DNA SECURITY

Pregledni članak

Magdalena Punčeva Marina[1]
[1]MIT University Skopje, Treta Makedonska Brigada 66a, Sjeverna Makedonija,
e-mail: magdalena.punceva@gmail.com

Abstract
Whole genome sequencing is easier than ever and helps advancing the medical research and personalized healthcare. As extracting the genomic data becomes faster and less complex, individual rights in privacy must be respected and secured. The main focus of this paper is to study and compare models and algorithms that protect the genomic data by using the most fundamental form of data protection known today: information-theoretic (I-T) security. A cryptosystem is I-T secure even when the adversary has unlimited computing power and knows the secret codebook. Information theory offers tools to design codes that maximize the useful communication rate while minimizing the "leak" of confidential information to unwanted parties. In order to study and protect genomic data one needs multidisciplinary approach using High Power Computing and genome analysis. We conclude the paper by advising how the structure of the genome may be used to make the data protection more efficient.

Introduction

Five nucleobases—adenine (A), cytosine (C), guanine (G), thymine (T), and uracil (U)—are called primary or canonical. They function as the fundamental units of the genetic code, with the bases A, G, C, and T being found in DNA while A, G, C, and U are found in RNA.
Deoxyribonucleic acid (DNA) is a polymer composed of two polynucleotide chains that form a double helix molecule. DNA sequence is a natural large-scale information storage system as it contains the unique genetic information for the development, functioning, growth and reproduction of all organisms and also many viruses. It is composed of sequences of only four nucleotide bases adenine (A), cytosine (C), guanine (G) and thymine (T). The order of these four bases is unique for each individual organism. The process of determining the order of bases is called sequencing. It is a laboratory procedure that determines the order of bases in the genome of an organism in one process [1].
In recent years, great progress has been made with regard to increasing the availability and lowering the cost of the process of DNA sequencing, and also storage and retrieval of DNA data. The availability of storing and retrieving such data, has contributed to great success and has opened up unprecedented research possibilities in biomedical fields. However, human DNA data consists of highly sensitive information, since the genome contains the entire DNA set, it therefore represents all the genetic information of an individual [2]. Thus, if the sequence of the nucleotide bases in an organism gets discovered, then also its unique DNA fingerprint, or pattern, becomes known, hence reveling not only information about an individual but also about her ancestors, descendants and relatives.
Some genomic databases are publicly available for research purposes [3]. Besides the above mentioned advantages brought by sequencing and digitization of DNA data, these technologies involve risks with respect to privacy and security, and therefore special attention is needed for protecting the security and privacy of these data. Since genomic data contain genetic information about an individual, potential leakage of such data, can also affect the privacy of his family and closed relatives. Furthermore, the impact of a leakage of such data is permanent since person’s genetic fingerprint can not be replaced like for example a credit card when it gets stolen.
Therefore we can regard that the information that is contained in DNA, is actually stored as a code made up of the four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T) and this code is the representation of the hereditary material in almost all living organisms. The sequence of these chemical bases indicates the information available for building and maintaining an organism, and it needs to be stored and retrieved in a secure way.
In this paper, we analyze the questions of security and privacy of DNA data, by outlining the potential problems, weaknesses and vulnerabilities of current methods and solutions for storage and transfer of such data. We focus on questions and challenges with respect to the sheer sizes of these datasets, and the scalability issues and propose directions for future research.

1. Problem formulation

In order to provide sufficient level of privacy and security of the DNA data storage and data transfer, we use the approach that is standard in information security and provides balanced data protection. It is known as CIA triad, that is short for confidentiality, integrity and availability (CIA). Below we give a short explanation about the three CIA attributes:

- **Data confidentiality** requires that only those individuals who are authorized to access the data should be able to access it, and no unauthorized access should be allowed. Data must be protected from reading or alterations.
- **Data integrity** means that the original data must be preserved in its entirety, and changes, either intentional or unintentional, should not happen, as that can have serious consequences on further use of the data.
- **Data availability** means that data must be continually available for access to the authorized individuals.

We will analyze and discuss how current solutions satisfy these conditions. In addition, it will be explored how these conditions but also data privacy of the stored DNA data are satisfied with current solutions. We shall also focus on the technological solutions that allow for efficient storage and retrieval of DNA data, especially in the presence of attacks such as SQL injection, XSS attacks, file leakage. Finally, we shall discuss another approach, namely DNA as a storage, which relies on synthetic DNA, as a means for storing data as large information storage system.

2. Related work

2.1. Storing DNA data

Storing sequenced DNA data requires addressing several challenges which are mostly related to scalability of the algorithms, compression, data security and privacy. In [4] the authors propose a hash based data structure for compression of the DNA data. The algorithm performs efficiently both on repeated and non-repeated patterns in DNA sequences.
In [2] several attacks and threats on the DNA databases have been documented. They focus on the fact that the use of digital technologies have increased the risk to DNA and genomic data with respect to threats such as social engineering and exploits, genomic dossiers, DNA theft, genomic data theft, and genetic warfare. On such documented attack is a famous DNA testing service which was breached by cybercriminals in 2017 where hackers were able to breach 92 million accounts. In this case the attackers only accessed encrypted ID and passwords, and the original DNA or genetic data remained safe. However, such events emphasize the need to analyse and understand security weaknesses of DNA databases and to develop adequate solutions to protect them.

2.2. DNA as a storage

Skinner et al. in 2007 [5] demonstrated a simple DNA-based data storage scheme in which information is written using addressing oligonucleotides. Mardis [6] discusses the recent scientific discoveries that resulted from the application of next-generation DNA sequencing technologies. In 2012, Church [7] used a simple coding scheme to translate the bit sequence to ACGT sequence. In his scheme, A and C represent 0, and G and T represent 1. In this way, he encoded his new book, some image files, and a JavaScript programme, which resulted in 700 *kB* of information. In 2013, Goldman et al. [8] have encoded similar amount of data, however with a different scheme. In their system, every byte (eight bits) is converted into four DNA letters (A,C,G,T). This is logical as for encoding four letters one needs two bits. In this way they achieve lower error rate when reading the data compared to Church’s method. Grass et al. [9]

in their work show that digital information can be stored on DNA and recovered without errors for considerably longer time frames. To allow for the perfect recovery of the information, they encapsulate the DNA in an inorganic matrix, and employ error-correcting codes to correct storage-related errors. Specifically, they encoded 83 *kB* of information to 4991 DNA segments, where each one is 158 nucleotides long. More importantly, they showed that data can be archived on DNA for millennia under a wide range of conditions. As an alternative storage media, DNA surpasses the information density and energy of operation offered by flash memory. In [10] Erlich and Zielinski demonstrated the huge potential of the DNA to provide large-capacity information storage. They presented a method called DNA Fountain, which approaches the theoretical maximum for information stored per nucleotide. They demonstrated efficient encoding of information, including a full computer operating system into DNA, that could be retrieved at scale after multiple rounds of polymerase chain reaction. Industry is also quite interested in DNA data storage. In May 2017, Microsoft Research has formalized a goal of having an operational storage system based on DNA working inside a data center toward the end of this decade. The aim is a proto-commercial system in three years storing some amount of data on DNA in one of their data centers. Microsoft also harbors the even more ambitious goal of replacing tape drives, a common format used for archiving information. These plans signal how seriously some high-tech companies are taking the seemingly strange idea of saving videos, photos, or valuable documents in the same molecule our genes are made of. DNA is the densest known storage medium in the universe, just based on the laws of physics. In July 2016, Microsoft publicly announced it had stored 200 megabytes of data in DNA strands, including a music video, setting a record. Major obstacles to a practical storage system remain. Converting digital bits into DNA code (made up of chains of nucleotides labeled A, G, C, and T) remains laborious and expensive due to the chemical process used to manufacture DNA strands. Today, the main issue with DNA storage is the cost. According to Microsoft, the cost of DNA storage needs to fall by a factor of 10,000 before it becomes widely adopted. Automating the process of writing digital data into DNA will also be critical. The rate of moving data into DNA was only 400 bytes per second. Microsoft says that needs to increase to 100 megabytes per second. Reading out the data is much easier. That is done using a high-speed sequencing machine, including to recall specific parts of the files, analogous to random access memory on a computer. Even a twofold improvement in DNA reading would make that aspect of the system effcient enough for commercial use. Because writing and retrieving data into DNA is slow, any early use of the technology will be restricted to special situations. That could include data that needs to be archived for legal or regulatory reasons, such as police, bodycam video, banking or medical records. Microsoft currently works with Twist Bioscience, a DNA manufacturer located in San Francisco. Twist is one of a number of newly formed companies trying to improve DNA production, a list that now includes startups: DNAScript, Nuclera Nucleics, Evonetix, Molecular Assemblies, Catalog DNA, Helixworks, and a spin-off of Oxford Nanopore called Genome Foundry. One exciting possibility being pursued by some of the startups is to replace the 40-year old chemical process used to make DNA with one that employs enzymes, as our own bodies do. Technicolor Research, in Los Altos, is funding such work at Harvard University, in the laboratory of the genomic expert G. Church. The molecule is so stable that it is frequently recovered from mammoth bones and ancient human remains. But its most important feature is density. DNA can hold quintillion, that is 10^{18} bytes of information in a cubic millimeter. In addition to being dense and durable, DNA has a further advantage that’s not often mentioned — its extreme relevance to the human species. Think of those old floppy

disks you can’t read anymore or clay tablets with indecipherable hieroglyphs. Unlike such media, DNA probably won’t ever go out of style. We’ll always be reading DNA as long as we are human [11].

Recently there was a great deal of work related to different ways of DNA storage. Some of them include random-access DNA storage [17], nucleid acid memory [18], error free data storage [19], asymmetric lee distance codes for DNA-based storage [20], or mutually uncorrelated DNA data storage [21]. Several researchers focused on DNA based image storage [24, 25, 26].

In 2019, Chandak et al. [22] and Fei and Wang [23] explored the use of low density parity check codes (LDPC) for improving the cost of read and write operations during DNA storage. Several results published in 2020, made further progress regarding multicomponent molecular memory [27], information storage in small-molecule mixtures [28], coded traced reconstruction [29], DNA punch cards for storing data via enzymatic nicking [30] and nanopore-based DNA hard drives for rewritable and secure data storage [31].

In [32], Pan et al. propose a two-dimensional molecular data storage system that records information in both the sequence and the backbone structure of DNA and performs nontrivial joint data encoding, decoding and processing.

Lim et al. in 2023 went a step further by proposing a biological camera that captures and stores images directly into DNA [33].

3. Methods and algorithms for achieving secure DNA data

3.1. Cryptography

Standard cryptographic methods are roughly classified as either symmetric or asymmetric. Generally, cryptographic methods use a key in order to encrypt the data and they are considered safe, as long as the key is safe. However, the major challenge with the symmetric algorithms is the exchange of the key between the two communicating parties, which needs to be done via separate secure channel.

3.2. Differential privacy

There exist several methods for preserving data privacy. Classical methods for privacy preserving are based on anonymization of the data, which is a procedure that removes or modifies personally identifiable information such that the data cannot be linked to the person to whom they belong. However, it has been shown that with the reverse process known as de-anonymization, surnames can be recovered from a genetic genealogy databases [12]. Therefore, the challenge lies in the fact that removing identifiable data on one hand and retaining useful information for further research on the other hand are sometimes conflicting goals.

Differential privacy [13] is relatively novel method that is proposed to addresses the issues of privacy by adding controlled amount of noise to the data. Computations and requests can thus be performed against the database without revealing the privacy of individuals that are part of it. Differentially private algorithms provably resist against re-identification and linkage attacks.

3.3. DNA as medium for data storage

In this paper, we rely on the original expertise in the area of reliable (error less) storage and the potential of the DNA as a natural storage medium of the genetic information over the generations. DNA is a universal and fundamental data storage mechanism in biology since it keeps and transfers the genetic information and it is long lasting. DNA data storage broadly refers to any scheme for storing digital data in the base sequence of the DNA.

It uses a synthetic DNA made using commercially available oligonucleotide synthesis machines for storage and DNA sequencing machines for retrieval. This type of storage system is more compact than current magnetic tape, hard drive storage systems or solid-state drive (SSD) devices, due to the data density of the DNA. It also has the capability for longevity, as long as the DNA is held in cold, dry and dark conditions, as it was discovered with the DNA of frozen animals found in permafrost. It is, however, a slow process, as the DNA needs to be sequenced in order to retrieve the data, so the method might be used for low access rate such as long-term archival of the big data. A team of scientists lead by N. Goldman has encoded and stored all Shakespeare’s sonnets in DNA. The researchers say their method could easily be scaled up to store all of the data in the world. DNA packs information into much less space than other media, the data density per cube centimeter is million times higher than the state-of-the-art storage media. In other words, the entire current world data could be stored in few kilograms of DNA. At the same time, the data longevity is much higher and the power usage per gigabyte is around 100 million times lower. However, before DNA can become a viable competitor to conventional storage technologies, many challenges must be solved, from reliably (without errors) encoding information in DNA and retrieving only the information a user needs, to making nucleotide strings cheaply and quickly enough. We explore and study all the advantages and issues for the potential use of error correcting codes to store large amounts of data in the DNA nucleotide. As any data is transformed into a sequence of binary digits (zeros and ones), it can be easily mapped into a sequence of the DNA base pairs (adenine-00, cytosine-01, guanine-10 or thymine-11). Then, this sequence will be used to produce the synthetic DNA that will be the storage medium. As we mentioned above the synthesis, or production of such DNA is still a slow and expensive process. By sequencing this DNA, we can retrieve the data that was stored. The open issues include the time and cost of synthesizing DNA strings, the error made by writing/reading, and the time of DNA sequencing. This is indeed a novel approach that will open many new applied research possibilities on the cross road of these interdisciplinary activities.

4. Information-theoretic security

If a cryptosystem is secure in an information theoretic sense, it means that even when the adversary has unlimited computing power and knows the secret codebook, it will be not able to get any information about the protected information. Information theory offers tools to design codes that maximize the useful communication rate while minimizing the “leak” of confidential information to unwanted parties [14].

In information theory, there exists a notion of perfect secrecy or information-theoretic secrecy. In contrast to standard cryptography, information-theoretic secrecy is secure even if the intruder has unlimited computing power and/or unlimited memory. In order for an algorithm to be informationtheoretically secure, the ciphertext produced by it, should not provide any

knowledge about the plaintext, without using the key. More formally, a cryptographic algorithm provides perfect secrecy if and only if for any probability distribution from which the plaintext is drawn, and for any plaintextciphertext pair (M,E) it holds that: $Pr(m = M|K(m) = E) = Pr(m = M)$ for all ciphertext E, and all corresponding plaintext M [15]. In other words, the adversary cannot learn more about the original plaintext by observing the ciphertext. Equivalently, observing the ciphertext will do the same as guessing the message among all possible messages. On possible encoding that reaches perfect secrecy is the Vernam cipher [16] or the so called one-time pad. In this algorithm if we provide a random binary key that is the same size as the binary plaintext, by adding them modulo two (XOR), we are getting a ciphertext that is equivalently random as the random key. The drawback of such an algorithm is that the key must be the same size as the message, and cannot be re-used. Also the key has to be completely random and the probability of 0's and 1's must be equal.

5. Secure DNA storage

In our line of work we want to explore the secure DNA data storage. It will use not only optimal storage codes that provide reliability, i.e., low error probability, but also secrecy using informationtheory approach. In their work [31] Chen et al. discuss nanopore-based DNA hard drives for rewritable and secure data storage, however they do not use information-theoretic security. By proposing information-theoretic secure storage codes, we will not only have an elegant and long lasting way of storing the data, but the data will be stored in a fundamentally secure way such that only legitimate users will have access to it. Note however, that finding such codes will be a big challenge.

6. Conclusion and future work

Huge challenge of the present day, will be the storage of data, since the amount of produced information may soon exceed the capacity of traditional storage media (magnetic tapes, hard drives, flash memories). DNA is a molecule that carries the genetic information used in the growth, development, functioning and reproduction of all known living organisms and many viruses. Most DNA molecules consist of two biopolymer strands coiled around each other. The two DNA strands are called polynucleotides since they are composed of simpler units called nucleotides. Each nucleotide is composed of one of four nucleobases — cytosine (C), guanine (G), adenine (A), or thymine (T).

We focus here on several features that make DNA a serious competitor to the present day storage media such as the data density per cubic centimeter, the power usage and data retention in years, and outline future directions in order to address the existing challenges.

References

[1] Center for Disease Control and Prevention, *Whole Genome Sequencing*, www.cdc.gov/pulsenet/pathogens/wgs.html, 2022.

[2] S. Arshad, J. Arshad, M. Mubashir Khan and S. Parkinson, *Analysis of security and privacy challenges for DNA-genomics applications and databases*, Journal of Biomedical Informatics, 2021.

[3] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, J. Ostell, K.D. Pruitt and E. W. Sayers, *Genbank*, Nucleic Acids Res. 46 (2018) D41–D47.

[4] A. Mehta and B. Patel, *DNA Compression Using Hash Based Data Structure*, International Journal of Information Technology and Knowledge Management July-December, Volume 2, No. 2, pp. 383-386, 2010.

[5] G. M. Skinner, K. Visscher and M. Mansuripur, *Biocompatible Writing of Data into DNA*, Journal of Bionanoscience, Volume 1, Number 1, pp. 17-21(5), June 2007.

[6] E. R. Mardis, *Next-generation DNA sequencing methods*, Annual Review on Genomics and Human Genetics, 2008;9:387-402, 2008.

[7] G. M. Church, Y. Gao and S. Kosuri, *Next-generation digital information storage in DNA* Science, vol. 337, no. 6102, pp. 1628–1628, Sep. 2012.

[8] N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E.M. LeProust, B. Sipos and E. Birney, *Towards practical, high-capacity, low-maintenance information storage in synthesized DNA Nature* vol. 494, no. 7435, pp. 77–80, 2013..

[9] R. N. Grass, R. Heckel, M. Puddu, D. Paunescu, W. J. Stark, *Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes*, A Journal of teh German Chemical Society, Volume 54, Issue 8, pp. 2552 – 2555, 16.02.2015.

[10] Y. Erlich and D. Zielinski, *DNA Fountain enables a robust and efficient storage architecture*. Science. March, 355(6328):950-954, 2017.

[11] A. Regalado, *Microsoft Has a Plan to Add DNA Data Storage to Its Cloud*, MIT Technology Review, 22.05.2017.

[12] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin and Y. Erlich, *Identifying personal genomes by surname inference*, Science 339, (6117) 3214-4, 2013.

[13] C. Dwork and A. Roth, *The Algorithmic Foundations od Differential Privacy*, Foundations and Trends in Theoretical Computer Science, 2014.

[14] M. Bloch, O. Gu¨nlu¨, A. Yener, F. Oggier, H. V. Poor, L. Sankar and R. Schaeffer, *An Overview of Information-Theoretic Security and Privacy: Metrics, Limits and Applications*, IEEE Journal on Selected Areas In Information Theory, VOL. 2, NO. 1, March 2021.

[15] C. E. Shannon, *Communication Theory of Secrecy Systems*, Bell System Technical Journal, vol. 28(4), page 656–715, 1949.

[16] G. S. Vernam, *Secret Signaling System*, United States Patent, United States Patent Office, Patent US-1310719, 22.07.1919.

[17] S. M. H. T. Yazdi, Y. Yuan, J. Ma, H. Zhao and O. Milenkovi´c, *A rewritable, random-access DNA-based storage system*, Scientific Reports 5, 14138, 2015.

[18] V. Zhirnov, R. M. Zadegan, G. S. Sandhu, G. M. Church and W. L. Hughes, *Nucleic acid memory*, Nature Materials 15, 366-–370 (2016).

[19] S. M. H. T. Yazdi, R. Gabrys and O. Milenkovi´c, *Portable and error-free DNA based data storage*, Scientific Reports 7, 5011, 2017.

[20] R. Gabrys, H. M. Kiah and O. Milenkovi´c, *Asymmetric lee distance codes for DNA-based storage*, IEEE Transactions on Information Theory 63, 4982–4995, 2017.

[21] S. M. H. T. Yazdi, H. M. Kiah, R. Gabrys and O. Milenkovi´c, *Mutually uncorrelated primers for DNA-based data storage*, IEEE Transactions on Information Theory 64, 6283 – 6296, 2018.

[22] S. Chandak, K. Tatwawadi, B. Lau, J. Mardia, M. Kubit, J. Neu, P. Griffin, M. Wootters, T. Weissman, H. Ji, *Improved read/write cost tradeoff in DNA-based data storage using LDPC codes*, 57th Annual Allerton Conference on Communication, Control, and Computing, 147–156, 2019.

[23] P. Fei and Z. Wang, *LDPC codes for portable DNA storage*, IEEE International Symposium on Information Theory (ISIT), 76–80, 2019.

[24] M. Dimopoulou, M. Antonini, P. Barbry and R. Appuswamy, *A biologically constrained encoding solution for long-term storage of images onto synthetic DNA*, IEEE 27-th European Signal Processing Conference (EUSIPCO), 1–5, 2019.

[25] C. Pan, S. M. H. T. Yazdi, S. K. Tabatabaei, A. G. Hernandez, C. Schroeder and O. Milenkovi´c, *Image processing in DNA*, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 8831–8835, 2020.

[26] M. Dimopoulou and M. Antonini, *Image storage in DNA using Vector Quantization*, IEEE 28-th European Signal Processing Conference (EUSIPCO) 516–520, 2021.

[27] C. E. Arcadia, E. Kennedy, J. Geiser, A. Dombroski, K. Oakley, S. L. Chen, L. Sprague, M. Ozmen, J. Sello, P. M. Weber, S. Reda, C. Rose, E. Kim, B. M. Rubenstein and J. K. Rosenstein *Multicomponent molecular memory*, Nature Communications Volume 11, Number: 691, 2020.

[28] J. K. Rosenstein, C. Rose, S. Reda, P. M. Weber, E. Kim, J. Sello, J. Geiser, E. Kennedy, C. Arcadia, A. Dombroski, K. Oakley, S. L. Chen, H. Tann, B. M. Rubenstein, *Principles of information storage in small-molecule mixtures*, IEEE Transactions on NanoBioscience 19, 378–384, 2020.

[29] M. Cheraghchi, R. Gabrys, O. Milenkovi´c, and J. Ribeiro, *Coded trace reconstruction*, IEEE Transactions on Information Theory 66, 6084–6103, 2020.

[30] S. K. Tabatabaei, B. Wang, N. B. M. Athreya, B. Enghiad, A. G. Hernandez, C. J. Fields, J.-P. Leburton, D. Soloveichik, H. Zhao and O. Milenkovi´c, *DNA punch cards for storing data on native DNA sequences via enzymatic nicking*, Nature Communications, Volume 11, Number: 1742, 2020.

[31] K. Chen, J. Zhu, F. Boˇskovi´c and U. F. Keyser *Nanopore-based DNA hard drives for rewritable and secure data storage*, Nano Letters 20, 3754–3760, 2020.

[32] C. Pan, S. K. Tabatabaei, S. M. H. T. Yazdi, A. G. Hernandez, C. M. Schroeder and O. Milenkovi´c, *Rewritable two-dimensional DNA-based data storage with machine learning reconstruction*, Nature Communications 13, 2984, 2022.

[33] C. K. Lim, J. W. Yeoh, A. A. Kunartama, W. S. Yew and C.L. Poh, *A biological camera that captures and stores images directly into DNA*, Nature Communications, 14 (1): 3921, 2023.