DNA COMPUTING

The yin–yang codec for archival DNA storage

A robust and reliable codec is the backbone for any digital DNA storage. A recent work introduces a codec based on ancient Chinese philosophy, yin–yang, that outperforms other codecs in terms of reliability and physical information density.

Manish K. Gupta

In the modern world, whenever we use our mobile phones to click on a picture and post it on social media accounts, we do not worry about the availability of storage, as we assume that companies such as social media platforms have enough storage space. However, data storage is a burning issue, not only for these companies, but also for our society at large, as we produce a substantial amount of data daily. More specifically, we produce data in the range of a couple of exabytes (1 exabyte = 1 billion gigabytes) every day[1], and the pace is accelerating: the Internet of Things (IoT) will potentially produce data in the order of geopbytes (1 geopbyte = 1 trillion exabytes). Ultimately, this poses immediate challenges related to cost and space: current data storage technologies will require a substantial amount of space to store our data, which can be fairly expensive. A promising avenue for improving our data storage capabilities is using DNA as a storage medium. For instance, one gram of DNA can store up to 455 exabytes of data[2], meaning that 2 grams of DNA can store the entire Internet (which takes approximately 700 exabytes), and 1 kg of DNA can store all the data produced by humankind. However, while DNA storage might be a natural choice for our current needs, there are still some challenges to be addressed, including finding an optimal codec that can achieve high information density and provide good error tolerance. Writing in *Nature Computational Science*, Zhi Ping and colleagues[3] present a codec based on the ancient Chinese philosophy of yin and yang that achieves an in vivo physical information density that is close to the theoretical maximum.

The DNA storage process can be divided into six simple steps (Fig. 1). First, one must convert all media files (byte strings of zeros and ones) into DNA strings (strings of A, C, G, and T) using the encoding module of a codec. This ensures that the DNA strings that store our data are protected from certain types of errors (for example, insertion, deletion, substitution, secondary structure formation, and so forth). In step 2, the DNA strings are synthesized (written) using a machine, and then stored in test tubes (step 3). To access the data, one would need to retrieve the test tube (step 4) and use a DNA sequencer machine to read the data in the form of DNA strings (step 5). These DNA strings would then need to be further decoded to get the actual media files, again using the codec, in step 6.

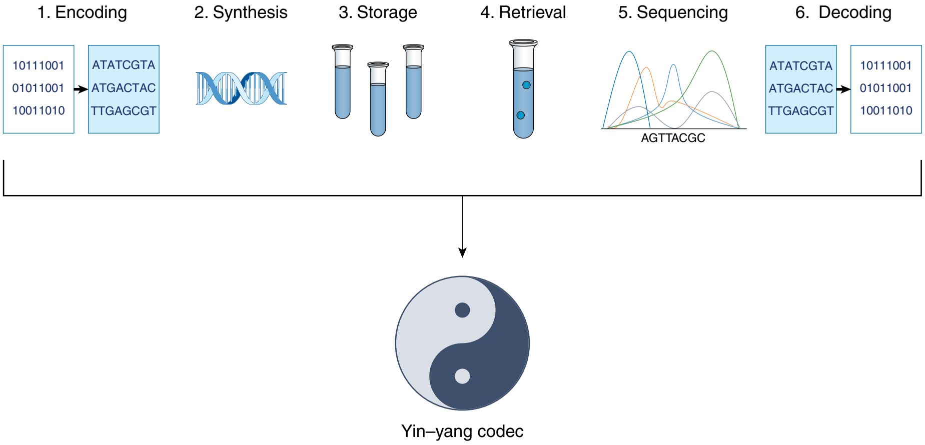

Fig. 1 | The six steps of the DNA storage process. Steps 1 (encoding) and 6 (decoding) are performed by the codec. Machines carry out DNA writing (synthesis) and DNA reading (sequencing).

The complete process is prone to various types of errors. The most common error is deletion, which may occur if the designed DNA strings from the codec have repeated symbols of A, C, G, or T, known as homopolymers. Similarly, if the designed DNA strings form a secondary structure, there will be a problem. The codec's job is then to create the DNA strings in such a manner to minimize these errors. Another important task (and challenge) for the codec is that, ideally, the information density per nucleotides should be maximized. The information density achieved so far by previous work range from 1 to 1.98, and most importantly, the best physical density previously reported is of 215 petabytes per gram of DNA[5], which is yet far from the maximum physical density of 455 exabytes per gram of DNA.

The yin–yang codec[3] is motivated by Goldman's rotating encoding strategy (where binary strings are converted to ternary strings using Huffman coding and then using a specific table[4] to get homopolymer-free DNA strings) and the DNA fountain coding strategy (erasure codes involving Luby transform and screening[5]), and it has three steps. First, byte strings are partitioned into segments of equal length. Then, two binary segments are selected randomly and are combined bit by bit by using the yang rule first, and then by using the yin rule next, returning a final nucleotide as an output. In the yang rule, [A,T] represents bit 0 and [G,C] represents bit 1. In the yin rule, the current nucleotide to be encoded is represented by involving the previous nucleotide and the corresponding bit. In the last step, the generated DNA strings are screened for constraints of GC-content (40–60%), maximum homopolymer length (<4), and secondary structure free energy (≥−30 kcal mol^{-1}). Using their approach, there are 1,536 options for encoding the binary sequence, which ultimately results in high information density (1.965). The robustness of their approach was tested by introducing random and systematic errors: the data recovery percentage can be maintained at 98% with a sequence loss rate of <2%. Most importantly, the scheme also produced, for the first time, an in vivo physical (experimentally measured) information density of ~432.2 exabytes per gram of DNA, which is substantially better than previous work.

Although the yin–yang codec gives a physical information density close to the theoretical maximum, the success of DNA data storage still depends on many other factors such as an efficient error correction for insertion/deletion and sequence loss, the DNA synthesis (writing) step, and the DNA sequencing (reading) step. As a matter of fact, the speed for both these processes is very important for practical utility. At present, new technologies, such as enzymatic DNA synthesis and nanopore sequencing, can substantially increase the overall speed of the workflow depicted in Fig. 1. As a consequence, this will likely provide a boost to produce a commercial product for DNA storage. Recently, an international consortium[6] of more than 50 universities and companies has been formed to frame the uniform standards and protocols for DNA data storage technology. Overall, we will likely see more progress towards the current challenges of optimal codec design, cost, and speed of digital DNA storage very soon. ❐

Manish K. Gupta ✉

Laboratory of Natural Information Processing, Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, Gujarat, India.

✉*e-mail: mankg@guptalab.org*

Published online: 25 April 2022
https://doi.org/10.1038/s43588-022-00235-y

References

1. Marr, B. How much data do we create every day? The mind-blowing stats everyone should read. *Forbes* (May 2018); https://go.nature.com/3qJusli
2. Church, G. M., Gao, Y. & Kosuri, S. *Science* **337**, 1628–1628 (2012).
3. Ping, Z. et al. *Nat. Comput. Sci.* https://doi.org/10.1038/s43588-022-00231-2 (2022).
4. Goldman, N. et al. *Nature* **494**, 77–80 (2013).
5. Erlich, Y. & Zielinski, D. *Science* **355**, 950–954 (2017).
6. https://dnastoragealliance.org

Competing interests

The author declares no competing interests.