International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Raze the Silicon to Raise the DNA using Data Mining Techniques

D. Ramya[1], Dr. S. Prabhu[2], Dr. Rajeshkumar M[3]

[1]Assistant Professor, Department of Computer Science, Sasurie College of Arts & Science, Vijayamangalam, Tirupur, Tamil Nadu, India.
[2]Assistant Professor, Department of Computer Science, Government Arts & Science College, Thittamalai, Nambiyur, Tirupur, Tamil Nadu, India
[3]Assistant Professor Department of Computer Science KPR College of Arts Science and Research Coimbatore.

ABSTRACT

DNA Mining is an energetic technique which is used in Data Mining. DNA Mining is an effectual process for Storing and Bring back information's in the form of DNA. There are several kind of electronic storage mediums are used as a stock, but we looking forward to the oldest storage device named as DNA Mining.

DNA Mining will be used as next generation's Digital information storage medium that has immense of storage capacity. The future scope of this DNA Mining is not only focusing on the stock, but it also satisfies some factors such as: Space complexity, Time and Predict accurate solution to the require problem.

I. Introduction

Data Mining means exploit the enlightment from enormous data file. DNA Mining one of the manners used in data mining for stocking purpose. DNA Mining is the system to store and fetch data from DNA Database.

One of the current steal of all Industry is immeasurable storage. To blown away these inconvenience, Researchers discovered our oldest storage medium DNA(Deoxyribose Nucleic Acid). DNA is an upperhand way to store information, because we can extract data from bones of Woolly Mammoths,which data backtens of thousands of years and make sense of it, says researcher Nick Goldman, a molecular and evolutionary biologist and a mathematician at the European Bioinformatics Institute(EBI) in Hinxton, England.[1]

DNA molecules which encode genetic information are for all living things on computer. The accomplished DNA is the certain combination of DNA molecules are interpreted as a particular result to a combinational problem encoded in the original molecules present. There are some good reasons for researchers preferring for this is,[2]

1. DNA is, "Incredibly dense (you can store one bit per base and a base is only a few atoms large)".

2. DNA is, "Volumetric (beaker) rather than planner (hard disk)" again meaning you can pack it in efficiently.

3. Finally DNA is, "Incredibly stable- where other bleeding edge storage mediums need to be kept in sub-zero vacuums, DNA can survive for hundreds of thousands of years in a box in your garage."

This means that unlike magnetic hard drives which have a comparatively short lifespan; DNA could store data for a long time.

So, DNA can Store data more efficiently than anything we've invented, But DNA wasn't designed.

II. History

Excessive Data Storage is a major sidestep by Humanity. Our Society has an explosion on the volume of information that we are producing on a daily basis, says recent surveys conducted by IDC Digital Universe Perfusion of Technology.[3]

The world's information is doubling every two years by Texts, Videos, Tweets, Facebook updates, Unsolicited Farm Ville requests, Instagram posts and various other forms of digital data production. All the information's require to be stored, a good deal of it will be inconspicuous away somewhere for Succeeding generations. We need to find new storage solution for progression of generating information's.

On August 16, 2012 Sriram Kosuri, a Harvard geneticist and member of the Wyss Institutes Synthetic biology platform and another geneticist Yuan Gao and also Synthetic biology pioneer George Church, denominates a new procedure for using DNA to encode digital information that included on HTML draft of a 53,400 word book written by the lead researcher , eleven JPG Images and one Java script program, Multiple copies for redundancy were added and 5.5 petabitsto be stored in each cubic millimeter of DNA.[4]

A polished structure was reported in the journal Nature in January 13, in an article organized by lead researchers Nick Goldman and Ewan Birney at the European Bioinformatics Institute of European Molecular Biology Laboratory (EMBL-EBI)- were successfully stored and perfectly retrieved and also reproduced the combination of Text files and Audio files. The encoded information's consisted of all 154 of Shakespeare's sonnets, a twenty-six-second audio clip of the "I Have a Dream" speech by Martin Luther King, the well-known paper on the structure of DNA by James Watson and Francis Crick, a photograph of EBI headquarters in Hinxton, United Kingdom, and a file describing the methods behind converting the data. All the DNA files reproduced the information between 99.99% and 100% accuracy.[4] The sequences of the individual strands of DNA overlapped in such a way that each region of data was repeated four times to avoid errors.

But the idea and the general considerations about the possibility of recording, storage and retrieval of information on DNA molecules were originally made by Mikhail Neiman and published in 1964–65 in the *Radiotekhnika* journal, USSR.[4]

III. Past

To encode files in DNA, Birney and Goldman started by converting text, image, or audio data into binary code. Then, in several steps using software that Goldman wrote, they converted that into A, T, G, or C code, which stand for the four DNA bases. Working from that string of letters, they drew up the blueprints for thousands of pieces of DNA , each containing a snippet of a file, and sent their designs to Agilent Technologies, which manufactures custom DNA for biologists. Agilent sent back the completed DNA fragments—just a smidge of white dust in the bottom of a plastic tube, Goldman recalls. To open the files, the team used a standard DNA sequencer, a process that took about 2 weeks. They then used Goldman's software to reassemble the sequenced DNA into coherent, readable files. With the exception of two small gaps in the DNA, the sonnets, photo, speech, PDF, and text file re-emerged from the white dust almost completely unscathed. After the scientists performed a little repair work, all of the information—about 739 KB worth—was retrieved with 100% accuracy.[5]

The *Nature* team stored slightly more data, and Goldman avoided one of the sources of error in the Science paper—strings of repeated bases that DNA sequencers have trouble handling—by adjusting the way his software converts the information into A, T, G, and C. But on the whole, the ideas are similar, and represent a big step forward from earlier, smaller studies.[5]

IV. Present

iSGTWrecently interviewedEwan Birney, associate director of the European Bioinformatics Institute (EBI), regarding his keynote talk at the EUDAT 2nd Conference. In this interview, Birney raised the exciting prospect of using DNA as an organic data storage device. But could DNA storage really replace tapes and hard disks for long-term preservation of data? Charles Harvey investigates…

DNA is the world's oldest data storage device. The technology to read and write DNA has become commonplace since bacteria were first genetically engineered in 1973. And, while it's possible to store petabytes of data in a microscopic space, might it ever be worthwhile to store information as DNA, rather than on hard drives or magnetic tape?

A. Structure

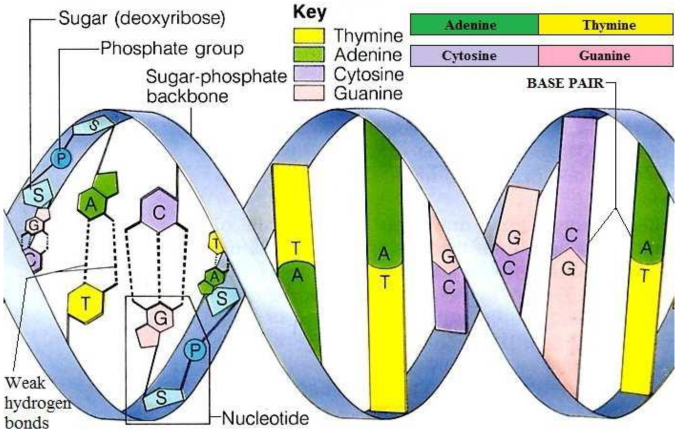


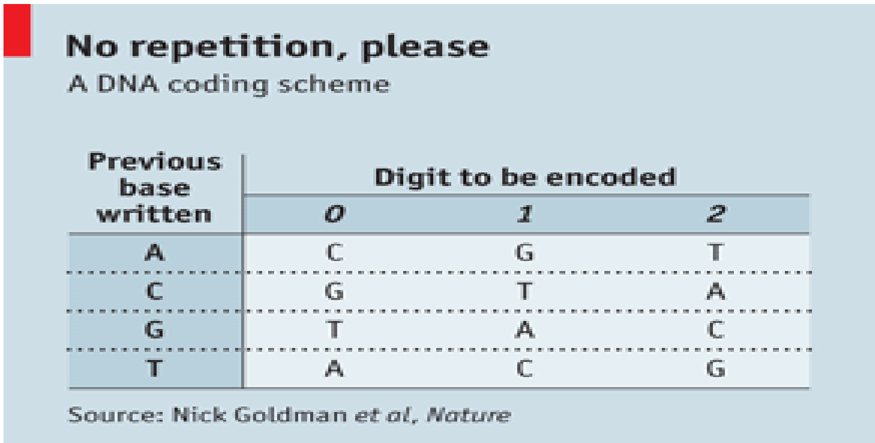
Fig 1: Structure of DNA

To store information, DNA uses 4 bases (Adenine, Cytosine, Thymine, and Guanine — often simply referred to as A, C, T, and G). In 2003, Pak Chung Wong from the Pacific Northwest National Laboratory, USA, encrypted text into DNA by converting each character into a base-4 sequence of numbers, each corresponding to a certain base. Using genetic engineering, these sequences were inserted into the genome of a bacterium, once repetitive sequences of numbers — space-consuming and potentially harmful for the bacteria — had been removed. Special beginning and end tags were also added to the strands to allow indexing and to prevent the bacteria expunging the inserted sequences as viral invaders.

Bacteria are an obvious option when it comes to storing data as DNA: they replicate quickly, creating numerous copies of the data in the process. Also, should a mutation occur within an individual bacterium, the remaining bacteria will still contain the original information, allowing researchers to recover the original sequence with close-to-perfect accuracy. [6]

B. Faithful reproduction

The trick to this fidelity lies in the way the researchers translate their files from the hard drive to the test tube. DNA uses four chemical “bases”—adenosine (A), thymine (T), cytosine (C) and guanine (G)—to encode information. Previous approaches have often mapped the binary 1s and 0s used by computers directly onto these bases. For instance, A and C might represent 0, while G and T signify 1. The problem is that sequences of 1s or 0s in the source code can generate repetition of a single base in the DNA (say, TTTT). Such repetitions are more likely to be misread by DNA-sequencing machines, leading to errors when reading the information back.



No repetition, please

A DNA coding scheme

Previous base written	Digit to be encoded		
	0	*1*	*2*
A	C	G	T
C	G	T	A
G	T	A	C
T	A	C	G

Source: Nick Goldman *et al*, *Nature*

The team’s solution was to translate the binary computer information into ternary (a system that uses three numerals: 0, 1 and 2) and then encode that information into the DNA. Instead of a direct link between a given number and a particular base, the encoding scheme depends on which base has been used most recently (see table). For instance, if the previous base was A, then a 2 would be represented by T. But if the previous base was G, then 2 would be represented by C. Similar substitution rules cover every possible combination of letters and numbers, ensuring that a sequence of identical digits in the data is not represented by a sequence of identical bases in the DNA, helping to avoid mistakes.

The code then had to be created in artificial DNA. The simplest approach would be to synthesise one long DNA string for every file to be stored. But DNA-synthesis machines are not yet able to do that reliably. So the researchers decided to chop their files into thousands of individual chunks, each 117 bases long. In each chunk, 100 bases are devoted to the file data themselves, and the remainder used for indexing information that records where in the completed file a specific chunk belongs. The process also contains the DNA equivalent of the error-detecting “parity bit” found in most computer systems.

To provide yet more tolerance for mistakes, the researchers chopped up the source files a further three times, each in a slightly different, overlapping way. The idea is to ensure that each 25-base quarter of a 100-base chunk was also represented in three other chunks of DNA. If any copying errors did occur in a particular chunk, it could be compared against its three counterparts, and a majority vote used to decide which was correct. Reading the chunks back is simply a matter of generating multiple copies of the fragments using a standard chemical reaction, feeding these into a DNA-sequencing machine and stitching the files back together.When the scheme was tested, it worked almost as planned.[7]

V. Scope on Future

Some centers of research in this area are developing new branches in this young field. Advancements are being made in cryptography. Researchers are working on decreasing error in and damage to the DNA during the computations/reactions. There are models for universal DNA computers, while others have described methods for doing addition and matrix multiplication with these computers. The field of DNA computing is truly exciting for the revolution it implies will occur within the next few years.[8]

VI. Future

DNA storage has its limitations. As I mentioned earlier, it's not re-writable, and it's not random access. Its latency is also too high for it to be practical for anything other than archival storage, but we've already established that we're in dire need of space for archiving, anyway. The only other big limiting factors, at present, are synthesis and sequencing technologies — and those won't be an issue for much longer.

According to Kosuri, the costs of DNA synthesis and sequencing have been dropping much faster than Moore's law. In the supplementary information section of their paper, Kosuri and his colleagues imagine what a petabyte of storage would require, from the standpoint of synthesis and sequencing costs, and conclude that they would need a roughly 6 order of magnitude drop in sequencing, and 7-8 in synthesis for storage media of that capacity to become feasible.

"To give perspective," explains Kosuri, "costs have been dropping for the past 5-10 years at 10x and 5x per year for sequencing and synthesis respectively." In other words: this tech is right around the corner.[3]

VII. Applications

Applications of DNA Mining used in many industries including Auditing, Health industry, Telecommunications, Retail industry etc.

Other Applications of DNA Mining used in Cryptography. RSA and DES are the two algorithm used in cryptography. Out of these DES has been broken using DNA computers. Tremendous parallel processing and enormous data capabilities are the key to breaking cryptographic algorithm. Generate all possible 64 bit keys (DES) using DNAmemory strands. For each of 24^64 possible keys, computethe cipher text using the current key. The use of a defined strength of phosphoric acid for themanufacture of artificial gene for data storage will alsoneed the exact strength of the chemicals during decoding orthe gene will be destroyed or corrupted.[8]

VIII. Characteristics

The major reason why this would have been difficult in the past is that it is really difficult to construct a large stretch of DNA with exact sequence, and make it cheaply. We took an approach that allows us to use short stretches of DNA (basically by having an address (19 bits) and data block (96 bits), so each short stretch can be stitched together later after sequencing. Using short stretches allowed us to leverage both next-generation synthesis [for writing data]… and next-generation sequencing [for reading data] technologies to really lower cost and ease.[3]

IX. How DNA Computers will work

Today the microprocessor and computer chip manufacturersare in a race to develop the next generation IC's that willtopple the speed records, but this race will soon hit theboundary of limits of speed and miniaturization. To avoid thischip manufacturer need to find a new material that willproduce faster computing speeds. This new material wasfound out to be a natural DNA that the scientists claim wouldbe our next-gen material for making of IC's instead of siliconused currently. There are millions of natural supercomputersthat exist inside living organisms. DNA (deoxyribonucleicacid) molecules, the complex chemical compound the genesare made of, have the prospective to carry out calculationsmany times faster as compared to the world's fastest andpowerful man made computer [8]. DNA will be incorporatedin a computer chip to make a biochip that will speed up thecomputers even faster. DNA molecules have already beenharnessed to compute complex mathematical problems.

X. Speed

The speed of read and write on an artificial gene is quite high and the method to make and decode is extremely costly so can be used by limited number of people as it will need proper labs for it. This technique also involves the immobilization and manipulation of combinatorial mixtures of DNA on a support.[8]

XI. Advantages

DNA has significant advantages over both printed text and electronic media. For one thing, it can stay behind stable for long periods of time with a minimum of care. Intact DNA has been extracted from bones (andother organic matter) tens of thousands of years old, and its sequence reconstructed with as much detail as if it had come directly from a living organism.

Another advantage of DNA over electronic media is that it assess no power supply to maintain its integrity, which makes it easy to transport and store, and potentially less vulnerable to technological failure.

Perhaps the greatest advantage of DNA as a storage medium is its minuteness. For example, EMBL-EBI'sofficial press release claims that more than 100 million hours of high-definition video could be stored in roughly a cup of DNA.

XII. Disadvantages

This is cost-effective only for archives intended to last hundreds or even thousands of years — something few of us contemplate.

The main cost of maintaining electronic archives over such a long period of time is that the media have to be periodically replaced and the data copied, whereas DNA has merely to be stored somewhere cool, dry and dark.

But if the cost of synthesizing DNA can be reduced by one or two orders of magnitude — which, judging by current trends could occur within a decade — DNA archives intended to last less than 50 years would become feasible.

The major drawback is the current cost of synthesizing DNA in the quantities required, estimated at around $US12, 400 per megabyte of data stored.[9]

XIII. DNA Sequence Mining

Sequence Mining means finding sequential patterns among the large dataset. It finds out frequent substring as patterns from adataset. With massive amounts of data continuously being gathered , many industries are becoming interested in diggingsequential patterns from their databases. Sequential pattern mining is one of the most well-known methods and has broadapplications including web-based analysis, customer procure behavior analysis and medical record analysis. In the retailingbusiness, sequential patterns can be mined from the transaction records of customers. For example, having bought a breadpacket, a customer comes back to buy a butter and a milk packet next time. The seller can use all this information foranalyzing the behavior of the customers procure , to understand their interests, to satisfy their demands, and to predict theirrequirements. In the medical field, sequential patterns of symptoms of any diseases exhibited by patients to identify strongsymptom/disease correlations that can be a valuable source of information for medical diagnosis and preventive medicine. InWeb log analysis, the exploring behavior of a user can be extracted from member records or log files. For example, havingviewed a web page on "Data Extraction", user will return to evaluate "Business Perception" for new information next time.These sequential patterns give huge profits, when acted upon, increases customer royalty.The goal of sequential data mining is to discover frequently occurring patterns but not identical. The challenge in discoveringsuch patterns is to allow for some *noise* in the matching process. To find such a method first is to find the definition of apattern, and then definition of similarity between two patterns. This similarity definition of the two patterns can vary from oneapplication to another.[10]

Reference

[1]. Chromosome Analysis Using Laplacian Based Centromere Detection, International Journal ofRecentTechnologyAndEngineering2019,pp.265-2705.92.

[2]. Ahmed, F, Bari, A.H, Hossain, E, Al-Mamun, H.A and Kwan, P, (2011). Performance analysis of support vector machine and bayesian classifier for crop and weed classification from digital images. World Applied Sciences Journal,vol.12,no.4,pp.432-440.

[3]. Alam, M, Alam, M.S, Roman, M, Tufail, M, Khan, M.U and Khan, M.T, (2020), Real-time machine-learning based crop/weed detection and classification for variable-rate spraying in precision agriculture. In 7th International Conference on Electrical and Electronics Engineering (ICEEE),pp.273-280.

[4]. Alamdar,F and Keyvanpour,M,(2011).A new color feature extraction method based on QuadHistogram.Procedia Environmental Sciences,vol.10,pp.777-783.

[5]. Albawi, S, Mohammed, T.A and Al-Zawi, S, (2017), Understanding of a convolutional neural network. In 2017 International Conference on Engineering and Technology (ICET),pp.1-6.

[6]. Ali,H,Lali,M.I,Nawaz,M.Z,Sharif,M and Saleem,B.A,(2017).Symptom based automated detection of citrus diseases using color histogram and textural descriptors.Computers and Electronics in agriculture,vol.138,pp.92-104.

[7]. Ali,M,Guru,D.S and Suhil,M,(2018).Classifying Arabic Farmers Complaints Based on Crops and Diseases Using Machine Learning Approaches. In International Conference on Recent Trends in Image Processing and Pattern Recognition,pp.416-428.

[8] P. Lichter, "Multicolor fishing: what's the catch?" Trends Genet., vol. 13,pp. 475–479,1997.

[9] K.R.Castleman,"Match recognition in chromosome band structure,"Biomed. Sci.Instrum.,vol.4,pp.256–264,1968.

[10].K. Paton, "Automatic chromosome identification by the maximumlikelihoodmethod,"Annals of Human Genetics,vol.33,PP.177-184,1969.