Data storage in the genetic code

MADE-TO-MEASURE DNA MAY BE THE SOLUTION FOR LONG TERM STORAGE OF MOUNTING DATA LOADS, WRITES **S ANANTHANARAYANAN**

Data storage in magnetic tape or hard discs needs high maintenance so that it is not corrupted and that it stays in a format that is readable in the future. An emerging alternative is storage in the way the genetic code is preserved in DNA for millions of years. It is reported that Twist Bioscience, a California-based start-up company that specialises in building bits of DNA from scratch, has been engaged by the Microsoft Corporation to supply 10 million DNA strands and help try out the new medium.

Dr Emily M Leproust, CEO of Twist Bioscience, was, in fact, an author of seminal papers — one in 2010 that reported building DNA strands and the other in 2013, where the strands were used to store digital data.

Emily M Leproust

Data storage has grown to become an area of great importance. The output of very expensive scientific programmes, like space exploration, or CERN's Large Hadron Collider, is immense data. While the data gets processed over months, either in large, dedicated facilities or even through crowd sourcing, a good part of it could be needed again and needs to be preserved. Another field that gives rise to huge data is the digitising of civil documents, like land records, topographic data, layout of drainage or communication lines. As actual drawings and plans deteriorate and the data is growing, most of these have been converted to digital formats, as are also many other records that need archiving.

But digital storage also deteriorates and, what is worse, the decoding technology, like DVD format, also changes. The digital data, hence, needs to be periodically renewed, both to prevent deterioration as well as to bring the storage format up to date. Such conversion and verification can be a very time-consuming and expensive and with increasing loads of data the investment on maintenance can be comparable to that of data acquisition.

The microscopic DNA macro-molecule, in contrast, is able to pack huge data within very small volume and the stability of the record is legendary. Apart from being the vehicle of faithfully transmitting the mammoth genetic data of living things across generations, it is routine that fossils and organic remains from prehistoric times contain DNA in good enough condition for research. Finding a way to encode digital, computer-generated data on to the DNA structure would, hence, permit very compact and hardy storage.

Digital format

Computers consist of electronic devices that ultimately take only the states of "on" or "off", represented by the numbers "0" and "1". All data, be it of text, images or sounds, hence have to be coded with the help of only these two numbers. This coding is done with the help of a form of counting that is based on only the numbers "0" and "1" and is called binary arithmetic, as opposed to the usual decimal arithmetic, based on the number 10.

The decimal system has symbols for the numbers from "0" to "9" and when we reach the number ten we write it as "10", to say that it is one "ten" and no units. In the same way we write twenty as "20", one more as "21", and so on till a hundred is "100", to indicate ten 10s and no more. In binary arithmetic, we do the same with the number, two taking the place of the number 10. Thus, we have the two symbols, "0" and "1" and when we count one more, we say "10" to indicate "one time the number, two and no more". One more, or the number "3", would become "11" to mean "one times two and one more", and so on. The number "4" would be written as "100", the number "5" as "101", "six" as "110", and so on.

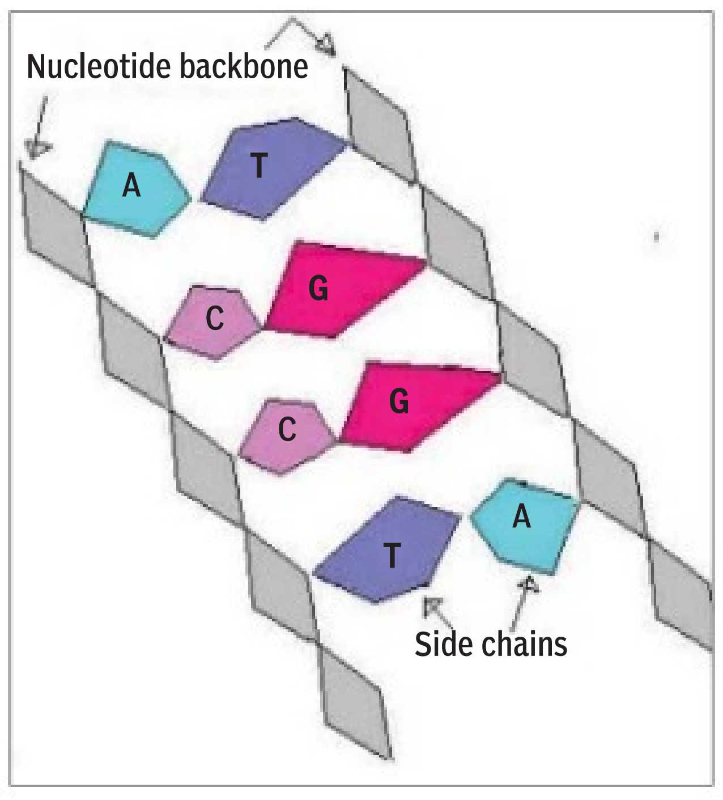
When we wish to represent features of text, like the alphabet, digits, punctuation and other symbols, there is a convention known as the American Standard Code for Information Interchange, where all text characters are represented by the 128 numbers from zero to 127. Thus, the characters A, B, C… Z are coded as the numbers "65" to "90", small letters, a,b,c… z are coded as numbers "97" to "122", the full stop is "96", the comma is "44", etc. But computers still cannot recognise these numbers, and the numbers are again converted to binary, like "65" for "A" becomes "1000001", "122" for "z", becomes "1111010", etc.

Even long texts, along with the additional information of font, size, colour, etc, as we type into the keyboard, get converted into binary representation and that is the way text is stored and processed in the computer. There are again standards like "jpg" or "bmp", which specify how the data of pixels that make up an image would be represented by eight-digit binary numbers, and there are similar conventions for audio files. This article, for instance, consists of 9,523 characters (including spaces) and three image files and was represented in MS Word format, along with font, margins, etc, data, in 141,000 eight-digit binary numbers.

DNA representation

The DNA is a string of chemical units, each one of which can take on and be differentiated by one of four types of side markers. The units of the string are called nucleotides and they attach to each other like the carriages of a train. The four kinds of "side chains" are denoted as C, G, A and T. These side chains also form bonds with the side chains of a parallel train of nucleotides, but with a rule that A pairs with T and C pairs with G. The parallel train thus forms with the side chains that correspond to those of the first train, and with bonds forming right through the length of the string the DNA molecule has remarkable stability and resilience, till special enzymes cause it to separate for reproduction, etc.

It is the order in which C, G, A, T are attached that specifies the different amino acids of the different proteins of a species that the DNA codes for. The scheme, in fact, is based on groups of three nucleotides, each of which can have any of four kinds of side chains, and in the group of three there can be 4x4x4=64 different combinations. With provision of redundancy to take care of any errors, and also for markers to indicate the start and the end of the code for a protein, these 64 possible combinations code for 20 amino acids, which are the constituents of all the millions of proteins.



The same idea can also be extended to represent characters of text, digits, distribution of pixels, etc, to form coding that stores computer records. While the successive units in magnetic tape or on a hard disc can take only the two forms of "on" or "off" or "up" or "down", to represent the numbers "0" or "1", the units in DNA can take four forms. In practice, it is found that strings of more than three units with the same side chain are not stable. The side chain "G" is hence not used in the coding, but is used only as a filler to break any chain of the repeated letters, and we are left with only three forms of each unit for coding.

Just as the two forms of units in conventional storage lead to binary arithmetic based on the number two, the three forms in DNA coding give rise to a number system based on the number three, known as "trinary". Here, we have three symbols, "0", "1" and "2" and when we count "3", we write "10" to show "one three and no more", and so on:

Decimal	binary	trinary	Decimal	binary	trinary
0	0	0	5	101	12
1	1	1	6	110	20
2	10	2	7	111	21
3	11	10	8	1000	22
4	100	11	9	1001	100

Creating the DNA

In principle, the digital, or binary representation of all kind of documents can be converted to trinary and then coded on to a string of nucleotides in DNA. We would then have all our data preserved in the billion-nucleotide-long DNA molecules, compact and secure, able to last centuries!

The trouble is that it is no simple task to create DNA in the way we like. Even in nature, present-day DNA has evolved from simpler forms and does not form unit by unit, but assembles when the strings of DNA separate for reproduction.

Artificial synthesis is carried out by actual attachment of successive nucleotides, using materials with nanometre pores as a scaffold. The process, however, is limited by side reactions and a chain of more than 100 nucleotides was not possible. The firm, Agilent Technologies Inc, where Dr Leproust was a researcher, has refined the process to create stretches of over 150 nucleotides. These find application in biomedical research and were used in the DNA data storage trial reported in 2013.

The trial used a large text sample, over 700 kilobytes (this article is about 141 KB) and transferred the data on to 153,335 DNA strings, each one 117 nucleotides long. The data on each string also carried information that identified the portion of text, where in the whole text the portion belonged, check digits, to detect and possibly correct any errors and also a large overlap. DNA itself consists of a pair of complementary strings that act as alternate copies. Along with the overlap, the coding thus provided ample redundancy and security. The paper of 2013 reported error-free reproduction of the coded material when the mass of DNA strings were decoded using methods of genetic engineering.

We can see that DNA coding is not as simple as writing to the disc by hitting the "save" button. And then retrieval is a task too. But the application is for data that needs to be saved for a long time, which would have recurring costs in the normal way. The costs of DNA coding and decoding would also come down. The result of the Microsoft Corporation trials may set the course.

THE WRITER CAN BE CONTACTED AT
response@simplescience.in

THE GENE EXCHANGE

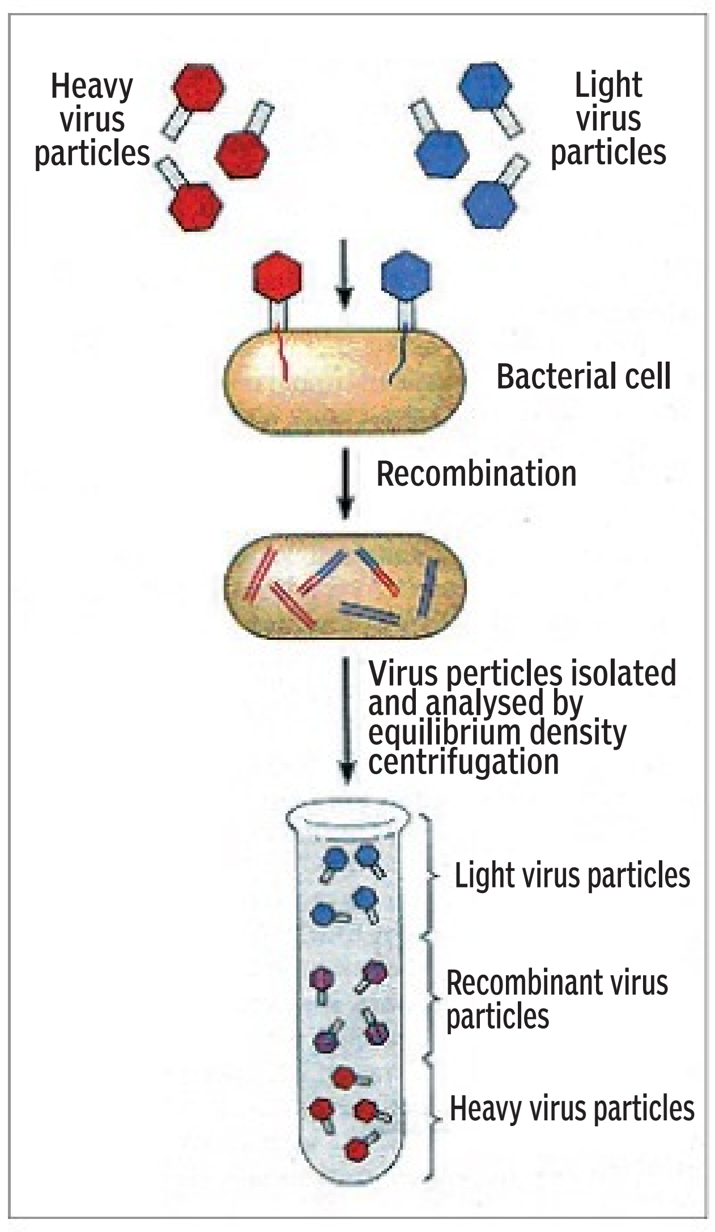
TAPAN KUMAR MAITRA EXPLAINS THE MOLECULAR MECHANISM OF HOMOLOGOUS RECOMBINATION

There are five different situations in which genetic information can be exchanged between homologous DNA molecules, namely: prophase I of meiosis associated with gametogenesis in eukaryotes; co-infection of bacteria with related bacteriophages; transformation of bacteria with DNA; transduction of bacteria by transducing phages; and bacterial conjugation. In spite of their obvious differences, all five situations share a fundamental common feature — each involves *homologous recombination* in which genetic information is exchanged between DNA molecules exhibiting extensive sequence similarity.

Since the principles involved appear to be quite similar in prokaryotes and eukaryotes, we will use examples from both types of organisms. Shortly after it was first discovered that genetic information was exchanged between chromosomes during meiosis, two theories were proposed to explain how this might occur. The *breakage-and-exchange model* postulated that breaks occurred in the DNA molecules of two adjoining chromosomes, followed by an exchange and rejoining of the broken segments. In contrast, the *copy-choice model* proposed that genetic recombination occurred while DNA was being replicated.

According to the latter view, DNA replication begins by copying a DNA molecule located in one chromosome and then switches at some point to copying the DNA located in the homologous chromosome. The net result would be a new DNA molecule containing information derived from both chromosomes. One of the more obvious predictions made by the copy-choice model is that DNA replication and genetic recombination should happen at the same time. When subsequent studies revealed that DNA replication took place during S phase while recombination typically occurred during prophase I, the copy-choice idea had to be rejected as a general model of meiotic recombination.

The first experimental evidence providing support for the breakage-and-exchange model was obtained in 1961 by Matthew Meselson and Jean Weigle, who employed phages of the same genetic type labelled with either the heavy (^{15}N) or light (^{14}N) isotope of nitrogen. Simultaneous infection of bacterial cells with these two labelled strains of the same phage resulted in the production of recombinant phage particles containing genes derived from both phages. When the DNA from these recombinant phages was examined, it was found to contain a mixture of ^{15}N and ^{14}N. Since these experiments were performed under conditions that prevented any new DNA from being synthesised, the recombinant DNA molecules must have been produced by breaking and rejoining DNA molecules derived from the two original phages.



Evidence for DNA breakage-and-exchange during bacteriophage recombination. In this experiment, bacterial cells were infected with two strains of the same phage, one labelled with ^{15}N and the other with ^{14}N. After recombination, the DNA from the recombinant phages was found to contain both ^{15}N and ^{14}N, supporting the idea that recombination involves the breaking and rejoining of DNA molecules.

Subsequent experiments involving bacteria whose chromosomes had been labelled with either ^{15}N or ^{14}N revealed that DNA containing a mixture of both isotopes was also produced during genetic recombination between bacterial chromosomes.

Moreover, when such recombinant DNA molecules are heated to dissociate them into single strands, a mixture of ^{15}N and ^{14}N is detected in each DNA strand; hence, the DNA double helix must be broken and rejoined during recombination.

A similar conclusion emerged from experiments performed shortly thereafter on eukaryotic cells by J Herbert Taylor. In these studies, cells were briefly exposed to ^{3}H-thymidine during the S phase preceding the last mitosis prior to meiosis, producing chromatids containing one radioactive DNA strand per double helix. During the following S phase, DNA replication in the absence of ^{3}H-thymidine generated chromosomes containing one labelled chromatid and one unlabelled chromatid. But during the subsequent meiosis, individual chromatids exhibited a mixture of radioactive and non-radioactive segments, as would be predicted by the breakage-and-exchange model.

Moreover, the frequency of such exchanges was directly proportional to the frequency with which the genes located in these regions underwent genetic recombination. Such observations provided strong support for the notion that genetic recombination in eukaryotic cells, as in prokaryotes, involved DNA breakage and exchange.

These experiments also showed that most DNA exchanges arose between homologous chromosomes rather than between the two sister chromatids of a given chromosome. This selectivity is important because it ensures that genes are exchanged between paternal and maternal chromosomes.

THE WRITER IS ASSOCIATE PROFESSOR, HEAD, DEPARTMENT OF BOTANY, ANANDA MOHAN COLLEGE, KOLKATA, AND ALSO FELLOW, BOTANICAL SOCIETY OF BENGAL, AND CAN BE CONTACTED AT
tapanmaitra59@yahoo.co.in

A new way forward

DOUG BOLTON REPORTS ON A STUDY THAT SAYS VIRTUAL REALITY CAN BE USED TO CURE SEVERE PARANOIA

Oxford University researchers have found that virtual Reality can be hugely effective in treating people suffering from severe paranoia. Using top-of-the-range VR headsets that can track users' movements and immerse them in a simulated world, the Oxford team virtually placed paranoia sufferers in environments they found stressful, like crowded trains or cramped lifts.

By experiencing these places in a controlled setting, the patients were able to practice how to deal with them. By the end of the groundbreaking tests, many of the patients reported a marked decrease in their paranoia.

The number of those suffering from severe paranoia, usually as a central feature of mental health disorders like schizophrenia, around the world may represent a negligible percentage. They may incorrectly believe that others are deliberately trying to attack, mock or upset them, and the feeling can be so profound in some patients that they are unable to leave the house.

A woman uses one of the VR headsets used in the tests.

Defensive behaviours such as avoiding eye contact or limiting social interactions appear to provide temporary relief, but only exacerbate the problem.

In the tests, the participants donned a VR headset and entered a virtual Tube train, which became more and more crowded in each session. One group of participants was encouraged to use their normal defence mechanisms, but were told that the situation would become more tolerable as they got used to it.

The others were told to drop all of their usual behaviours and instructed to approach and look at the human-like avatars inside the train, making eye contact or standing toe-to-toe with them. This second group experienced the biggest improvements —(a) over 50 per cent of them said they no longer had severe paranoia at the end of the testing day.

Even in the first group, things got better — 20 per cent reported their paranoia had gone after the VR experience.

Professor Daniel Freeman, from Oxford's Department of Psychiatry, said in a statement, "Paranoia all too often leads to isolation, unhappiness and profound distress. But the exceptionally positive immediate results for the patients in this study showed a new route forward in treatment."

The treatment isn't easy for patients, since dropping long-held defences takes time and courage. However, he said, "As they relearned that being around other people was safe we saw their paranoia begin to melt away. They were then able to go into real social situations and cope far better. This has the potential to be transformative."

VR has been used in similar ways to treat veterans suffering from Post-Traumatic Stress Disorder. Skip Rizzo, a University of Southern California psychologist and leading VR therapy researcher, has been able to help sufferers by getting them to virtually relive their traumatic battlefield experiences in a safe environment, where they can better process their painful memories and emotions.

Further research will be needed to see if the benefits the Oxford participants experienced lasted beyond the testing day. The therapy should also become more accessible as VR technology lowers in price.

Dr Katherine Adcock, from the Medical Research Council, which funded the study, explained, "There is a lot of work to be done in testing the approach in treating delusions but this study shows a new way forward."

THE INDEPENDENT

PLUS POINTS

'Finding' da Vinci

An international team of scientists is bidding to track down the real remains of Leonardo da Vinci, extract his DNA to shed new light on his character and create a model of what the great Renaissance genius would have looked like. Cracking the real da Vinci code is expected to be difficult and some of the world's leading experts in genetics — including some who worked with the FBI to identify those killed in the 9/11 attacks — will be working on the case.

A cleaner vacuums in front of a Leonardo da Vinci self-portrait drawn around 1515 or 1516 at an exhibition in Brussels.

The project could see hairs taken from paintings known to be his work, the owners of some of his hand-written journals — such as Queen Elizabeth, the Vatican and Bill Gates — could be asked to submit them for fingerprint tests and suspected living relatives will be asked to provide samples of their DNA for analysis.

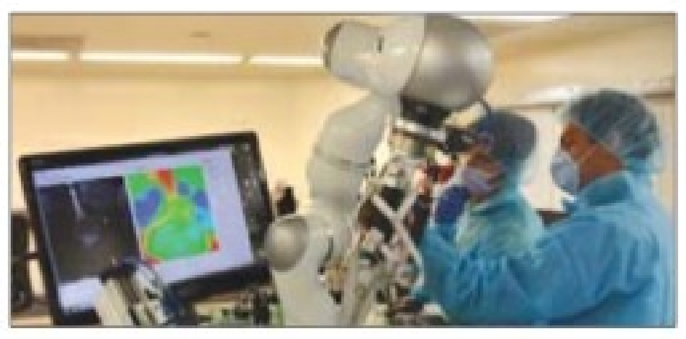
Jesse Ausubel, vice-chairman of the US-based Richard Lounsbery Foundation, which has helped to create the project, said, "I think everyone in the group believes that Leonardo, who devoted himself to advancing art and science, who delighted in puzzles, and whose diverse talents and insights continue to enrich society five centuries after his passing, would welcome the initiative of this team — indeed would likely wish to lead it were he alive today."

Announcing The Leonardo Project in the journal *Human Evolution*, the researchers said they hoped to mirror the success of projects to identify the remains of people like King Richard III of England and Miguel de Cervantes, the author of *Don Quixote* It is believed da Vinci, who died in 1519 at the age of 67, is buried in Amboise, France.

THE INDEPENDENT

Robot surgery

A team of US doctors has shown for the first time that soft tissue surgery can soon be performed entirely by a robot on humans, putting surgery one step closer into the realm of intelligent machines. The so-called Smart Tissue Autonomous Robot succeeded in suturing and reconnecting bowel segments in living pigs — a procedure known as intestinal anastomosis — and all the animals survived with no complications.



The robotic sutures were compared with the work of five surgeons completing the same procedure using three methods — open, laparoscopic and robot-assisted surgery with the well-known da Vinci Surgical System. The robot's time was longer than open and robot-assisted surgery but comparable to the laparoscopic procedure. By all other measures, the robot's performance was comparable to or better than the surgeons. "No significant differences in erroneous needle placement were noted among all surgical techniques," the researchers wrote, "suggesting Star was as dexterous as expert surgeons in needle placement."

According to Peter Kim, associate surgeon-in-chief at Children's National Health System in Washington, they tweaked what the robot was doing about 40 per cent of the time. The other 60 per cent of the time, the machine did it by itself without any interference.

IANS

Plant longevity

Compared to humans' century-long life span, some plants — evergreens in particular — have the capacity to live for an exceptionally long time, even millennia. In a study published in *Current Biology* on 5 May, scientists from the University of Bern in Switzerland presented evidence for a potential mechanism that could help explain some plants' everlasting longevity: minimal stem cell divisions to avoid "mutational meltdown".

The team zeroed in on the formation of axillary meristems — stem cells that give rise to branches — in *Arabidopsis thaliana* and tomato, finding few cell divisions between the apical meristem located at the very top of a plant and the axillary meristems. With such little proliferation comes less opportunity to accumulate potentially deleterious genetic mutations in somatic cells that could kill the organism, the authors reasoned.

The research team was led by Cris Kuhlemeier, who studies plant development at the University of Bern.

THE SCIENTIST