Turkish Journal of Science & Technology
Volume 12(1), 25-32, 2017

A Schematic Overview of Securing Precision Medicine Data with a DNA Database System

Shannon CONLEY, Mehmet KAYA, Asaf VAROL
Department of Software Engineering,Firat University Elazig, Turkey.
Electrical and Electronics Engineering Dept.Adiyaman UniversityAdiyaman, Turkey
Department of Software Engineering ,Firat University, Elazig, Turkey
shconley@syr.edu

(Received: 14.07.2016; Accepted: 01.03.2017)

Abstrack

This work explores the security challenges facing the implementation of the National Institute of Health's proposed Precision Medicine Initiative® Cohort Program. The program is planning to recruit over a million participants and collect their health data over an extended period of time. Data will be made available to researchers to drive the development of more "precise preventive care and medical treatment". However, the ability of the program to keep participants' personal health information private is being scrutinized and might deter volunteer recruitment. This is due in large part to recent security breaches experienced by the US federal government in which over 20 million people's security clearance data was stolen including credit card numbers and fingerprints. To secure the personal information of the volunteers, we propose the use of a natural computing nature-inspired hardware technique: molecular data storage. In this scheme, personal patient identifiers would be encoded and stored in DNA molecules, thus protected behind a "biological firewall". The DNA database would only be accessed in the same laboratory environment(s) used to collect/process patient health samples and data would be retrieved using enzyme-based biological reactions. An anonymous volunteer id randomly assigned to each volunteer would be stored both in the DNA database and a traditional network connected database(s) used by program approved researchers and serve as the cross-reference key between databases. By adopting the technique, the new initiative could not only revolutionize the way medicine is practiced, but also the field of data storage and security

Keywords: Natural Computing; Data Storage in DNA; Precision Medicine.

Hassas Tıp Verilerini Bir DNA Veritabanı Sistemiyle Güvenceye Almaya Şematik Bir Bakış

Özet

Bu çalışma Ulusal Sağlık Enstitüsü tarafından önerilen Hassas Tıp Projesi'nin (Precision Medicine Initiative) uygulanmasında karşılaşılabilecek güvenlik sorunlarını araştırmaktadır. Bu projede bir milyondan fazla katılımcıya ulaşıp belli bir zaman sürecinde bu katılımcıların sağlık verilerinin toplanması planlanmaktadır. Bu veriler daha sonra "daha hassas önleyici bakım ve tıbbi müdahale" teknikleri geliştirmek amacı ile araştırmacın erişimine sunulacaktır. Ancak bu proje katılımcıların kişisel ve özel sağlık bilgilerinin gerektiği şekilde korunması yönünden eleştirilmektedir ve insanlar kişisel bilgilerinin güvenliğinden şüphe ederek programa katılımda gönülsüz kalabilirler. Birleşik Devletler Federal Hükümetinin yakın zamanda tecrübe ettiği ve 20 milyondan fazla insanın kredi kartı numaraları ve parmak izi gibi güvenlik bilgilerinin çalındığı bir güvenlik sorunu da aslında kısmen bu duruma neden olmaktadır. Bu çalışmada sunulan şemada, kişisel hasta numaralarının kodlanıp DNA molekülleri içinde saklanarak bir "biyolojik güvenlik duvarı" ile korunması öngörülmektedir. Bu DNA veri tabanı sadece hasta örneklerinin toplanıp işlendiği laboratuvarda erişilebilecek ve veriler sadece enzim-bazlı biyolojik reaksiyonlar vasıtası ile tekrar elde edilebilecektir. Her katılımcıya rastgele atanan anonim bir katılımcı numarası hem DNA veri tabanında hem de programda akredite edilmiş araştırmacıların kullandığı ve veri tabanları arasında çapraz başvuru sağlayan geleneksel elektronik bir veri tabanında saklanacaktır. Bu tekniğin kullanılması ile, bu hassas tıp projesi sadece tıp uygulamalarında değil, veri depolama ve güvenlik alanlarında da çok büyük yenilikler başarabilir.

Anahtar Kelimeler: Doğal Kodlama; DNA içerisinde Veri Saklama; Hassas Tıp.

1. Indroduction

President Obama has proposed a new initiative to collect health information from over a million volunteers and build a data repository that will drive translational research geared towards the creation/implementation of precision medicine based disease prevention measures and treatments [1]. "As an emerging approach for disease prevention and treatment that takes into account people's individual variations in genes, environment, and lifestyle, [2]" the precision medicine model includes the following desired research outcomes outlined in the Precision Medicine Initiative® Cohort Program's proposal:

- Create disease risk scoring systems based on the interplay between environmental and genetic factors [1].
- Enable researcher to create drugs and doctors to customize drug prescriptions based on a patient's genetic and/or environmental makeup [2]. (a.k.a. Pharmacogenomics)
- Develop new diagnostic tools using discoveries of "biological markers that signal increased or decreased risk" of diseases [1].
- Deploy mobile health (mHealth) technologies "to correlate activity, physiological measures and environmental exposures with health outcomes [1]" and determine if they are effective in improving the user's health/promoting healthy lifestyle changes [2].

As described in the initiative's proposal, for researchers to discover and confirm the associations needed to propel the above goals, they need the diversity of data and statistical power that such a large scale study would provide. Assuming that Congress will approve the program and allocate the necessary funding in 2016, the program will need to attract and retain one million US volunteers from diverse backgrounds. Otherwise, as the current director of the National Institute of Health (NIH), Dr. Francis Collins, has stated: "There are two ways this could go wrong with the creation of this cohort. One would be that no one signs up; the other would be that everybody in the country signs up. We want a Goldilocks opportunity but that may be difficult to dial in [3]."

Unfortunately, the risk of data compromisation might deter many from participating and prevent the initiative from proceeding. Although the program will remove personal identifiers from the data given to researchers [4], it might still be possible to identify the person behind a record. For instance, if DNA information is provided as a field in the record, it could be matched to DNA information in a police or insurance database. This type of risk can be mitigated by restricting access to the raw records to only trusted research groups. However, the threat of database attack/hack is a more formidable issue. Between 2014-2015, the federal government's human resource department, the Office of Personnel Management (OPM), was hacked and the security clearance data "including fingerprints, Social Security numbers, addresses, employment history, and financial records" of around 21.5 million people were stolen [5]. The question is being raised that if the government cannot protect security clearance information, how can it secure a volunteer's confidential health data? Dr. Collin's has responded that, "in this current climate," the program cannot guarantee that a security breach will not occur.

If these records are leaked, they could cause major financial repercussions for a volunteer. For instance, a company might not hire a candidate if he or she has a pre-existing condition/is at risk for a disease that would be difficult to insure or if he or she has a history of drug or alcohol addiction. Although such hiring discrimination is illegal, it would be difficult for the job candidate to prove. Furthermore, it is not illegal for other types of organizations/businesses such as life insurance companies to take such information into consideration [4]. Even for healthy volunteers, it is hard to anticipate what the study might reveal about their possible risk for a disease-thus, making an accurate assessment regarding the risk of participating in program where such information could be revealed/leaked extremely difficult. Rather than accept the possibility of a security breach, the initiative could explore and pioneer an alternative approach to storing data digitally that is accessed via the Internet or other type of computer network. This paper will explore the

possibility of securing volunteer data via natural computing data storage solutions.

2. Overview of Natural Computing

Natural Computing or natural computation has been defined as the "process of extracting ideas from nature to develop 'artificial' (computational) systems, or using natural made media (e.g. molecules) to perform computation. [6]" And, this field is typically subdivided into three branches or areas of research as shown in Fig. 1.6 [6, 7].

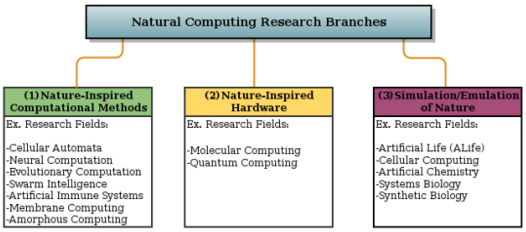

Figure 1 Three major branches natural computing research and popular sub-fields within each branch

A. Nature-inspired computational methods

This branch is considered to be the oldest branch of natural computation and focuses on the application of algorithms inspired by natural mechanisms such as "natural selection" to find acceptable solutions to optimization problems where there is not enough information or computational power to find the optimal solution [7, 8]. Several nature-inspired meta-heuristic algorithms are briefly described as follows:

1) Artificial Neural Networks (ANNs) [7,9]

As neural computation technique, ANNs are modeled after connections between neurons or nerve cells known as neural networks. Each connection between nodes or "artificial neurons" is assigned a value that is repeatedly adjusted until the difference between the desired and actual outputs is sufficiently lowered. This process, referred to as back-propagation, allows the algorithm to "learn" (supervised learning method) and makes it suitable for complex nature-like tasks such as "computer vision and speech recognition."

2)Genetic Algorithm (GA)[7,10]

Falling under the evolutionary computation category, GAs use the process similar to the concept of "survival of the fittest" to choose which candidate solutions should be remain in the solution pool (survivor selection) and which candidates should be selected to create new candidate solutions (reproduce) based on fitness value produced by the optimization problem's objective function. Candidate solutions must be represented "genetically" typically with a fixed-length binary array. During reproduction, two or more candidate solutions are aligned and sections are from each are combined to produce children, thus imitating the process of genetic recombination. Genetic mutation-like process can be achieved by randomly flipping bits in a candidate solution. GAs tend to be easy to implement and widely applicable, but have a hard time finding the optimal solution.

3)Particle Swarm Optimization (PSO)[7,11]

As a swarm intelligence approach based on the interaction or movement of animal groups such as an insect swarm or flock of birds, PSO is used to find solutions in a multidimensional search space by modeling "particles" as possible solutions. Particles are iteratively repositioned in the search space in an attempt to "improve upon" their solution. Repositioning is based on the particle's velocity, its all-time best position, and the best positions of its neighbors. Like GA, after a certain number of iterations candidate solutions are expected to converge or reach a single solution. (This could be a local optimum as opposed to the global optimal or best solution.)

4) Dendritic Cell Algorithm (DCA)[7,12]

Classified as an artificial immune system (AIS) algorithm, DCA is based on cells that trigger the immune system to release antibodies by detecting and presenting a foreign molecule or antigen. (Dendritic cells are sometimes nicknamed "crime scene investigators.") All dendritic cells start out in the immature stage and process safe, danger, and pathogen associated molecular signals (PAMPs). If the cell receives mostly safe signals, then the antigens it has gathered will be considered safe and it will enter the semi-mature stage. Otherwise, if the other two types are in the majority, the antigens will be

considered "harmful" and the cell will transition into the mature stage. The algorithm has been successfully used to detect anomalies and in such applications as bot detection on a compromised host machine where the three signal types were derived from statistics associated with possibly suspicious API function calls such as the rate of change between a key logging call and antigens were defined as the process associated with the suspect function call**.**

B. Nature inspired hardware

As alternatives to electronic hardware, the "natural" hardware alternatives discussed below have the potential to solve security and storage issues that plague the current electronic paradigm. However, since quantum computing is much less developed (still in its "infancy") than molecular computing, the rest of this paper will focus on the exploration of Biocomputing techniques to securely store confidential health records [7, 13]. The two approaches are described as follows:

1)Molecular Computing [7,14]

In this computing model, data is stored/encoded into biomolecules such as deoxyribonucleic acid (DNA) molecules. (DNA molecules are composed of four different types of building blocks or nucleobases: Adenine (A), Cytosine (C), Guanine (C), and Thymine (T), which can be paralleled to a base-4 or quaternary numeral system. Additionally, A-T and C-G can hydrogen bond to create the base pairs responsible for structures such as DNA's formation of a double helix.) And, computational operations such as arithmetic and logic are carried out using "standard protocols and enzymes" [7, 14]. In a "proof-of-concept" study published in 1994, researchers demonstrated the computational possibility of working with molecules by encoding a directed graph in DNA (Fig. 2) [14] to provide a simple illustration/introduction to molecular computing; the study is described in detail below.

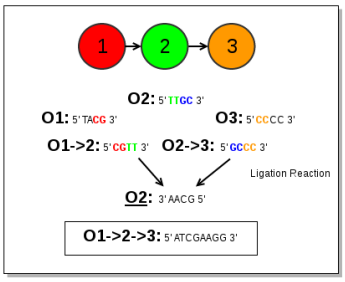

Figure 2 DNA Encoded Directed Graph

This is a demonstrative example of how a directed graph is represented using DNA. The three vertices O1 O2 O3 are represented by random 4-omer oligonucleotides and the last two nucleobases of O1 and the first two nucleobases of O2 are combined to form the directed edge O1->2. (For edges containing the input and output vertices, they were the exact copy of the input or output vertex.) The same logic is used for edge O2->3. To create possible graphical paths, the complement of a vertex (O2) is used to bring together two edges via base pairing so that the ligation enzyme can create a covalent bond between the 3' end of O1->2 and 5' end of the O2->3 sequences. (DNA strands are directional with a 3' atom sugar backbone and hydroxyl group and 5' sugar backbone with a phosphate group.) And then they used laboratory procedures to solve the directed Hamiltonian graph problem as outlined below [14].

Step 1: To generate random paths through the graph, the researcher used a ligation reaction and vertices complementary oligonucleotides to join fragments together.

Step 2: To select only paths that begin with the input vertex and end with the output vertex, polymerase chain reaction (PCR), which is used to generate many copies of sequence within the sequence, where the two primers- to indicate where replication should occur- were the sequence of the input vertex and the complementary sequence of the output vertex.

Step 3: Since a graph is Hamiltonian when there is a path that enters each vertex aside from the input and output only once, only paths that enter the number of vertices within the graph are kept by selecting sequences with a length of n*base size using agarose gel separation.

Step 4: To keep only sequences where every inner vertex has been entered, the complement of the first inner vertex was used to attract/separate out only those sequences containing that vertex and repeated for each inner vertex.

To solve the problem, the study required a week's work of laboratory work/manual labor [14], an issue that still persists today. Regardless, many advances have been made in this discipline over the past twenty years especially in relation to data encoding and both advances and current limitations/research opportunities will be discussed later in this paper.

2) Quantum Computing [7,13,15]

In this model, data is stored in quantum bits or qubits. Unlike a classical bit, one qubit can be used to encode/transmit two classical bits via superdense coding and the principle of superposition where the addition of two or more quantum states produces a new valid quantum state. It is possible for n-qubit system to represent up to 2n different states simultaneously whereas an n-bit can represent only one of these states at a given point in time. Theoretically, this property will allow computations with integers too large to be stored/manipulated by traditional computers. For instance, integer factorization capabilities needed to break public key cryptographic system could be made possible. (The US government is currently funding projects to create a quantum computer to break encryption.) And, although quantum computing has the ability to create security vulnerabilities, another quantum phenomenon known as entanglement could provide an alternative to mathematical complexity-based cryptography. Since reading/measuring a quantum state alters or breaks it, encoded data cannot be surreptitiously copied or read such as public key exchange message without the exchangers detecting the attack.

C. Simulation/emulation of nature

Unlike the first two branches of natural computing research, the third is focused on studying the natural world by modeling it as traditional computational system [7]. (The opposite of creating new computational methods and techniques from nature-inspired models.) This type of research is outside of the scope of this study/not directly applicable to the research question at and will not be elaborated on further. However, it is noted that the results of research regarding a more thorough understanding of natural systems is likely to advance/create new nature-inspired computing techniques.

3. Molecular Data Storage Using DNA

In the field of molecular computing, much focus has been placed on the use of DNA molecules as storage medium for the following reasons [16, 17].

High storage density-DNA can store an enormous amount of information in an extremely small amount of space, making it superior to current electronic storage systems. It is approximated that one gram of single-stranded DNA can hold up to 455 Exabytes. (As a reference, a single Exabyte could hold 500 to 3000 copies of the entire collection of materials stored in the US's Library of Congress [18].)

- Proven longevity-DNA has been preserved for up to 60,000 years in certain conditions and has a projected lifespan or longevity of around 1 million years with specific set of storage conditions [19]. In contrast, today's hard disks and SSDs have around a ten year life span. Additionally, by integrating the sequence into an organism's genome, the information could be passed down from generation to generation as long as the species survived and the information was not mutated. (Currently, there is a research project to integrate an encoded poem into the genome of D. radiodurans, a radiation-proof bacterium [20].)
- Processing tools included - Unlike synthetic macromolecule storage alternatives, DNA storage comes with a natural set of reading, writing, and repair mechanisms that can be manipulated by scientists to perform desired computational operations such as solving the Hamiltonian graph problem in [14].
- Multi-layered encoding - DNA can form 3-dimensional or tertiary structures and the

repeating structural patterns or motifs can add an additional encoding layer. (A database column could be physically as opposed to representationally.)

- Heightened data security - As an off-line system, it is immune to network-based attacks. (However, traditional computer-based tools to encode/decode the data could easily introduce such vulnerabilities [21].)

5. DNA Data Storage Current Capabilities and Limitations

In a 2012 landmark study, researchers successfully stored and retrieved 5.27 megabits (658.75 kB) of data [16], one 53,426 word book, 11 JPG images, and a JavaScript program using inkjet printed DNA microchips. Prior to the study, researchers had a difficult time working with large DNA molecules and the largest published amount of data encoded into DNA was 7920 bits [16]. To address this issue, the study used thousands of small DNA sequences (oligonucleotides) that were given 19 bit addresses and a 96 bit storage capacity. And, the study was able to reduce the cost of encoding/decoding by approximately 100,000 times with only 10 bits of error out of 5.27 million bits [16]. Since this study, items such as all of Shakespeare's sonnets have been encoded and decoded with 100% accuracy and further reduction in cost [21, 22]. It took $12,400/MB to write the data and $220/MB to read the data, but the study also suggested that with the current pace of advancement in this field the approach could become cost-efficient in less than ten years. (The current scientific cost is projected to be around $500/MB to write the data [23].) Due to the high read/write cost and the slow read/write speed, data storage in DNA is currently being advocated/proposed as long term storage solution. However, with the right investment/application opportunity ("necessity is the mother invention"), a DNA storage solution that could behave more like/serve as a database management system (DBMS), as opposed to an archival approach, might be possible. As pointed out in [21], the development of read/write technologies especially suited for DNA storage as opposed to the use of traditional lab techniques could bring down the cost, while chemical processes could be developed/designed specifically to increase the read/write speed. In the following section, we will identify an opportunity where a DNA storage database solution could be developed and adopted as a more secure alternative to a network connected database and propose a scheme for how such a system would work.

6. DNA-Based Storage Approach to Secure the Personal Information of Volunteers in the Precision Medicine Initiative Program

As discussed in the introduction, the success of the precision medicine initiative relies on the successful recruitment of the target number of volunteers. And, with recent hacking incidents where millions of citizen's private data was stolen from governmental databases, the security risk involved in participation is a major challenge facing recruitment. The program has promised to strip personal information from data provided to researchers. However, personal information will be kept "behind a firewall," [4] and thus still vulnerable to network based computing attacks. The use of a "DNA database" to instead store personal information "behind a biological firewall might assuage potential participants fears," while still using a traditional database to store the anonymised health data would be more affordable and minimize experimental uncertainties/maintain a low learning curve. And, since volunteer data such as urine or blood will be collected and processed in a "clinical/laboratory environment," the setting would already be conducive to the use of DNA storage system

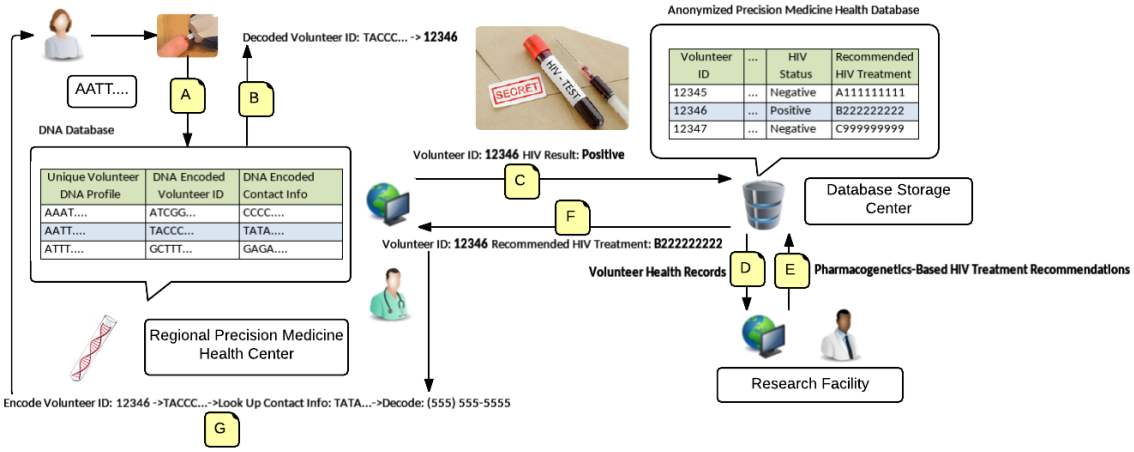

Figure 3 DNA Database Overview for Precision Medicine

And, processing times associated with sample interpretations would mean that the time associated with reading/accessing DNA database would not necessarily be the rate limiting step/**G.** disproportionately time-inefficient. Figure 3 illustrates how such a system might operate.

This provides a schematic overview of how a DNA database system could protect volunteer identity. The following steps are described as follows:

A. Volunteer (female figure in top left corner) visits a regional precision medicine health center to receive an HIV test. A sample of her blood is drawn and DNA profiling is performed. Using a sequence(s) unique to her as primer, PCR is performed to extract/read only her encoded DNA sequence from the DNA database.
B. Either a precision medicine employee or automated tool would then decode the DNA to determine her volunteer id.
C. Once the HIV test results have been returned, the volunteer id will be used to record the result into the precision medicine database.
D. During this time, a research group is using the health data to perform a pharmacogenetics study to tailor HIV treatments according to a particular subset of genetic variations.
E. The study is successful and the research group updates the database to recommend particular treatments for each patient.
F. For each volunteer id with a positive HIV status, every regional center receives a push notification along with ids and recommended treatment information. Instead of a sending the information to each locality, a central lab might be responsible for receiving push notifications and contacting patients.

Either manually or through an automated system, volunteer ids from the push notification(s) would be encoded into DNA and server as a lookup key for locating encoded contact information. This would then be decoded and only volunteers living within that region's boundaries would be contacted to come into the facility for an important update. Important information would only be disclosed at a facility.

7.Conclusion

By advancing the precision medicine research, the NIH's proposed program could significantly improve the health of people worldwide. However, to successfully launch and carry out the program, the program must address current security issues and ensure that volunteers are not harmed by having their personal data compromised. Clearly, the current firewall approach to guarding data is not sufficient and an alternative approach is needed. The natural computing field of data storage in DNA has made significant advances in the last several years and could serve as a viable alternative. Our suggested approach to securing the personal information of Precision Medicine Initiative® Cohort Program volunteers might come with a higher activation cost. However, it might have both short-term benefits by protecting participants and long-term benefits by making

invaluable contributions to both the field of data storage and security.

8. References

1. "Scale and Scope | National Institutes of Health (NIH)." U.S National Library of Medicine. Accessed January 5, 2016. https://www.nih.gov/precision-medicine-initiative-cohort-program/scale-scope.

2. Accessed January 5, 2016. http://syndication.nih.gov/multimedia/pmi/infographics/pmi-infographic.pdf.

3. "Q&A: Francis Collins on High Hopes and a Mad Schedule for Precision Medicine." POLITICO. December 27, 2015. Accessed January 5, 2016. http://www.politico.com/story/2015/12/francis-collins-nih-health-precision-medicine-initiative-217150.

4. "NIH Head Francis Collins On New Efforts To Use Medical Records Of Volunteers To Treat Diseases - The Diane Rehm Show." The Diane Rehm Show NIH Head Francis Collins On New Efforts To Use Medical Records Of Volunteers To Treat Diseases Comments. Accessed January 5, 2016. http://thedianerehmshow.org/shows/2015-09-28/nih-head-francis-collins-on-new-efforts-to-use-medical-records-of-volunteers-to-treat-diseases.

5. Koren, Marina. "About Those Fingerprints Stolen in the OPM Hack." The Atlantic. September 23, 2015. Accessed January 5, 2016. http://www.theatlantic.com/technology/archive/2015/09/opm-hack-fingerprints/406900/.

6. Castro, Leandro N., and Leandro Nunes De. Castro. Fundamentals of Natural Computing: Basic Concepts, Algorithms, and Applications. Boca Raton: Chapman & Hall/CRC, 2006.

7. "Natural Computing." - Wikipedia, the Free Encyclopedia. Accessed January 6, 2016. https://en.m.wikipedia.org/wiki/Natural_computing.

8. Wikipedia. Accessed January 6, 2016. https://en.wikipedia.org/wiki/Metaheuristic.

9. "Artificial Neural Network." - Wikipedia, the Free Encyclopedia. Accessed January 6, 2016. https://en.m.wikipedia.org/wiki/Artificial_neural_network.

10. "Genetic Algorithm." - Wikipedia, the Free Encyclopedia. Accessed January 6, 2016. https://en.m.wikipedia.org/wiki/Genetic_algorithms.

11. "Particle Swarm Optimization." - Wikipedia, the Free Encyclopedia. Accessed January 6, 2016. https://en.m.wikipedia.org/wiki/Particle_swarm_optimization.

12. Al-Hammadi, Yousof, Uwe Aickelin, and Julie Greensmith. "DCA for Bot Detection." 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence).

13. Wikipedia. Accessed January 6, 2016. https://en.wikipedia.org/wiki/Quantum_computing.

14. Adleman, L. "Molecular Computation of Solutions to Combinatorial Problems." Science, 1994, 1021-024.

15. Wikipedia. Accessed January 7, 2016. https://en.wikipedia.org/wiki/Quantum_key_distribution.

16. Church, G. M., Y. Gao, and S. Kosuri. "Next-Generation Digital Information Storage in DNA." Science, 2012, 1628.

17. Wikipedia. Accessed January 7, 2016. https://en.wikipedia.org/wiki/DNA_digital_data_storage.

18. Wikipedia. Accessed January 7, 2016. https://en.wikipedia.org/wiki/Exabyte.

19. "Data-storage for Eternity." Data-storage for Eternity. Accessed January 7, 2016. https://www.ethz.ch/en/news-and-events/eth-news/news/2015/02/data-storage-for-eternity.html.

20. "The Prose at the End of the Universe." Engadget. Accessed January 7, 2016. http://www.engadget.com/2015/12/30/christian-bok-the-xenotext-bacteria-poetry/.

21. Zakeri, Bijan, and Timothy K. Lu. "DNA nanotechnology: new adventures for an old warhorse." Current opinion in chemical biology 28 (2015): 9-14.

22. Nature.com. Accessed January 8, 2016. http://www.nature.com/nature/journal/v494/n7435/nature11875/metrics.

23. "DNA Storage Could Preserve Data for Millions of Years." DNA Storage Could Preserve Data for Millions of Years. Accessed January 8, 2016. http://www.gizmag.com/dna-data-storage/36151/.