REPORT

DNA STORAGE

DNA Fountain enables a robust and efficient storage architecture

Yaniv Erlich[1,2,3]* and Dina Zielinski[1]

DNA is an attractive medium to store digital information. Here we report a storage strategy, called DNA Fountain, that is highly robust and approaches the information capacity per nucleotide. Using our approach, we stored a full computer operating system, movie, and other files with a total of 2.14×10^6 bytes in DNA oligonucleotides and perfectly retrieved the information from a sequencing coverage equivalent to a single tile of Illumina sequencing. We also tested a process that can allow 2.18×10^{15} retrievals using the original DNA sample and were able to perfectly decode the data. Finally, we explored the limit of our architecture in terms of bytes per molecule and obtained a perfect retrieval from a density of 215 petabytes per gram of DNA, orders of magnitude higher than previous reports.

Humanity is currently producing data at exponential rates, creating a demand for better storage devices. DNA is an excellent medium for data storage, owing to its demonstrated information density of petabytes of data per gram, high durability, and evolutionarily optimized machinery to faithfully replicate this information (*1*, *2*). Recently, a series of proof-of-principle experiments has demonstrated the value of DNA as a storage medium (*3*–*9*).

To better understand its potential, we explored the Shannon information capacity (*10*, *11*) of DNA storage (*12*). This measure sets a tight upper bound on the amount of information that can be reliably stored in each nucleotide. In an ideal world, the information capacity of each nucleotide could reach 2 bits, as there are four possible options. However, DNA encoding faces several practical limitations. First, not all DNA sequences are created equal (*13*, *14*). Biochemical constraints dictate that DNA sequences with high GC content or long homopolymer runs (e.g., AAAAAA…) are undesirable, as they are difficult to synthesize and prone to sequencing errors. Second, oligonucleotide (hereafter "oligo") synthesis, polymerase chain reaction (PCR) amplification, and decay of DNA during storage can induce uneven representation of the oligos (*7*, *15*). This might result in dropout of a small fraction of oligos that will not be available for decoding. In addition to biochemical constraints, oligos are sequenced in a pool and necessitate indexing to infer their order, which further limits the number of available nucleotides for encoding information. Quantitative analysis shows that the biochemical constraints reduce the coding potential of each nucleotide to 1.98 bits. After combining the expected dropout rates and barcoding demand, the overall Shannon information capacity of a DNA storage device is ~1.83 bits per nucleotide for a range of practical architectures (*12*) (figs. S1 to S5 and tables S1 to S3).

Previous studies of DNA storage realized about half of the Shannon information capacity of DNA molecules. In addition, most of the previous studies reported challenges in perfect information retrieval (Table 1). For example, two previous studies attempted to address oligo dropout by dividing the original file into overlapping segments so that each input bit is represented by multiple oligos (*4*, *6*). However, this repetitive coding procedure generates a loss of information content and is poorly scalable (fig. S6). In both cases, these studies reported small gaps in the retrieved information (*4*, *6*). Other studies explored the use of Reed-Solomon (RS) code on small blocks of the input data to recover dropouts (*5*, *7*). Although these studies were able to perfectly retrieve the data, they were still far from realizing the capacity. Moreover, testing this strategy on a large file size highlighted difficulties in decoding the data due to local correlations and large variations in the dropout rates within each protected block (*7*), which is a known issue of blocked RS codes (*16*, *17*). Only after employing a complex multistep procedure and high sequencing coverage was the study able

[1]New York Genome Center, New York, NY 10013, USA. [2]Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY 10027, USA. [3]Center for Computational Biology and Bioinformatics (C2B2), Department of Systems Biology, Columbia University, New York, NY 10027, USA.
*Corresponding author. Email: yaniv@cs.columbia.edu

Table 1. Comparison of DNA storage coding schemes and experimental results. For consistency, the table describes only schemes that were empirically tested with pooled oligo synthesis and high-throughput sequencing data. The schemes are presented chronologically on the basis of publication date. Coding potential is the maximal information content of each nucleotide before indexing or error correcting. Redundancy denotes the excess of synthesized oligos to provide robustness to dropouts. Error correction/detection denotes the presence of error-correction or -detection code to handle synthesis and sequencing errors (RS, Reed-Solomon codes). Full recovery indicates whether all information was recovered without any error. Net information density indicates the input information in bits divided by the number of synthesized DNA nucleotides (excluding adapter annealing sites). Realized capacity is the ratio between the net information density and the Shannon capacity of the channel. Physical density is the actual ratio of the number of bytes encoded and the minimal weight of the DNA library used to retrieve the information. This information was not available for studies by Bornholt *et al*. (*6*) and Blawat *et al*. (*7*), as indicated by the dashes. See (*12*) for more information.

Parameter	Church *et al*. (*3*)	Goldman *et al*. (*4*)	Grass *et al*. (*5*)	Bornholt *et al*. (*6*)	Blawat *et al*. (*7*)	This work
Input data (Mbytes)	0.65	0.75	0.08	0.15	22	2.15
Coding potential (bits/nt)	1	1.58	1.78	1.58	1.6	1.98
Redundancy	1	4	1	1.5	1.13	1.07
Robustness to dropouts	No	Repetition	RS	Repetition	RS	Fountain
Error correction/detection	No	Yes	Yes	No	Yes	Yes
Full recovery	No	No	Yes	No	Yes	Yes
Net information density (bits/nt)	0.83	0.33	1.14	0.88	0.92	1.57
Realized capacity	45%	18%	62%	48%	50%	86%
Number of oligos	54,898	153,335	4,991	151,000	1,000,000	72,000
Physical density (Pbytes/g)	1.28	2.25	25	–	–	214