How to Store Wikipedia into a Forest Tree: Initial Idea

Karin Ljubič*, Iztok Fister Jr**

*University of Maribor, Faculty of Medicine, Taborska 8, 2000 Maribor, Slovenia
**University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, 2000 Maribor, Slovenia

Abstract: DNA represents a possibility of highly compact and extremely resistant storage of information in the future. DNA is composed of two complementary strands. These complementary strands are built of basic building block called bases. These are adenine, guanine, cytosine and thymine. In the last year, this type of encoding information gained with the help of nanotechnology. With the help of artificial sequencing of bases in different sequences, the DNA can encrypt the digital information that is recorded in a computer readable form-in a form of bits. 1 bit is a basic unit of information. With last year´s successes in this field began the expansion of this field. Future promises even commercially available methods for storing and reading information. A significant reduction of costs of this type of coding is needed. Prospects are promising, because the prices are falling exponentially. In this article we discuss about using DNA as information storage and systematically present the first successes in this field. We propose new ideas for storing data, which are based on currently developed technologies and out-of-the-box thinking.

1. INTRODUCTION - DNA DATA STORAGE

DNA (Deoxyribonucleic Acid) is a molecule with genetic instructions used for development of a living organism. It represents a possibility of highly compact and extremely resistant storage of information in the future [1]. DNA is composed of two complementary strands. DNA molecules are double-stranded helices. Each strand consists of units called nucleotides-each nucleotide contains a nucleobase (guanine,

adenine, thymine or cytosine). We record these nucleobases by using the letters G, A, T and C. Each type of nucleobase on one strand bonds with just one type of nucleobase on the other strand. This principle is called complementary base pairing, with adenine bonding only to thymine and cytosine bonding only to guanine. This is very important for coding information because each of the two strands contains identical information. The two nucleotides binding together across the double helix are called a base pair. A backbone of each strand is made of alternating sugars and phosphate groups. Both strands of the double-stranded structure store the same biological information. Therefore, DNA is well-suited for biological information storage and this information is replicated as the both strands are separated as we explain in the following paragraphs. It is important to emphasize that the significant portion of DNA is non-coding. In humans, for example, more than 90% of DNA does not serve a function of encoding proteins. The situation is similar in other living organisms and therefore this non-coding part of DNA represents a possibility of storing information which is not connected to the development of the organism [2].

DNA nanotechnology deals with designing and manufacturing of artificial DNA for technological use [3]. Today synthetic DNA can be produced with the use of commercially available apparatus for synthesis of nucleotides. Decoding of information written in DNA is called DNA sequencing. It determines the precise order of nucleotides within a DNA molecule. The advent of rapid DNA sequencing methods has greatly accelerated the medical and biological research and discovery [4]. There are many methods for DNA sequencing, some of them will be presented in this article. Storing information into DNA is more compact than any currently available magnetic tapes or hard drives. It also offers the possibility of extremely long storage of information, as long as it is stored in cool, dark and dry conditions. This can be solved by storing into DNA sequence of a living plant though this brings some ethical concerns.

Despite all the options, DNA storage is yet relatively expensive and slow process. The major drawback is a current cost of synthesizing DNA in the quantities required. Estimated price per megabyte of data stored in specialized laboratories is US$12,400. Another issue is the cost of decoding the information stored in the DNA, estimated at about US$220 per megabyte. If the costs of synthesizing DNA could be reduced by

one or two orders of magnitude, personal DNA archives could become feasible. According to current trends, this might occur within a decade.

In this article we made a step forward in presenting a new idea of storing information. In the first section we present the work that has already been done in this field. Then we describe a basic principle of storing information into DNA and coding schemes. Based on all facts, we present a new idea of a Wiki tree, pros and cons and ethical concerns.

2. FIRST SUCCESS

George McDonald Church, American chemist, molecular engineer and geneticist, in collaboration with colleagues made an experiment in the middle of 2012. They converted a HTML file containing 53.400 words 11 images in JPG format and JavaScript program (multiple copies of all these documents) into a DNA sequence. Thus, in 1 mm^3 of DNA they saved 5,5 petabytes of information. They used a simple coding system, one bit corresponding to one nucleobase. Adenine or cytosine is represented with 0, guanine or thymine is represented with 1. This experiment is presented in the article Towards the Next Generation Digital Information Storage in DNA which was published in the journal Science. One disadvantage of this type of coding are long stretches of the same nucleotide and this might lead to several mistakes in decoding process [5].

In January 2013 Nick Goldman and his colleagues from European Bioinformatics Institute introduced a new system of coding. The information was decoded without mistakes. They used a more complex coding pattern. 1 byte, made of 8 bits, was encoded in the word of five letters using A, G, C, T. To further reduce the possibility of errors in decoding, the DNA was cut into sections 117 characters (bases) long. They stored more than 5 million bits of data into a DNA sequence. The synthetically made helix stored all 154 of Shakespeare´s sonnets and a 26 seconds long section of the famous Martin Luther King speech entitled “I Have a Dream”. In addition they added a copy of Watson and Crick reports on the structure of DNA, photography of the institute and a file with the description of this experiment. Recipients of encrypted information read the message with 100% accuracy [6, 7].

3. THE PRINCIPLE OF STORING INFORMATION INTO DNA

The principle of storing data into DNA is theoretically simple. All information stored in computer is in the form of bits, consisting of zeros and ones. The computer program then parses the code in the letters A, C, G, T. As mentioned. We can use different coding techniques. The most accurate is the one already mentioned in the previous section. One byte is coded with a word of five letters using A, G, C, T. That way the artificial DNA is created and sent to recipient. With the help of a sequencing machine then the information is decoded and converted back into zeros and ones. For easier understanding the process is presented in picture 1.

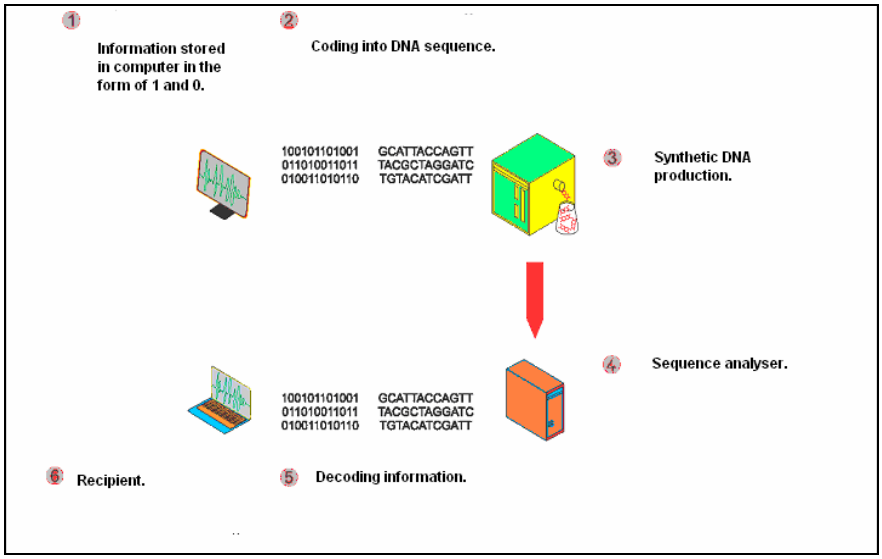

Fig. 1:The process of coding digital information in the form of ones and zeros into DNA. After determination of the DNA sequence, synthetic DNA is produced and sent to a recipient. After the sequence analysis, the information is prepared for decoding. Decoded information can be presented to the recipient.

4. STORING INFORMATION INTO A LIVING CELL

So far, the data has not been stored into a living cell. A seed is an embryo with two points of growth, one of which forms the root and other the stem. An embryo is enclosed in a seed coat with some food reserves. The embryo is formed from the zygote. So the seed is composed of the embryo (the result of fertilization) and tissue from the mother plant, which also forms a cone around the seed in coniferous plants. The information should be stored into the DNA of a zygote, since the zygote represents the first cell of a newly formed plant. Each following cell would afterwards contain identical genetic information as the first cell. The procedure could be carried out by using genetic engineering techniques. The result is genetically modified organism/plant. Genetically engineered plants are generated in a laboratory. Mostly the genetic material of a cell is modified by using biolistic method (particle gun) [8]. A gene gun or a biolistic particle delivery system is a device for injecting cells with genetic information. The device is able to transform almost any type of cells. The target of a gene gun is a callus of undifferentiated plant cells growing on gel medium in petri dish. By using the gene gun the injected DNA eventually migrates to and integrates into a plant chromosome. Selected single cells from the callus are then treated with a series of plant hormones, such as auxins and gibberellins. Each cell may divide and differentiate into the organized, specialized, tissue cells of an entire plant. Totipotency is a term that explains the capability of total re-generation. Therefore the new plant originated from a successfully shot cell may have new genetic and heritable traits. There are also other less frequent methods of inserting new genetic material into the living cell. One of them is with the use of Agrobacterium tumefaciens and its Ti plasmid. We can see that there are different possibilities of genetic transformation and a number of methods are available to transfer DNA into a plant cells.

5. GROWING A WIKI TREE

For easier explanation, we chose Wikipedia. Wikipedia is a collaboratively edited, free internet encyclopedia. There are over 4,4 million articles in the English

Wikipedia and the estimated size is 42 gigabytes. Comparing to 90 petabytes of information currently stored in CERN, the European particle-physics lab near Geneva, which could be compressed into 41grams of DNA, the Wikipedia seem like a small task. The stored information should be indexed by a specific initial and ending sequence of bases for the sequencing machine to recognize it. It is important to emphasize that all tissues of this tree would contain the same information. The process by which a cell duplicates itself into two daughter cells is called mitosis. Each leaf would be carrying the whole encoded information.

Imagine having a whole Wikipedia stored in a tree in a park. It would be possible to have all the data of this world stored in a one information self-sustaining forest. The biggest problem of this idea is probably decoding of information. Imagine picking a leaf from a tree in order to immediately get information. There are many decoding strategies available today and it is a question of time when the technology will allow small portable computers reading information from a tree leaf. According to current sequencing techniques this might not be only an imaginary situation. There are several developed techniques for DNA sequencing, some of them are under development. We shall present some innovative ideas that could evolve next-generation information browsing. The high demand for low-cost sequencing has driven the development of next-generation sequencing technologies that parallelize the sequencing process [9].Methods already used are: Massively parallel signature sequencing (MPSS), Polony sequencing (PS), 454 pyrosequencing, Illumina (Solexa) sequencing, Solid sequencing, Ion Torrent Semiconductor Sequencincg, DNA Nanoball Sequencing. Heliscope single molecule sequencing and Single molecule real time sequencing. We present some new solutions that will be cost-effective, portable and extremely faster that today´s methods. One of them is Nanopore DNA sequencing. This method is based on electrical signals occurring when nucleotides pass through pores. The DNA passing through the nanopore changes its ion current. The pore contains a detection region capable of recognizing different bases, with each base generating various time specific signals corresponding to the sequence of bases as they cross the pore. Another approach uses measurements of the electrical tunneling currents across single-strand DNA as it moves through a channel. Each base affects the tunneling current differently, allowing the differentiation among different

bases. This method is called Tunnelling currents DNA sequencing. Another possible innovative methods are Microfluidic Sanger Sequencing, RNAP sequencing and others. It is a matter of time, when each one of us will possess a small portable DNA sequencer [10].

6. COMPUTATIONAL CHALLENGES

It is important to evaluate the raw sequence data from sequencing technologies described in the previous section. So far, the evaluation is done mostly by algorithms such as Phred and Phrap. Phrap is a widely used program for a DNA sequence assembly. A detailed description of the Phrap algorithm can be found in the Phrap documentation. It is a challenge to deal with repetitive sequences that can occur in many places in DNA when coding information into a base sequence. For more detailed bioinformatical analysis, new methods for correcting errors have been developed. Good example are trimming programs. However, this is an open filed for research and optimization [11].

7. CONCLUSION AND ETHICAL CONCERNS

DNA is a God´s masterpiece. Interference in its structure faces our society with many questions. However, at the current social state of mind, sincerely, our society is not ready for the advent of enormous possibilities of data storage and data browsing. Imaging that on every step there is a possibility of accessing information is exciting but also raises questions about exploitation. As already mentioned, our society should slowly grow up to immense technological possibilities which will become reality in the future. Not to mention the ecological aspects of that type of storing information. There are many pros and also many drawbacks and we can only raise our next-generation in a spirit that will not allow tempting drawbacks for exploitation.

REFERENCES:

1. Mashaghi A, Katan A (2013). "A physicist's view of DNA". *De Physicus* 24e (3): 59–61.
2. Alberts, Bruce; Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts and Peter Walters (2002). *Molecular Biology of the Cell; Fourth Edition*. New York and London: Garland Science.
3. Seeman, Nadrian C. (June 2004). "Nanotechnology and the double helix". *Scientific American* 290 (6): 64–75.
4. Seeman, Nadrian C. (2010). "Nanomaterials based on DNA". *Annual Review of Biochemistry* 79: 65–87.
5. Church, G. M.; Gao, Y.; Kosuri, S. (2012). "Next-Generation Digital Information Storage in DNA". *Science* 337 (6102): 1628
6. Goldman, N.; Bertone, P.; Chen, S.; Dessimoz, C.; Leproust, E. M.; Sipos, B.; Birney, E. (2013). "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA". *Nature* 494 (7435): 77–80.
7. Yong, E. (2013). "Synthetic double-helix faithfully stores Shakespeare's sonnets". *Nature*.
8. Klein, TM et al (1987) High-velocity microprojectiles for delivering nucleic acids into living cells. Nature 327:70-73.
9. Ten Bosch, J. R.; Grody, W. W. (2008). "Keeping Up with the Next Generation". *The Journal of Molecular Diagnostics* 10 (6): 484–492
10. Liu, Lin; Li, Yinhu; Li, Siliang et al. (1 January 2012). "Comparison of Next-Generation Sequencing Systems". *Journal of Biomedicine and Biotechnology* (Hindawi Publishing Corporation) 2012: 1–11.
11. Del Fabbro C, Scalabrin S, Morgante M and Giorgi FM (2013). "An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis". *PLoS ONE* 8 (12).