DECODING LIFE THROUGH DNA IN IT

Carina CONSTANTINOVA

Technical University of Moldova

Abstract: *Buried deep inside nearly every cell of your body, tiny strands of DNA help determine who you are, how you look and explain in detail how to build and operate a human being. Over the last sixty years, scientists and researchers have learned how to decipher more of human genome. Moreover, diverse encoding models for reading and writing data onto DNA, codes for encrypting data which addresses issues of error generation, and approaches for developing codons and storage styles have been developed over the recent past. DNA has been identified as a potential medium for secret writing, which achieves the way towards DNA cryptography and stenography. DNA utilized as an organic memory device along with big data storage and analytics in DNA has paved the way towards DNA computing for solving computational problems. This article stresses the possibilities and scientific breakthroughs that genomic revolution brought to modern humans and armed them with technological tools and huge knowledge.*

Key words: *DNA-storage, genetics, code of life, cryptography.*

1. Unravelling the mysteries of genome

Human body contains about 100 trillion cells. Cells are building blocks of the human body. In the center of most cells lies the nucleus that is the control center, giving every other part of the cell instructions what to do. Within the nucleus of a human cell are twenty-three chromosomes, together, these hold all of a person's genetic code (information). Chromosomes are made up of very tightly coiled deoxyribonucleic acid. DNA forms an interconnected, spiral known as a double helix. It contains the human genome – all of the data required to make and operate a human being. In the early 1950s, three scientists: Francis Crick, a British biophysicist, James Watson, an American biologist, and Rosalind Franklin, a British chemist, went into adventurous genomic exploration. They produced first X-ray image of a molecule of DNA.

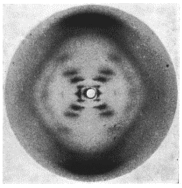

Fig. 1. Rosalind Franklin *photo 51.*

Once scientists knew this much about the genome, their next task was clear. They needed to decode the genome to find out how DNA directs cells to make proteins. There was a problem, though - a huge one. No one - not even Crick, Watson, or Franklin - knew how to determine the order of the nitrogenous bases (A) - adenine, (T) - thymine, (G) - guanine, (C) – cytosine, making up the genome. As scientists continued to study DNA, they realized that certain groups of base pairs work together - genes. The process of decoding genes is known as sequencing. It amounts to listing the sequence, or order, of base pairs in a gene, from start to finish.

The genome contains tens of thousands of genes, and the instructions for building a single protein could be one, 1.000, 100.000, or even 1 million base pairs long. Eventually, the Human Genom Ptoject, headed by the pioneers of DNA, had sequenced the whole human genome that was approximately three billion nucleotide bases long. To type them all, a person would have to work eight hours a day for *fifty years,* typing at a speed of sixty words per minute. The extreme density (and durability) of DNA laid the foundation of new experimental science, which will be described in the next chapters.

2. DNA as a data storage

Humanity is creating information with lambent speed. The "digital universe" grows at an unprecedented rate – about 16 Zettabytes every year (a ZB is the equivalent of one trillion gigabytes).

By 2025, the Global Datasphere will grow from 33 ZB in 2018 to 175 ZB, according to predictions from analyst firm IDC. Where will we store the entire world's data?

The answer is DNA.

Sometimes it is hard to get our minds around such a big number. Here are some illustrations of what 175 ZB represent:

- One zettabyte is equivalent to a trillion gigabytes.
- If you were able to store the entire Global Datasphere on DVDs, then you would have a stack of DVDs that could get you to the moon 23 times or circle Earth 222 times.
- If you could download the entire 2025 Global Datasphere at an average of 25 Mb/s, today's average connection speed across the United States, then it would take one person 1.8 billion years to do it, or if every person in the world could help and never rest, then you could get it done in 81 days[1].

The excursion of data storage initiated from bones, rocks, and paper. Then this journey deviated to punched cards, magnetic tapes, gramophone records, floppies, and so forth. Afterwards with the development of the technology optical discs including CDs, DVDs, Blu-ray discs, and flash drives came into operation. All of these are subjected to decay. Being nonbiodegradable materials, all these objects pollute the environment and release high amounts of heat energy while using energy for operation.

Demand for data storage is growing exponentially, but modern archiving technology is not keeping up with the growing tsunami of bits. Despite improvements in optical discs, storing a zettabyte of data would still take many millions of units, and use significant physical space. If we want to preserve the world's data, we need to seek significant advances in storage density and durability.

The team of enthusiastic researchers from Microsoft Inc. and the Washington University is just one of a number of research groups around the globe pursuing the potential of DNA as a vast digital attic. It has the needed advantages as a storage medium. One selling point is durability – capable of lasting for a very long time if kept in good conditions (DNA from woolly mammoths was recovered several thousand years after they went extinct, for instance) – and will always be current. Another advantage in using DNA to archive data is its attractive feature to be extremely dense (up to about 1 Exabyte per cubic millimeter or a single gram of DNA can hold 215 Petabytes) and durable (half-life of over 500 years).

The biotechnology industry made big advances in both "synthesizing" (encoding) and "sequencing" (decoding) data in recent years. It is a long way to go, however, the scientists are upbeat. They note that their diverse team of computer scientists, computer architects and molecular biologists already has increased storage capacity a thousand times in the last year. Storing digital data on DNA works like this:

First, the data is translated from 1s and 0s into the "letters" of the four nucleotide bases of a DNA strand — A, C, G, T. Reading the data uses a biotech tweak to random access memory (RAM), another concept borrowed from computer science. The team uses polymerase chain reaction (PCR), a technique that molecular biologists use routinely to manipulate DNA, to multiply or "amplify" the strands it wants to recover. Once they have sharply increased the concentration of the desired snippets, they take a sample, sequence or decode the DNA and then run error correction computations. DNA storage requires cutting-edge techniques in data compression and security to design a sequence both info-dense enough to realize DNA's potential and redundant enough to allow robust error checking to improve the accuracy of information retrieved down the line. DNA achieves this in two ways. One, the coding units are very small, less than half a nanometer to a side, where the transistors of a modern, advanced computer storage drive struggle to beat the 10 nanometer mark. Moreover, the increase in storage capacity is not just ten- or a hundred-fold, but thousands-fold. That differential arises from the second big advantage of DNA: it has no problem packing three-dimensionally.

Nick Goldman[2] and his EBI colleague Ewan Birney took the idea back to their labs, in 2013 announced that they had successfully used DNA to encode five files, including Shakespeare's sonnets and a snippet of Martin Luther King's 'I have a dream' speech. By then, biologist George Church and his team at Harvard University in Cambridge, Massachusetts, had unveiled an independent demonstration of DNA encoding. An improved system was reported in the journal *Nature* in January 2013, in an article led by researchers from the European Bioinformatics Institute (EBI) and submitted at around the same time as the paper of Church and colleagues. Over five million bits of data, were stored, retrieved, and reproduced. All the DNA files reproduced the information between 99.99% and 100% accuracy.[3]

3. Public key cryptography of binary data stored as DNA

As fast as the digital information grows that much this powerful and wanted good at any time needs protection. In order to protect information, cipher writing was used since ancient times and is used as well nowadays. Famous and widely applied techniques, which implement secret writing, are cryptography and steganography. These two sciences manipulate information in order to cipher or hide its main sense. Why do we have such an interest for DNA cryptography? It is a newborn cryptographic field based on the research of

DNA computing. This chapter investigates a variety of bio-informatic methods and proposes two different algorithms for encrypting and decrypting data stored in real or artificial DNA digital form.

The plaintext message is encrypted with RSA public key algorithm. The security of this algorithm is given by the computational difficulty of factoring large numbers. To be secure, very large numbers must be used as primes, 100 decimal digits at the very least. Product of such large prime numbers is an easy mathematical operation, but reverse process is a very hard task. It is extremely difficult, nearly impossible, to determine the original values from the product, at least it will take a lot of time.

This algorithm offers the public key (n, e) for encryption and the private key (n, d) for decryption (d is secret); n is the product of two primes, while e and d mathematically derive from n:

n=p*q (p and q are prime numbers:
they have only two divisors, 1 and itself);
φ (n) = (p-1) (q-1) is Euler's totient;
e coprime to φ(n);
d*e mod φ (n) = 1;
C = Pe mod n (encryption);
P = Cd mod n (decryption);
where P is the plaintext and C the cipher text.

The encrypted message with RSA is a set of numerical values. These numbers will be converted using substitution in artificial DNA strand. All resulted pieces of DNA strands are bind together using a special ligase protein and the complementary strand as a template. The encrypted message can be transmitted in a compact form on DNA chip.

Algorithm steps:

Step 1: Binary data, text or image, is visualized like ASCII cod or brightness levels.

For example original message: "my secret!" in ASCII will be:

109 121 32 115 101 99 114 101 116 33.

Step 2: This numeric values are arranged in a string and taken by several digits at once, number of digits rise together with the public keys length. In this example, we take seven digits at once and obtain:

1091213 2115101 9911410 111633.

Step 3: These numbers, seven digits long will be encrypted with public key (public key will be relatively short in order to make the example easer to follow) and the result is another set of numbers: 417310496328959; 129126952185213; 373906236380070; 367568882589235.

Step 4: Encrypted sequence is transformed in binary form:

417310496328959 → 01011110111000101010101011111001010000110011111 11.

Step 5: Binary sequence using substitution is transformed in DNA sequence:

A – 00
C – 01
G – 10
T – 11

0101111011100010101010101111100101000011001111111→
CCTGTGAGGGGGTTGCCAATATTTT

Step 6: All sequences are bind together in a single strand, the cipher text:
CCTGTGAGGGGGTTGCCAATATTTTCTCCCTAAG
TCGACTCGGTCCTTCCTCCCTAAGTCGACTCGGTCC
TTCCCCCAACAATCCACGCGACATGGCGCCCCAAC
AATCCACGCGACATGGCGCCATGCATCCATAGGTC CCTGATAT.

Decryption is a reverse process: the DNA strand is cleaved in original pieces using restriction enzymes and transformed in numerical values using the same substitution as for encryption. Using the private RSA key finishes the last step of decryption.

Conclusions

As long as life on the Earth is DNA-based, humanity will be interested in reading it. It is clear that data storage in DNA is not a part of science fiction anymore; it is being realized and improved at very promising rates by research teams all over the world.

Similar to all revolutions in technology, DNA-based data storage technology has to face major challenges to realize its full potential. A simple fact that 4 grams of deoxyribonucleic acid can hold inside itself the entire world's data created during an year. It is inevitable that DNA would be invariably used for archival purposes for its sheer density, robustness, stability and energy efficiency. In theory, grams of DNA can store all the information ever produced by mankind. Several breakthroughs will be required before it becomes commercially mainstream for data retrieval.

This field has had a million-fold improvement in the recent years. Digital Data Storage in DNA technology. Bioinformatics. Future work could include compression schemes; dealing with redundancy at all levels, checking for parity, correcting errors to enhance density and safety. DNA could also be substituted with polymers or be modified to suit the needs of digital storage. Furthermore, it will fuel research to look for alternative materials for information storage and to aid in realizing the need for a universal medium for data.

To top it off, this technology is here to transform the way we have ever looked at DNA and computing as different entities.

References

1. David Reinsel, John Gantz, John Rydning, *The Digitization of the World From Edge to Core.* International Data Corporation, November 2018, p. 5-10.
2. Nick Goldman, the leader of the research team studying molecular genome evolution for the European Bioinformatics Institute, in Cambridge, United Kingdom. His team created the modern field of "DNA-storage", the use of DNA to archive digital information.
3. S. Vlad, R.V. Ciupa, A.I. Nicu, *DNA Cryptographic Algorithms,* MEDITECH 2009, IFMBE Proceedings 26, p. 223–226.
4. Ron Fridell, *Unravelling The Mysteries of The Genome,* 2005, p. 7-99.
5. www.nature.com