plasmid.

We begin by discussing the generation of individual shares. In order to utilize the binding properties of DNA while maintaining security from sequencing, we propose the following. Each individual receives a vial, consisting of single-stranded fragments composed of five distinct components.

2.2.1 Identification Tags

First, each fragment has two regions denoted as identification tags. In order to account for which individuals have contributed to a plasmid, it is necessary to have a marker that is unique to each participant and independent of the secret. We implement this by generating n distinct DNA sequences of length $\lceil \log_4 n \rceil$ and assigning one to each participant as their ID tag x_i. For a given individual i, all of their fragments will have their personal tag x_i in the section labeled ID Tag 1, and the complement of one of the other $n-1$ tags, denoted x_j^C, in the section labeled ID Tag 2. This allows a fragment at the end of the growing plasmid with x_i and x_j^C to continue expanding the plasmid by binding to a fragment from the vial of individual j.

2.2.2 Antibiotic Resistance

Next, each fragment has a region containing antibiotic resistance genes. Each individual will be assigned a subset of the secret, which consists of a set of antibiotics, that their fragments will confer resistance to. Therefore this sequence is also conserved among all of the fragments in a given vial.

2.2.3 Tracking Individuals Present

Lastly, we introduce the idea of a history search tag and a memory search tag, both of which serve the purpose of helping keep track of which individuals have contributed to the plasmid as it forms. For simplicity, we start with the following proposed encoding of sets as DNA sequences: take the binary string of length n where there is a 1 in position i if the set represented by the string contains individual i and a 0 otherwise. Then convert this number into quaternary, and map each of the residues modulo 4 to one of the DNA nucleotides.

Then for each fragment with x_i and x_j^C, we let the history tag correspond to a subset s of $[n]$ containing i. Once the fragment binds a plasmid, it will have to bind a fragment from vial j next, so the memory tag encodes the complement of the encoding of $s \cup j$. We note two exceptions to this rule: fragments that start the plasmid (s would have size 1) have a universal start sequence as their history tag, and fragments that end the plasmid (s has size t) have the complement of the universal start sequence as their memory tag and no ID Tag 2.

2.2.4 Share Size

Now, we note that while the ID Tag 1 and Antibiotic Resistance regions are fixed for any given individual i, and the complementary ID Tag 2 region of a fragment is uniquely determined by the history search and memory tags, each possible combination of a history search and memory tag gives rise to a fragment that must be included in vial i.

As each subset of $[n]$ containing i of size k can further bind to any of the $n-k$ ID tags corresponding to individuals not in s, the total number of such combinations is upper bounded by

$$\sum_{k=0}^{n-1} \binom{n-1}{k} * (n-(k+1)) \approx \sum_{k=0}^{(n-1)/2} \binom{n-1}{k} * n$$
$$= n * 2^{n-2}$$

.

Therefore we synthesize at least one copy of each of the $O(n * 2^{n-2})$ fragments and combine them in a vial to produce a single share.

2.3 Plasmid Reconstruction

The next phase of the scheme is reconstruction, in which some number of individuals come together and mix their vials. When this occurs, regions of single-stranded fragments that contain complementary sequences will automatically bind upon interaction. In this section, we verify that a plasmid forms if and only if there are at least t participants.

Assume, for the sake of contradiction, that a plasmid forms after $k < t$ fragments have joined together.

Let s_i denote the subset encoded by the history search tag of the i^{th} fragment in a chain of DNA fragments found in the mixed vial, and let t_i similarly denote the corresponding fragment's memory tag. Anything that binds t_i must be equivalent to $t_i^C = s_{i+1}$, and as t_i corresponds to the complement of the encoding of $s_i \cup j$ for some $j \notin s_i$, $|s_{i+1}| = |s_i| + 1 \forall i \in [1, t-1]$.

Then the last fragment has a history search tag encoding s_k and a memory tag encoding s_{k+1}^C, which is not the complement of the universal start sequence.

The only exception to this forced linear chain growth is when t_i does not encode $s_i \cup j$, which is true for fragments with $|s_i| = t$. Therefore the chain can only become circular when t fragments are reached.

2.4 Authentication

The authentication process involves purification of plasmid, transformation of the reconstructed plasmid into bacteria, then the growth of the transformed bacteria in the presence of antibiotics. We will describe this scheme in detail below.

First, we purify the plasmid using gel electrophoresis, as in section 1.3.2, and replicate it using PCR, as in section 1.3.1. Then, the bacteria is transformed with the length-T plasmid using the method described in section 1.3.3. The transformed bacteria is then grown in the presence of the m secret antibiotics, where colonies that have resistance to all the antibiotics will survive, as detailed in 1.4. Finally, the step involves checking the status of the bacteria after one day of incubation. If the bacteria colony survives, then authenticate the group of people. If the colony dies, do not authenticate them.

The transformed bacteria will express the genes in the plasmid that they are given. Thus, if the length-T plasmid contains resistance genes to all m antibiotics, then the transformed bacteria will survive. In this scheme,

if $T \geq t$, then it is likely that the plasmid contains all resistance genes, and if $T < t$, if is unlikely that the plasmid contains all needed resistances. The details for this analysis will be given in section 3.1.

3 Evaluation

In this section, we will evaluate PlasmID on cryptographic correctness and security. We will also discuss the time and money cost of implementing PlasmID, and its practical limitations.

3.1 Cryptographic Evaluation

Correctness. We will show that in PlasmID, any group of $T \geq t$ people will authenticate with high probability. First, note that if a plasmid forms with resistance to all m antibiotics, then the bacteria colony that receives that plasmid will survive growth, so the group will authenticate. The linkage scheme described in section 2.2 guarantees that given enough time (around 12-24 hours), a length-T plasmid forms, which contains genes from everyone in the subset. TODO: Call such a plasmid a **length-T plasmid**. The frequency of the length-T plasmid can be amplified such that there is guaranteed to be bacteria that receive it and survive. Thus, we need to show that with high probability, the length-T plasmid contains resistance genes to all m antibiotics.

Recall that each person receives m genes (chosen randomly with replacement), where $m = e^{t-1}$. The length-T plasmid produced by the T people will contain $mT \geq mt = m(\log m + 1)$ genes. Consider the following game, where one repeatedly picks one gene randomly with replacement from the pool of m genes until the set of picked genes contains all m genes in the pool. Let random variable X represent the number of times we pick a gene in the game. Then the probability that the mt randomly chosen genes contain the whole pool of m genes can be expressed as $Pr(X \geq mt)$.

The game represents the coupon collector problem, and the distribution of X is well-known. By a theorem by Erdős and Rényi [ER61] (see theorem 5.13 in [MU17] for a proof based on Poisson approximation), $Pr(X < m(\log m + 1)) \to e^{-e^{-1}}$ as $m \to \infty$. As such, $Pr(X \geq mt) = 1 - Pr(X < m(\log m + 1)) \to 1 - e^{-e^{-1}} \approx 0.692$. Since this probability is greater than $\frac{1}{2}$, it can be made arbitrary close to 1 by repeating the authentication experiment, and authenticating the group if their bacteria colony survives a majority of the time, proving correctness.

Security. Now, we will show that any group of $T < t$ people will not authenticate with high probability. As mentioned in 2.3, ideally, no plasmids form upon mixing, so the transformation will not actually grant bacterial cells any resistance, and the authentication will always fail. However, in reality, DNA can bind to another sequence even if differs from the exact complementary sequence in some base pairs. This effect is particularly pronounced when the number of mismatches is relatively small compared to the length of the entire binding region.

While we could combat this by choosing sequence representations that differ in multiple positions, fragment synthesis is expensive – there is an inherent trade-off between optimizing synthesis costs and reducing the likelihood of erroneous binding. Therefore, in the case that we opt for shorter encodings of distinct tags that may differ by just a few base pairs, we still want to upper bound the probability that the group of $T < t$ people successfully authenticates. To do this, we must show that the length-T plasmid that forms from this group does not contain all m genes with high probability. First, note that the plasmid contains at most $m(t-1) = m \log m$ genes. Thus, with the same definition of random variable X from the correctness anaylsis, the probability that the plasmid does not contain all m genes can be expressed as $Pr(X < m \log m)$. By the Erdős-Rényi theorem, we have that as $m \to \infty$, $Pr(X < m \log m) \to e^{-e^0} = e^{-1} \approx 0.369$. As before, we can make this probability arbitrarily close to 0 by applying a majority algorithm, wherein we repeat the authentication experiment and do not authenticate if the bacteria colony dies a majority of the time.

We must additionally show a group of size less than t cannot learn the secret, which is the pool of m antibodies. Each person's share does not contain genes given to other people; the share only contains their own genes. Thus, by sequencing their own DNA, a person can only learn the set of distinct genes they were given, which is $O(\frac{m}{\log m}) = o(m)$ on expectation. A group of $T < t$ people can only learn the antibiotic whose resistance genes they possess by sequencing their combined vials, which is unlikely to be the full pool of antibodies, as shown in the above security proof.

Finally, we model the authentication process as performed by a third party who does not share the inter-

mediate results of the authentication progress, so we do not consider attacks which are based on analyzing the antibody solution to try to find the compounds which make up the antibodies.

Thus, any group of $T < t$ people cannot authenticate with high probability, and also cannot learn the secret with high probability.

3.2 Cost and Limitations

Now, we will estimate the time and monetary cost of performing share and authentication and discuss the practical limitations of PlasmID.

Sharing involves synthesizing at most $n2^{n-2}$ strands for each of n people, with each strand containing $m = e^{t-1}$ genes, along with auxiliary sequences, such as the history search tag, ID tags, and memory tag. On average, each resistance gene contains 800 base pairs [Sut78], so each strand contains about $800e^{t-1}$ base pairs (we ignore the auxiliary sequences since the number of base pairs in the auxiliary sequences is much less than $800e^{t-1}$). Thus, share involves synthesizing $n^2 2^{n-2}$ strands of length $800e^{t-1}$. The price of DNA synthesis is \$.39 per base pair at IDT [Tec], resulting in total cost of $312n^2e^{t-1}2^{n-2}$ dollars. This cost quickly explodes as n and t become large. For example, when $n = 8$ and $t = 4$, the cost of share is 25.7 million dollars. Synthesizing DNA also takes one week to one month at most companies, so Share has a significant time overhead as well.

Reconstruct and Authenticate involves plasmid formation, transformation of plasmid into bacteria, and growth of bacteria cultures. Plasmid formation has zero monetary cost, as it just involves mixing DNA vials. Transformation and bacterial growth requires an incubator and lab equipment, which cost around \$1000 [Sci]. Performing a single transformation is very cheap, since it only requires procuring commonly accessible lab equipment and antibiotic solutions. Transformation typically takes a few hours, and bacterial growth takes a few days, so authentication takes days to complete. Multiple authentication experiments can be ran in parallel by transforming and growing separate bacteria cultures.

The exponential number of strands that must be given to each person, and the exponential number of genes each strand contains causes the time and cost of share skyrocket exponentially with n and t, though the costs of reconstruct and authenticate are fixed. As such, PlasmID is only feasible when n and t are small, and to be practically applicable to larger groups, the exponential dependencies must be reduced.

4 Conclusion

In this paper, we gave a novel t-of-n authentication scheme based in DNA which uses only biological processes. Our new scheme is valuable since it departs from the previous paradigm of using DNA only as a base-four representation of data and not directly performing operations on DNA. Authentication in PlasmID can be performed independently of DNA sequencing and computers.

In the future, we hope to improve the exponential dependencies on n and t from share and authenticate to polynomial ones. We hypothesize this can be done by using a more clever scheme to enforce the formation of a length-T plasmid, and by changing the parameters around how genes are distributed to people. These improvements will allow PlasmID to scale well, in terms of cost, when n and t are large. We also hope that this scheme inspires other purely biological solutions to other cryptographic problems.

5 Individual Contributions

All three authors contributed to the final cryptographic scheme. In addition, Kim researched previous works and motivation and coordinated the work, Parthasarathi researched the biological processes involving DNA, and Xiu helped in coming up with the scheme, and analyzed the cryptographic correctness and cost of the scheme.

References

[BDL96] Dan Boneh, Cristopher Dunworth, and Richard J Lipton. “Breaking DES Using a Molecular Computer”. In: *DNA Based Computers* 27 (1996), pp. 37–66.

[ER61] Paul Erdős and Alfréd Rényi. “On a classical problem of probability theory”. In: *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei* 6 (1961), pp. 215–220.

[MU17] Michael Mitzenmacher and Eli Upfal. *Probability and computing : randomization and probabilistic techniques in algorithms and data analysis (2nd ed.)* United Kingdom: Cambridge University Press, 2017.

[Niu+20] Ying Niu et al. "Review on DNA Cryptography". In: *Bio-inspired Computing: Theories and Applications*. Ed. by Linqiang Pan, Jing Liang, and Boyang Qu. Singapore: Springer Singapore, 2020, pp. 134–148. ISBN: 978-981-15-3415-7.

[Sci] Southwest Science. *50 Liter (1.8 cuft) Premium Forced Air Incubators*. southwestscience.com/product/50-liter-premium-forced-air-incubator. Accessed: 2024-05-14.

[SS22] Manreet Sohal and Sandeep Sharma. "BDNA-A DNA inspired symmetric key cryptographic technique to secure cloud computing". In: *Journal of King Saud University - Computer and Information Sciences* 34 (1 2022), pp. 1417–1425.

[Sut78] J. Gregor Sutcliffe. "Nucleotide sequence of the ampicillin resistance gene of Escherichia coli plasmid pBR322". In: *Proc Natl Acad Sci USA*. 75.8 (1978), pp. 3737–3741.

[Tec] Integrated DNA Technologies. *Gene Synthesis | IDT*. https://www.idtdna.com/pages/products/genes-and-gene-fragments/custom-gene-synthesis. Accessed: 2024-05-14.