


Encoding information into polymers

Martin G. T. A. Rutten¹, Frits W. Vaandrager², Johannes A. A. W. Elemans¹ and Roeland J. M. Nolte¹ *

Abstract | Defined-sequence polymers have great potential as durable and high-density data-storage media. DNA already fulfils this role in nature, using the sequence of its four nucleobases to store genetic information. Synthetic DNA can be used to store binary codes, and it is both more durable and can store information at a much higher density than conventional silicon-based storage systems. Other defined-sequence synthetic polymers have properties that make them even more suitable for data storage, at least in principle, assuming that complete control over their composition, that is, their monomer sequence, can be achieved. This Review addresses the current status of data storage in DNA, proteins and synthetic polymers, with the objective to overcome the problems of current data storage technology.

Written records are crucial for our understanding of past civilizations. They are so important that history is commonly defined as the study of the past as it is described in written documents, with earlier events relegated to the category of prehistory. The main reason why we know so much about certain past civilizations is that they used durable media to store their writings and art. Thus, we learned about old civilizations in Mesopotamia through 5,300-year-old clay tablets from Uruk that have been preserved until today, we learned about the late Shang dynasty (c. 1200–1050 BC) from China through inscriptions on oracle bones and we learned about the Olmec civilization in Mexico through the Cascajal Block, a stone slab with 3,000-year-old writing made of serpentinite¹.

Digital data storage has completely changed the way we write, use and access information, and we now live in what is commonly referred to as the digital world. It is expected that the need for digital information storage will continue to grow, reaching the level of 44 trillion gigabytes in 2020 (REFS^{2–5}). However, current data storage suffers from digital obsolescence: although the bits and bytes of the digital world are eternal, at least in principle, the storage devices are not. They deteriorate over time, usually within a few decades. For instance, memory cards and chips are maintainable for only around 10 years, while standard hard drives are susceptible to magnetic fields, high temperatures and mechanical failures^{6–8}. Decay of the storage media would result in data loss, were it not for efforts to constantly shuffle data between different devices and facilities. The explosion of digital data means there is a constant need to migrate to new technologies, but these are not always backwards compatible⁹. As a result, much of the information that we currently have stored on floppy disks, tapes, CD-ROMS, spinning hard drives and flash memory will soon be lost

forever — and the challenges do not stop there. Current storage technologies require considerable space and enormous amounts of energy¹⁰. The world data centres currently consume annually ca. 420 terawatt hours of electricity, which is, for comparison, higher than the annual total energy consumption of the United Kingdom (300 terawatt hours)¹¹. It is clear, then, that new methods of writing and storing of information are required.

As alternatives to silicon-based devices, polymers show great potential for data storage because they are stable (at least the synthetic ones) and energy efficient and offer the possibility of high storage densities^{4,12–14}. Polymers are large macromolecules composed of many repeating units, that is, monomers. The most well-known polymers are synthetic plastics, such as polyethylene and polystyrene, and natural biopolymers, such as DNA and proteins, which are essential for all biological processes. In theory, polymers offer the intriguing possibility to durably store all the data of the world in only a handful of material, which could then be safely preserved in some cave or bunker on Earth or perhaps even on Mars!

Here, we review the current status of attempts to store data in natural and synthetic polymers. We focus on fundamental aspects, as the field has not yet sufficiently developed for practical applications to be possible. Current experiments, however, are promising and show great potential for the near future. We begin by discussing the units of information — bits and bytes — before outlining the basic principles and strategies to encode information in DNA. The potential for DNA-based computation, which has attracted a great deal of attention because of the potential to perform parallel calculations, is examined¹⁵. In addition, the potential of proteins as storage systems is briefly reviewed. The final section of this Review focuses on the most recent developments in

¹Institute for Molecules and Materials, Radboud University, Nijmegen, Netherlands.

²Institute for Computing and Information Sciences, Department of Software Science, Radboud University, Nijmegen, Netherlands.

*e-mail: R.Nolte@science.ru.nl

<https://doi.org/10.1038/s41570-018-0051-5>

alternative information storage, in particular synthetic polymers for both data storage and computation.

For the purposes of this Review, data are simply viewed as a sequence of bits, that is, a row of 0s and 1s. We do not care whether this sequence represents a text file, an audio file, a movie, a selection of files (so-called tar file) or something else, nor do we care whether the data are compressed and/or encrypted. We are interested in technology that can reliably store a sequence of bits in a polymer and, at some later point, reliably extract exactly the same sequence from the polymer again.

DNA as storage medium

DNA holds, in the form of a quaternary code (a specific sequence of four nucleobases), the information for the reproduction of a species in nature. DNA has several properties that make it a convenient medium for data storage. It is relatively robust, and the tools to write and read information — DNA synthesis and sequencing, respectively — are available and well understood. Currently, the synthesis of DNA is carried out using oligonucleotide arrays, enabling the synthesis of large libraries of DNA strands in parallel¹⁶.

The reading of DNA (DNA sequencing) has seen tremendous developments over the past 40 years¹⁷. For a long time, since its development in the mid-1970s, sequencing was achieved by methods developed by Sanger–Coulson¹⁸ and Maxam–Gilbert¹⁹. Both approaches are based on the division of a long DNA strand into different sections based on the incorporation of modified bases during copying. Increased demand for low-cost and rapid sequencing of large genomes motivated the development of alternative approaches, which began to take shape throughout the 1980s and 1990s, but these superseded the conventional methods only after completion of the human genome project in 2004. Massively parallel or next-generation sequencing (NGS), as these methods have become known¹⁷, allow for a much faster nucleobase readout by analysing in parallel large amounts of small DNA fragments immobilized on two-dimensional surfaces, using fluorescence-based detection and automated analysis¹⁷. The drawback of nearly all aforementioned methods, however, is that they require DNA template amplification, which is intrinsically prone to copying errors and information loss. To eliminate these deficiencies, fundamentally different approaches, which are based on reading sequences at the single-molecule level, are currently under active exploration. These new methods, also referred to as third-generation sequencing, allow longer reads and higher sequencing speeds and make use of smaller and often portable equipment¹⁷. In particular, nanopore sequencing, which monitors modulations in ion current that occur when a DNA molecule translocates a narrow (protein) channel and translates them into the primary sequence of the strand, is a revolutionary advance that has been commercialized recently¹⁷.

In addition to reading and writing, we also have the possibility to copy (PCR method), cut (with restriction endonucleases) and paste DNA (with DNA ligases), just as we would text in a word processing program^{20–25}. In addition, a DNA-based storage system is expected to be

~1 million times more energy efficient than the systems present in current computers, making it eco-friendly compared with energy-consuming data centres^{25–28}. It should be noted, however, that much of the energy consumed by data centres is used for writing, reading and copying, rather than for the data storage itself. The greatest advantage, though, is the data-storage density, which is much higher than that of conventional methods, overcoming the problem of limited space in which to store all our data. Currently, the largest magnetic hard drive has a capacity of 14 terabytes²⁹ — close to 1 TB in⁻², the predicted limit for this technology^{30,31}. By contrast, for DNA, the maximum storage density is 2 bits per nucleotide; hence, a much larger storage density of 455 exabytes (455×10^6 TB) per gram of single-stranded DNA can be achieved. This means that all the information produced in the world over 1 year could be stored in 4 g of DNA^{7,27,32}.

Encoding data in DNA. To store data in DNA, it must first be converted into a DNA sequence by a translational code. This code should be unambiguous and ideally also incorporate some method of error identification and correction. It is important to recognize that each DNA strand must contain, in addition to the data, a forward and reverse primer sequence at the beginning and end of the strand, which is necessary for DNA replication and reading (sequencing).

Several criteria are important in the design of an encoding algorithm for DNA. First of all, it should make efficient use of DNA. Although synthetic DNA is becoming less expensive to produce, the synthesis of long strands of DNA is still relatively expensive³³. The Shannon information capacity gives an upper bound on how much information can be stored in one unit of the code³⁴. The information capacity of DNA is, at most, 2 bits per nucleotide³⁴ — each of the four bases can encode 2 bits, for example, A = 00, C = 01, G = 10 and T = 11. However, this theoretical capacity is limited by several factors. First, because a G•C base pair has three H-bonds and an A•T base pair only two, the former requires more energy to break. Different double-stranded DNA sequences therefore have different melting temperatures depending on their A•T:G•C ratio, resulting in less efficient PCR amplification. Another difficulty is the occurrence of homopolymer runs (runs of two or more identical bases), which are associated with higher error rates during sequencing^{35,36}. These factors limit the storage capacity because not every nucleotide can be placed at every position. Even in the absence of homopolymer sequences, DNA replication and sequencing are prone to error, which leads to data corruption. To prevent this data corruption problem, multiple copies of the DNA strand are often used, with a resulting decrease in storage capacity. Conceptually, DNA storage can be viewed as a communication channel: we transmit information over the channel by synthesizing DNA strands and receive information by sequencing strands and then decoding the data. The channel is noisy owing to various types of errors, as explained above. Information theory, as developed by Shannon, defines the notion of capacity for a noisy channel and provides a mathematical model

by which one can compute it³⁷. This channel capacity provides a tight upper bound on the rate at which information can be reliably transmitted. Unlike classical information theory, in which noise is independently distributed, the error pattern in DNA data storage is heavily dependent on the input sequence. Nevertheless, Erlich et al., after combining the expected dropout rates and bar-coding demand, succeeded to derive an overall Shannon information capacity of 1.57 bits per nucleotide for a range of practical architectures for DNA storage devices³⁴.

A second important aspect in the design of an encoding algorithm for DNA is to use a code that allows easy and straightforward data retrieval. One aspect that increases the complexity of this problem is the impossibility of synthesizing arbitrarily long DNA strands, making it impossible to create one strand containing all the data. Instead, the data must be divided into multiple smaller fragments that each encode a part of the entire sequence. Aligning the fragments allows the retrieval of all data but also requires that the decoder knows what the order of all fragments should be. One option is to begin every DNA strand with a sequence that counts upwards, before the actual message starts; for instance, the first fragment has the binary code 00001, the second 00010, and so on. An alternative is to use an encoding strategy in which each new fragment also encodes part of the previous fragment; the direction then becomes clear by aligning the repeats³⁸. Another method, developed by Bancroft et al., requires the use of two different DNA classes: one containing the data and the other containing the polyprimer key (PPK)³⁹. In this strategy, every DNA strand containing data is composed of not only the data but also a unique sequencing primer (FIG. 1a). The PPK contains all sequencing primers in the correct order, holding the exact direction to align all DNA strands³⁹. Bancroft et al. used this model to encode and decode the opening line of a novel³⁹.

Error correction. Both DNA synthesis and sequencing are highly prone to error^{16,36}. In addition, mutations may occur during storage. Error correction is therefore an important goal in DNA-based data storage. The simplest method for data correction is to include multiple copies of the same message — in this case, multiple DNA strands with the same sequence. This allows for error correction by aligning and comparing the DNA sequences⁴⁰. The correct sequence can be retrieved using the regions conserved between the different strands. In order to reduce the computational power needed to align all the sequences, smart algorithms have been developed^{40,41}.

More recently, error correction codes used in computer technology have been adapted for data storage in DNA, one of which is XOR encoding. XOR encoding uses an exclusive-or operator for error protection⁴². Two bit sequences, named A and B, can together compose a third bit sequence: the exclusive-or $A \oplus B$. The exclusive-or compares the binary inputs of sequences A and B and gives an output of 0 or 1 based on the bits of strands A and B. The output of the XOR sequence is 0 if both bits of A and B are identical, whereas the output is 1 if both bits are different, for example:

$1110 \oplus 1001 = 0111$. The exclusive-or DNA strand also includes the addresses of the two input strands to clarify from which strands the XOR was taken (FIG. 1b). This encoding strategy gives overall three strands and allows failure of one of these strands, as only two out of three strands are needed to reconstruct the third. This encoding system enables error correction but also allows higher density information storage than multiple sequence alignment, which requires multiple copies of the same file.

A second error correction method adapted from computer technology is the use of Reed–Solomon codes, which were first introduced in 1960 and are applied in CD and DVD devices⁴³. The exact mathematical basis of this correction method goes beyond the scope of this Review but is briefly described here and illustrated in FIG. 1c. In principle, Reed–Solomon codes can detect and correct multiple symbol errors by the addition of parity symbols to the data. The latter symbols are calculated from the original data, which are divided into multiple pieces, for example, y_1 – y_4 . Each piece of data is given a coordinate point (x, y) that defines its location (x value) and the data (y value). The points are then fitted to a polynomial function $P(x)$, which is used to create the parity symbols. The parity symbols are additional data points that correspond to the original DNA sequence and are stored alongside the data. When some of the original data are lost, the remaining data points and parity symbols can be used to reconstruct the original polynomial function. Once the function is recovered, the original data points can be recalculated, and the data can be restored^{44,45}. The above-mentioned error correction methods require the use of extra nucleotides, which is often taken for granted. Some encoding strategies, however, have an inbuilt error correction, as outlined below.

Natural DNA encodes the amino acid sequence of proteins with each sequence of three nucleobases defining one of the canonical amino acids. In 1997, Doig pointed out that the coding efficiency of DNA (amount of nucleotides per amino acid) could be greatly improved if the codon length was varied⁴⁶. This strategy assigned a shorter codon length to more frequently occurring amino acids, whereas rare amino acids received a longer codon length⁴⁶. A similar encoding strategy was later applied in the storage of text files through the use of Huffman encodings⁴⁷. A Huffman encoding is a commonly used method for lossless data compression, in which the most frequently used letter gets the shortest code. The Huffman approach generates a compact DNA encoding for text files. Nevertheless, it possesses two major disadvantages. The first is that it is not possible to include any numbers, as the frequency of the numbers would be heavily text-dependent. This problem was solved by Ailenberg and Rotstein, who defined DNA codons for every character on the computer keyboard⁴⁸. A second disadvantage is the absence of a clear pattern. This poses a problem mainly for long-term storage, as the reader, unaware of the meaning, might confuse it with natural DNA and discard the message⁴⁹. The problematic absence of a clear pattern was overcome via the introduction of primers along the DNA chain containing the messages, for example, at every 500 nucleotides of

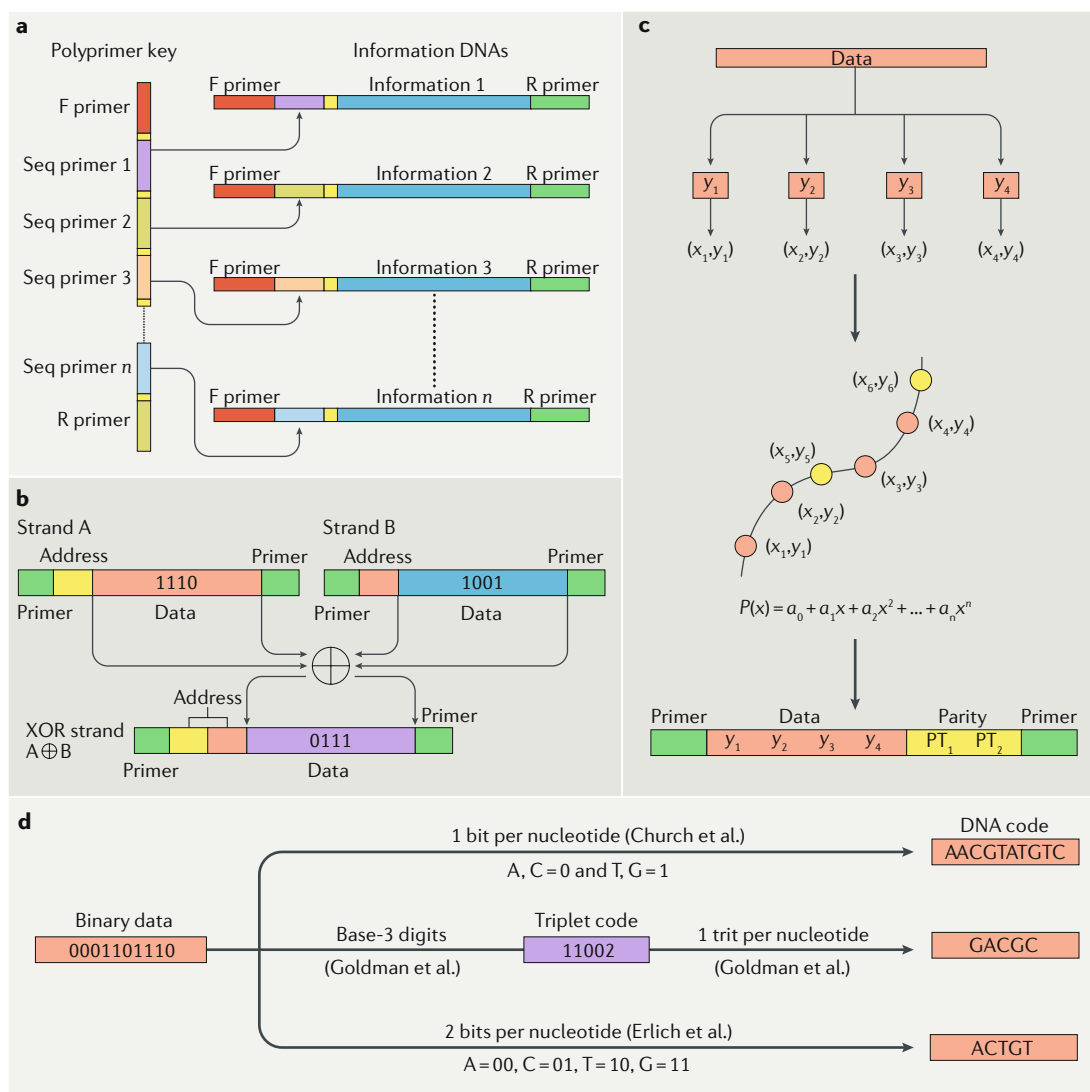


Fig. 1 | Data storage in DNA. a | DNA strands that encode data are composed of a forward (F) and reverse (R) primer flanking the information and a sequence primer. A polyprimer key holds all the sequence primers in the correct order, allowing easy alignment of the fragments during sequencing. **b** | Principle of XOR encoding, in which two strands (A and B) compose a third 'exclusive-or' strand by using an XOR operator, which compares the binary data on both strands. **c** | Principle of Reed–Solomon correction, in which data are split into fragments y_1 – y_4 . Each fragment is given an x coordinate to create coordinate points (x, y) (pink), which are used to generate a polynomial curve and function. Extra data points (yellow) are calculated as parity from the polynomial function. These parity points (PTs) are stored alongside the data and can be used to reconstruct the original polynomial function in case the original data points are lost. **d** | Different encoding strategies for the storage of binary data in DNA. Church et al. encoded binary data using a 1 bit per nucleotide strategy⁵¹. Goldman et al. converted binary data into base-3 digits, which were converted into DNA strands using a 1 nucleotide per trinary digit (trit) strategy⁵⁰. The exact nucleotide depended on the trit and the previous nucleotide⁵⁰. Erlich et al. used a 2 bit per nucleotide storage strategy by employing the maximal Shannon capacity of DNA³⁴.

data. This created a pattern comparable to the intron (primer) and exon (data) structure of natural DNA, allowing for easy pattern identification. The presence of a clear pattern also makes it possible to recognize mutations (errors) in the chain, which cause a shift in the reading frame⁴⁸.

Conversion of bits into nucleotides. The most obvious way to store binary data in DNA is by directly assigning 2 bits to every nucleotide (A = 00, C = 01, G = 10 and T = 11); this creates four different states (0, 1, 2 and 3) instead of two, achieving the maximal Shannon capacity,

as mentioned earlier. Storing 2 bits per nucleotide in this way makes optimal use of the four bases of DNA but offers no protection to errors^{34,35}. A ternary code (0, 1 and 2) can be used instead of a quaternary code to prevent the synthesis of homosequences, which can result in a high error rate during sequencing^{35,36}. In a ternary code, every nucleotide depends not only on the trinary digit (trit) but also on the previous nucleotide, preventing the occurrence of two identical consecutive nucleotides³⁰. Another variant to this code, used by Church et al.⁵¹, stores only 1 bit per nucleotide (for instance A, C = 0 and T, G = 1). However, this direct conversion offers no

protection against any form of errors and thus has to be used in combination with another error correction method, as discussed above.

Apart from the above-mentioned direct conversion methods, in the past, a number of other variants have been proposed, including the comma code, the comma-free code and the alternating code. These variants were specifically designed to encode words in a text and cannot be used for the encoding of other forms of data, for example, pictures or movies. This limited applicability makes it so that they are rarely used nowadays^{49,52}.

Storage in DNA in practice. In 1996, Davis was one of the first to store a message in DNA by encoding a binary file, representing a single image⁵³. The encoding scheme used, however, was inaccurate, as it made no distinction between a 0 and a 1. The four DNA bases were used to determine only how large a repeat of 0s and 1s was (C = 1, T = 2, A = 3 and G = 4), for instance, 100111 was encoded as CTA but could be decoded as either 100111 (correctly) or 011000 (incorrectly)⁵⁴. Two years later, the Genesis project was started by Eduardo Kac²⁷. One sentence was encoded in a two-step process, converting first into Morse code and subsequently into DNA (C = dot, T = dash, A = word space and G = letter space). The sentence was synthesized as a gene and fused into bacteria. It was found, however, that ultraviolet light caused mutations and altered the message²⁷. Another earlier study attempted the storage of a 23 character-long message hidden in a DNA microdot⁵⁴. A unique feature of this attempt was the use of two primers flanking the DNA code, which enabled the use of PCR to amplify the stored message.

A remarkable early study of large-scale data storage in DNA was performed by Church et al., who encoded the draft version of an entire book, including 53,426 words, 11 images and 1 JavaScript program⁵¹. Instead of constructing one long strand, several smaller fragments were made, together encoding the entire binary file. All the data were first converted into bits, which were then translated to DNA nucleotides using a simple encoding strategy of 1 bit per base (A, C = 0 and T, G = 1) (FIG. 1d). The entire sequence was split into non-overlapping strands. These strands included the data as well as a 19-nucleotide address label composed of bits counting upwards every strand to align all fragments in the correct order. Using this method, Church et al. were able to encode and decode 5.27 megabits, with a total of only 10 errors⁵¹.

Most of the errors encountered by Church et al. were caused by homopolymer runs and lack of coverage. To improve on this achievement, Goldman et al. added redundancy to the encoding scheme by creating overlapping fragments and were therefore able to encode and decode five files, including a written text, a picture and an audio file⁵⁰. The encoding strategy used a ternary code in combination with the Huffman encoding (see above) to compress the data. The original data were converted into base-3 digits (0, 1 and 2), and every trit was converted into a single nucleotide, in which the exact nucleotide depended on the trit and on the previous nucleotide,

preventing identical consecutive nucleotides and thus homopolymers (FIG. 1d). The obtained sequence was split into several DNA strands, containing data, indexing and one nucleotide to indicate the orientation. A parity check was included as an additional safety measure: it consisted of one nucleotide at the end of each strand and was the sum of the odd-positioned trits. When the message was decoded, the parity trit displayed the sum of odd trits in the original strand, and if an error occurred, this trit should not be in agreement with the actual amount of odd trits. Besides this parity check, overlapping segments were created from 75 nucleotides, meaning that every segment started with an offset of 25 nucleotides from the previous strand, which resulted in a fourfold redundancy⁵⁰. Of the five encoded files, four were recovered without errors. The fifth file contained two gaps of 25 nucleotides, in which none of the four overlapping segments was sequenced. By taking the neighbouring regions into account, the gaps could be manually filled with the missing nucleotides, after which the last file was also decoded successfully⁵⁰. Altogether, a storage density of 2.2×10^6 GB g⁻¹ was achieved^{50,55}.

Thus far, the highest information density has been achieved by Erlich et al., who stored 17.1 megabits of information in DNA oligonucleotides with a density of 1.57 bits per nucleotide (2.15×10^8 GB g⁻¹)³⁴. To realize this high density, an advanced erasure-correcting encoding algorithm was used: a so-called fountain code with Luby transform⁵⁶. The used encoding strategy fragmented the binary sequence into non-overlapping segments, which were randomly combined in a single bit stream, called a droplet, by XOR encoding (see above). An identification tag was added in the form of a seed to identify which segments were combined in the droplet. The droplets (the XOR code and the seed) were converted into a DNA sequence by translating 00, 01, 10 and 11 to A, C, G and T, respectively (FIG. 1d). To prevent the formation of homopolymers, sequences were scanned and should not contain more than three consecutive identical nucleotides, and the GC content should be between 45% and 55%. Invalid sequences were rejected, and droplets were made until 5–10% more fragments were obtained than actually needed to cover the entire sequence. Reed–Solomon codes were used to ensure that completely missing regions could be reconstructed efficiently. Decoding of the file was performed using a message-passing algorithm, which reversed the Luby transform and resulted in complete recovery of the input without errors. Decoding was still possible when the DNA molecules were diluted, which confirmed the robustness of the encoding strategy³⁴.

The above examples provide a proof of concept for data storage in DNA by applying commercial synthesis protocols and standard sequencing techniques. There is, however, much to gain from the development of faster writing and reading procedures. In addition, the material aspects of encoding require further attention, that is, the development of easy-to-use and fast commercial devices, which would allow big data companies to employ DNA storage as an alternative to silicon-based devices in the future. These goals provide challenging tasks for synthetic chemists and materials scientists.

Rewritable and random-access DNA storage. One of the major drawbacks of data storage in DNA is the time it takes to find and read the data, compared with the time needed in silicon-based devices. The reading rate of polymerase enzymes (~ 100 nucleotides s^{-1}) is approximately seven orders of magnitude slower than that of conventional hard drives (~ 10 Gbits s^{-1})^{38,57}. Furthermore, in the DNA data storage systems described above, it is necessary to decode the entire sequence in order to find a specific set of bases. In addition, rewriting of the stored data is problematic. The methods discussed thus far store the data in a read-only format, making it difficult to apply these systems to data storage that is subjected to change or needs frequent updates. Researchers have tried to overcome the two major drawbacks of DNA data storage, that is, random access and rewritability. The first problem can be addressed by using a barcode to store specific data in specific wells or pools³⁵. These DNA pools hold a random selection of different DNA strands, with each DNA strand containing an address label. When a specific data file is required, this strategy allows selection of the pool containing the desired data before decoding, thus limiting the number of DNA strands that need to be sequenced³⁵.

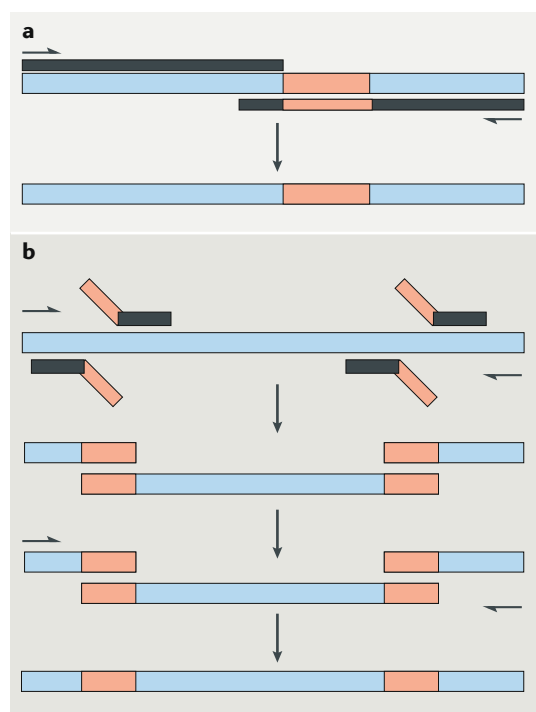


Fig. 2 | The DNA editing methods gBlock and overlap extension PCR. **a** | The gBlock methodology was used for short rewrites. A sequence containing the edited part of the fragment was synthesized via gBlock, and the remaining part of the strand was PCR amplified. An overlap of at least 30 nucleotides was present between the two strands in order to combine both strands into one. **b** | By using overlap extension PCR (OE-PCR), different parts of the DNA strand were amplified by PCR using primers with overhang containing the edited parts. All the different parts of the strands were finally combined into one strand using the overlap between the different segments. Adapted from REF.⁶⁴, CC-BY-4.0.

The second drawback, rewritability, might be addressed by the use of specific enzymes that invert and restore specific DNA sequences^{58,59}. In addition to the use of enzymes, chemical transformations of DNA bases, such as the selective modification of cytosine to uracil, have been employed^{60–63}. An interesting DNA-based coding system that allowed random access and rewritability was developed by Yazdi et al.⁶⁴. This storage system contained only written text stored in 1,000-nucleotide strands, including specialized address strings that could be used for selective information access. The encoding strategy used codons of 21 nucleotides, in which each codon corresponded to a single word. This fixed codon length was designed to make rewriting as easy as possible and to prevent propagation errors. Rewriting was made possible by two DNA editing techniques: gBlock⁶⁴ and overlap extension PCR (OE-PCR)⁶⁵. The gBlock method was used for short rewrites, in which a short section of the new strand containing the edit was synthesized by the gBlock methodology, while the remainder of the old 1,000-nucleotide strand was amplified by PCR. The new and old strands have a designed overlap of at least 30 base pairs, enabling combination of the two strands⁶⁴ (FIG. 2a). The gBlock method is very efficient but also requires the use of long and thus expensive primers. OE-PCR is more cost efficient and was used for the rewriting of longer blocks. Using OE-PCR, rewriting was performed in steps with short primers that contained the edits as overhangs (FIG. 2b). PCR was used to amplify all parts of the strand, which could then be combined by the overlap between the parts^{64,65} (FIG. 2b). If the segment to be rewritten was longer than 1,000 base pairs, entirely new strands were synthesized. The introduction of this DNA editing technique and the use of address strings allowed Yazdi et al. to select specific sequences and edit them successfully⁶⁴.

Storage of DNA. The way data encoded DNA strands are stored largely depends on the purpose of the data system; nonetheless, there are some generalities. DNA can be stored on a solid support, in which one end of the double-stranded DNA is immobilized, reducing the risk of unwanted strand aggregation⁶⁶. Storage in solution is also possible, however. The latter enables more rapid replication and sequencing, as the molecules are more flexible and more easily accessible. Furthermore, it allows for autonomous information processing and the possibility to store encoded DNA in microorganisms^{27,38}.

In solution at 4 °C, DNA decays within weeks. This is improved to a 3–5-year span at –80 °C in the solid state⁶⁷, but other options are essential for long-term storage. One possibility is to store the DNA in living microorganisms, as they can be chosen to withstand extreme conditions and the data can be retrieved after a long time^{38,68}. Data may even be stored in several different microorganisms to secure the highest possible chance of data recovery. Yachi and co-workers were the first to attempt to store data in bacteria by encoding the equation $E = mc^2$ in the genomic DNA of *Bacillus subtilis*⁶⁹. In later research, *Escherichia coli* was used as a storage device, affording a storage capacity of 1 kilobyte per cell⁷⁰. More recently, Church et al. used the CRISPR–Cas

gene editing technique to store a digital movie in bacterial DNA⁷¹. Although mutations occur in the genome of bacteria, the rate and amount should be sufficiently low to allow correct data retrieval⁷². In addition, data should always be stored in colonies of bacteria, providing many bacteria containing copies of the data and many data strands within all the bacteria.

Grass et al. have explored the use of synthetic silica matrices to store DNA⁷³. The use of such an inorganic material separates the DNA from the environment and thereby the effect of humidity from the storage environment. Besides protection against humidity, silica also offers protection against reactive oxygen species. Accelerated ageing experiments revealed that data could be correctly recovered after treating the DNA in silica at 70 °C for one week, equivalent to 2,000 years in central Europe or over 2 million years at the Global Seed Vault (−18 °C)⁷³.

DNA computation

Going beyond simple data storage, DNA can also be used to build synthetic biological circuits, akin to electric circuits. Biological circuits can be used, with the help of molecular biology, to solve computational problems. In order to do so, the computational problem must first be translated into biological terms, that is, DNA. The easy modification, amplification and stability of DNA molecules make them suitable for engineering circuits. Furthermore, DNA computation is energy efficient and allows parallel computations in the form of chemical reactions to be performed^{74,75}. As early as 1994, Adleman used DNA as a tool to solve a Hamiltonian path problem: the Travelling Salesman problem²⁸. This famous mathematical problem attempts to find the shortest route starting at any point for the eponymous salesman to visit several cities, once each (FIG. 3a). Computational solutions to the problem must consider all possible paths that visit all cities (eliminating any that visit the same city multiple times) — a very time-consuming process if each path is considered sequentially. As an alternative, Adleman used DNA computation to solve this problem for a collection of seven cities. Each city was represented by a unique oligomeric strand of 20 nucleotides. Paths between the cities were also represented by 20 nucleotide sequences such that they overlap with the last 10 nucleotides of the starting city and the first 10 nucleotides of the ending city (FIG. 3b). When paths and cities were mixed, the tendency of DNA to form double-stranded helices caused city sequences to combine with the complementary path sequences. The main advantage of DNA is that it can explore all the combinations in parallel, provided there is an excess of city and path strands to make the combinations²⁸. Adleman was able to generate all solutions in a few hours, after which the elimination of invalid paths could begin. Valid solutions should contain seven cities, meaning that any longer or shorter strands could immediately be eliminated. Strands with duplicated cities could also be eliminated, as each city may be visited only once. This need to eliminate invalid combinations immediately shows the major drawback of this DNA-based method: the elimination of all invalid paths took 7 days²⁸.

The work of Adleman and co-workers shows that although DNA can be used to solve computational problems, this approach cannot compete with conventional silicon-based computers. Nonetheless, the potential of parallel computation motivated continued research and development in this area. The Travelling Salesman problem is a classic example of a so-called nondeterministic polynomial time (NP)-complete computational problem (see BOX 1). These are problems for which the fastest known algorithms require a computational time that increases exponentially with the size of the inputs^{76,77}. In 2002, Braich et al. succeeded in solving a non-trivial example of another NP-complete problem, 3-SAT (see BOX 1), on a DNA computer. It involves 20 variables and 24 so-called clauses, leading to more than 1 million (2^{20}) truth assignments that must be checked. Each of the 20 variables was represented by two 15-base-pair sequences (one true, one false), and each of the possible solutions was represented by 300 base pairs (20 variables). The DNA computer itself consisted of an electrophoresis box with two chambers, one loaded with all the DNA sequences and the other containing one clause of the expression, with complementary base pairs for the correct variables. On starting the electrophoresis, strands moved from one chamber to the other, where sequences satisfying the clause would be captured and non-satisfying sequences moved through. Captured sequences from the first clause went through the same process again but now with the second clause of the expression, and so on. Eventually, this resulted in the retrieval of the correct answer, satisfying all clauses of the expression⁷⁷.

Other DNA-based computers have been developed by Shapiro et al., who used DNA and enzymes to solve computational problems autonomously^{78,79}. In these computers, the hardware consisted of enzymes, and the software and input were encoded by double-stranded DNA molecules. The automation process was based on processing the input molecule through a cascade of reaction cycles, producing an output molecule encoding the computational result^{78,79}. In FIG. 3c, an example is given: an enzyme (FokI) cuts a double-stranded DNA molecule containing a string of the letters a and b (each represented by four nucleotides and separated by spacer nucleotides). The resulting DNA molecule again acted as the input software for the enzyme, and this process was continued until the output result was obtained, that is, the answer to the question ‘does the DNA string have an even or odd number of bases?’⁷⁹.

Over the years, more techniques and tools have been developed to incorporate biology into the engineering of circuits. Developments include the design of a ring oscillator^{80,81} and DNA-based transistors⁸². Analogue computation was also shown to be possible using three different transcription factors to construct two cellular circuits, which could detect and compute compounds outside the cell⁸³. Recently, Lu et al. devised a way to combine data storage and circuit engineering by designing cells that express single-stranded DNA, induced by a chemical or light stimulus⁸⁴. These DNA strands were targeted to the genome, thereby converting cellular signals into DNA-encoded memory⁸⁴. Keinan et al. developed a more

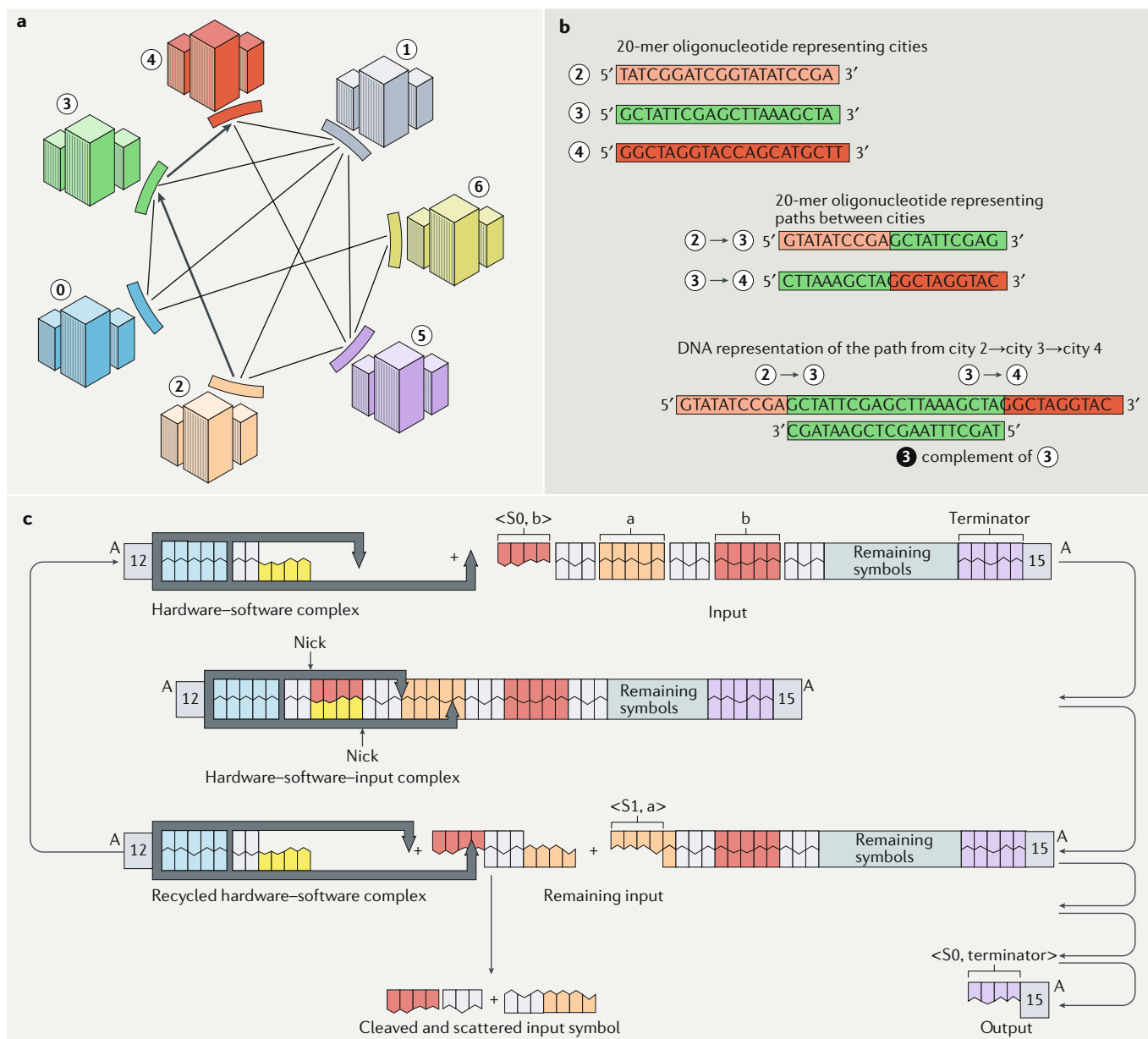


Fig. 3 | DNA computing. **a** | Graphical representation of the Hamiltonian path problem: the Travelling Salesman. The objective is to find the shortest route for a salesman who has to visit seven cities and visit them only once. **b** | DNA representation of the Travelling Salesman problem. Cities are represented by oligonucleotides of 20 base pairs, and the paths between the cities are also represented by 20 base pairs — 10 base pairs for the starting city and 10 base pairs for the ending city. The inclination of DNA to form double-stranded helices led to all nucleotide possible combinations, from which the correct solution can be derived. **c** | Autonomous DNA computing as performed by Benenson and Shapiro. An enzyme (FokI, grey rectangle with arms) capable of snipping a double DNA strand at positions 9 and 13 downstream of the DNA recognition site acts as the hardware. It binds an input DNA molecule (A, called software) and cleaves it, and the remainder of molecule A again acts as the input, and so on, eventually leading to an output result. The so-called machine states of the system are denoted S0 and S1. In the present case, the DNA machine was asked to calculate whether a DNA string containing two letters a (orange) and b (red) separated by spacer nucleotides (blue) contains an even or odd number of bases. Parts **a** and **b** are adapted with permission from REF.⁷⁴, Wiley-VCH. Part **c** is adapted with permission from REF.⁷⁹, PNAS.

complex DNA computer, capable of iterative computation, that is, using the output of one computation for a secondary computation process, and so on. This DNA computer used DNA plasmids as input and processed them using a predetermined algorithm. The output was written on the same plasmid used for the input, which could be further processed. In addition to the possibility

of iterative computation, this DNA computer produced biologically relevant results, opening ways to regulate and change biomolecular processes^{85,86}.

The abovementioned developments show that DNA has potential not only as a data storing device but also as a computer. The main drawback of DNA computation, however, lies in the extraction of the data,

Box 1 | The 3-SAT problem

Nondeterministic polynomial time, or NP, is a complexity class in computational complexity theory used to describe certain types of decision problems. In simple terms, NP is the set of all decision problems for which the instances in which the answer is yes have efficiently verifiable proofs. Within the class NP, NP-complete problems are the most difficult (in a precise mathematical sense) to deal with. The input for the so-called 3-SAT problem (an NP-complete problem) consists of a Boolean logical formula. Boolean logic (named after the 19th century mathematician George Boole) is a form of algebra in which all values of variables are reduced to either true or false. Boolean logical formulas are composed of variables x , y and z and connectives such as \wedge (and), \vee (or) and \neg (not). An example of a formula ϕ is $(x \vee \neg y) \wedge (\neg x \vee y)$. A truth assignment for a formula specifies for each variable whether it is true or false. A formula is satisfiable if there is a truth assignment that causes the formula to evaluate to true. For example, formula ϕ is satisfiable, as it evaluates to true under the truth assignment in which both x and y are true. The formula ψ given by $(x \vee y) \wedge (x \vee \neg y) \wedge \neg x$ is an example of a formula that is not satisfiable. The 3-SAT problem asks for the satisfiability of formulas that are required to have a particular structure. A literal is either a variable or the negation of a variable and a clause is a disjunction of literals (or a single literal). The input of the 3-SAT problem is a Boolean formula that is a conjunction of clauses in which each clause is limited to at most three literals (this represents the 3 in 3-SAT). The above formulas ϕ and ψ are examples of valid inputs for 3-SAT¹⁷⁴.

which still takes a large amount of time compared with silicon-based devices.

Data storage in proteins

Most research on alternative data storage has evolved around DNA. However, DNA is not the only molecule that is suitable for storing information. Proteins, being natural polymers composed of amino acids, also have the potential to act as storage devices. For use in data storage, the main focus has been on photoswitchable proteins, in which the specific state of the protein represents a binary 0 or 1.

Hirshberg et al. were the first to propose a photochemical memory model, based on colour transformation, triggered by absorption of a photon⁸⁷. New possibilities were opened with the discovery of photoconvertible fluorescence proteins and photoswitchable fluorescent proteins, which included Kaede, Dronpa and EosFP, in which the bits 0 and 1 were represented by the colours green and red, respectively^{88–90}. Using IrisFP (a mutant of EosFP), colour switching between red and green could be combined with switching between a dark and bright state^{91,92}. Another protein used for data storage is bacteriorhodopsin (bR), a light-activated protein from the membrane of the microorganism *Halobacterium salinarum*. Upon irradiation, light is converted into chemical energy, which sets the molecule into an intermediate state for a maximum of a few days⁹³. For data storage purposes, the protein was modified such that this intermediate state was stable for a few years⁹⁴. Binary values 0 and 1 were represented by the bright state and the dark state of the protein, respectively. Encoding was performed using a laser with a specific wavelength to set the protein into a shape representing a 0. A laser of another wavelength was used to convert the protein into a shape representing a 1. For reading, a low-power laser beam was used to detect the conformation of the protein without disturbing the conformation itself. The ability of bR to shift between different states also allows for rewritable data storage^{95,96}.

Rotaxane

A mechanically interlocked molecular architecture in which a macrocycle is kinetically trapped on a thread by the presence of two large 'stoppers'.

Writing information into proteins is far less efficient than writing information into DNA and, as it stands, does not allow for the storage of large amounts of data. It is doubtful, therefore, whether this type of approach to data storage will have a future.

Storage in synthetic polymers

Synthesis. In addition to DNA and proteins, synthetic polymers are also suitable for data storage, at least in principle. As early as 1986, Richard Dawkins suggested that, at least in theory, any polymer composed of at least two different monomers could be used to store data⁹⁷. Although the controlled synthesis of polymers with more than two monomers is possible — and although such polymers would potentially provide a more economical solution for data storage — most data-encoding polymers employ only two different monomers (directly representing 0 and 1 in the binary code). The main advantages of synthetic polymers are the possibility of having full control over their synthesis and the greater flexibility, meaning that one is no longer restricted to four monomers, as in the case of DNA. Instead, the monomers can be selected and tuned for the purpose of the application. In such synthetic data-encoding copolymers, it is essential to achieve perfect control over the monomer sequence. For example, DNA can be used as a template for the assembly of free nucleotides (including non-natural ones) before chemical or enzymatic polymerization^{98–103}. The drawbacks of this approach are the low efficiency and difficulty of removing the synthesized polymer from the template. Recently, molecular machines that mimic biological polymerization have been developed, such as the peptide synthesis machine designed by Leigh and co-workers¹⁰⁴. The machine is a rotaxane system in which the macrocycle sequentially picks up amino acids from the thread to assemble a peptide of known sequence. The current rates and yields of these reactions make practical applications difficult — and thus far, these systems are limited to the synthesis of natural polymers, that is, polypeptides¹⁰⁴. To work around biological polymerization techniques, Liu and co-workers designed a DNA translation system to synthesize sequence-controlled polymers not based on natural monomers¹⁰⁵. The polymerization in this case depends on the hybridization of DNA base pairs to a template. Synthetic building blocks are attached to these DNA base pairs via a cleavable linker. In this system, the DNA base pairs perform a very similar function to tRNA, that is, they serve to bring together the desired building blocks in the correct order. After polymerization, cleavage of the linker results in the release of the synthetic polymer¹⁰⁵.

Complete chemical polymerization has the advantage of a much wider range of available building blocks, but achieving perfect sequence control remains challenging because classical chain-growth and step-growth polymerizations do not allow for precise control over the monomer position. Sequence control can be improved by using living chain polymerization methods, in which each polymer chain grows in a more uniform way. For instance, by using controlled radical polymerization techniques, that is, reversible addition–fragmentation chain-transfer (RAFT) polymerization, Houshyar et al.

Monodisperse polymers

Polymers composed of uniform molecules with the same structure and mass. Naturally occurring polymers are frequently monodisperse, while synthetic polymers are usually not.

constructed new sequence-controlled macro-RAFT agents by inserting two monomers in a sequential manner¹⁰⁶. The low yield of the monomer insertion, however, made this process suboptimal for the synthesis of long chains. Atom-transfer radical polymerization (ATRP) was used by Tong et al. to construct vinyl polymers by an iterative process of single monomer addition¹⁰⁷. However, this method also suffered from low yields (as a result of side reactions), making the synthesis of long chains impossible.

In addition to controlled radical polymerization, sequence control in chain polymerization can also be obtained by using living ionic polymerization techniques. Living cationic polymerization was used by Minoda et al. to create sequence-regulated polymers by the addition of monomers one by one in order of decreasing reactivity¹⁰⁸. Nevertheless, defects occurred during the polymerization, necessitating purification after each monomer addition. Living anionic polymerization has also been used to obtain sequence-controlled chains composed of two different monomers (the choice of one bulky monomer prevented homopolymerization). In this way, an alternating pattern of two different monomers could be obtained¹⁰⁹. This type of kinetic control was also applied in radical polymerization to obtain an alternating pattern, that is, the high affinity between two different monomers was used to create regions with a specific sequence in the polymer chain¹¹⁰. Lutz et al. built upon this alternating method by tuning the sequence through time-controlled monomer addition^{111,112}. In this strategy, one monomer (an electron-rich donor) present in excess is polymerized by a radical reaction, while a second monomer (an acceptor) is added in small amounts at specified times. A highly favourable donor–acceptor interaction between the two monomers results in the incorporation of acceptor monomers in small, well-defined regions of the polymer backbone^{111,112}. Using automated protocols, Lutz et al. were able to construct well-defined polymer chains containing up to eight precisely positioned blocks with a specific sequence¹¹³. A similar approach was adopted by O'Reilly et al. in ring-opening metathesis polymerization, in which the position of four different functional moieties could be relatively well controlled along a growing polymer chain¹¹⁴. However, in all these methods, variations in the length and precise composition of each segment may occur, and some polymer chains might contain defects^{101,115}.

To minimize the number of defects, long building blocks in multiblock copolymerization may be used. Gody et al. employed this strategy in combination with degenerative transfer radical polymerization to construct well-defined multiblock copolymers¹¹⁶. In addition, Engeliš et al. used long blocks for the synthesis of well-defined multiblock copolymers by emulsion polymerization, in which monomers and catalysts were separated in micelles to isolate the growing polymers from one another and reduce unwanted side reactions¹¹⁷.

Despite the reported improvements in the synthetic procedures, it can be concluded that chain polymerization always results in polymers with deviations in chain length and composition¹¹⁵. This limitation

means that chain-growth polymerization is, at present, not a good method to prepare well-defined polymer sequences. Nonetheless, it could be employed for easy copying of already synthesized sequences by template polymerization¹¹⁸.

In addition to chain-growth polymerization, step-growth polymerization techniques can be used to synthesize polymer sequences with periodic monomer patterns. Conventional step-growth polymerization has been used for the synthesis of polyamides and polyurethanes. Although these methods are relatively straightforward, they do not allow for perfect sequence control. New step-growth polymerization techniques using radical polymerization^{119,120} or click chemistry¹²¹, however, do allow for such a sequence-controlled polymerization. The latter can also be achieved by applying multistep-growth synthesis, which involves the stepwise chemical attachment of monomers attached to a support¹²². This procedure results in highly monodisperse polymers. One such method is iterative solid-phase synthesis, similar to the well-known solid-phase peptide synthesis methodology. It employs an insoluble support on which the polymers are grown by the stepwise addition of monomers¹²³. The method is highly efficient but also very time consuming, and furthermore, the efficiency of the coupling steps decreases with increasing polymer length, making it best suited to the synthesis of short polymers. Despite these disadvantages, solid-supported synthesis remains the most frequently applied and most reliable method for the synthesis of sequence-controlled polymers¹²². An alternative is the use of a soluble polymer chain as a support¹²⁴, which improves the process efficiency, but the synthesis of long sequences is still not possible¹⁰¹. The Lutz group has investigated numerous strategies to exploit this multistep-growth methodology for the production of sequence-controlled polymers, including those encoding data¹²⁵. The previously described step-growth synthesis of polyurethanes, for instance, could be improved by applying a multistep-growth approach¹²⁶. This strategy relied on a sequence of two chemoselective steps: the reaction of an alcohol with an *N*-hydroxysuccinimide (NHS) moiety and then the reaction of an amine with NHS¹²⁶. Data were encoded using two different amino alcohol monomers (serving as 0 and 1), while *N,N'*-disuccinimidyl carbonate, containing two NHS moieties, was used as a linker¹²⁶. Another method developed by Lutz and co-workers is based on phosphoramidite coupling, a method already widely used for oligonucleotide synthesis¹²⁷. The synthesis uses a solid support, and the monomers are coupled one by one in three steps (FIG. 4a). First, *N,N*-dimethyltryptamine (DMT) deprotection of the monomer occurs, allowing the connection of the next monomer by phosphoramidite coupling, followed by oxidation of the phosphite to a phosphate. Optimization of this method allows each three-step cycle to be completed within a few minutes¹²⁸. Lutz et al. used this approach to synthesize a polymer with a controlled sequence from two monomers containing either a propyl moiety (representing 0) or a 2,2-dimethylpropyl moiety (representing 1)¹²⁹. In addition, another monomer containing a 2,2-dipropargylpropyl group (representing 2)

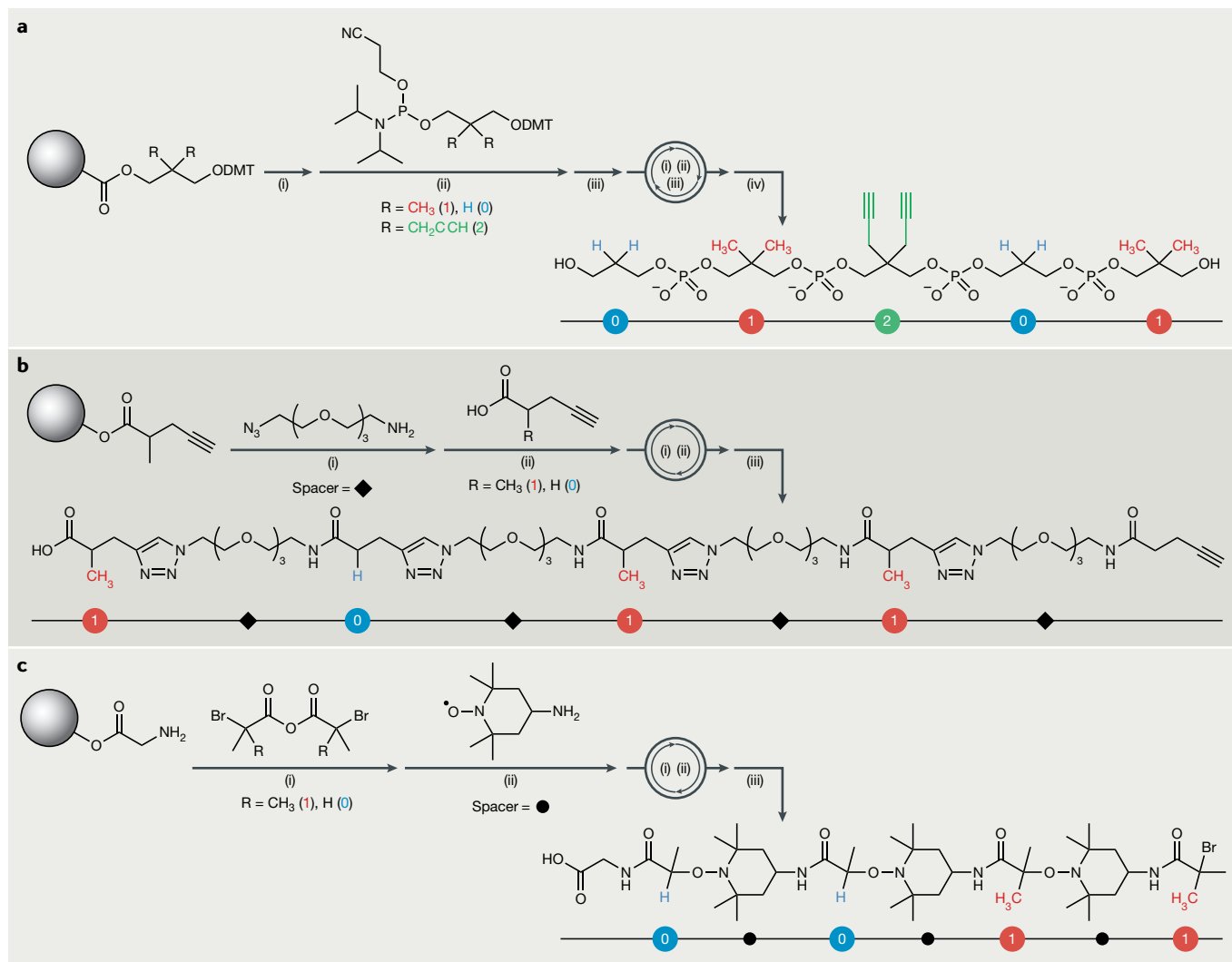


Fig. 4 | **Different strategies for the synthesis of information-containing macromolecules.** **a** | Phosphoramidite coupling was performed in four steps: (i) deprotection with *N,N*-dimethyltryptamine (DMT); (ii) coupling of the next monomer, representing 1, 0 or 2; (iii) oxidation of the phosphite bond to a phosphate; and (iv) cleavage from the resin. **b** | The AB + CD method, involving 3 different monomers representing 0, 1 and a spacer. (i) Coupling of the spacer (CD, black diamond) by an azide–alkyne copper-catalysed cycloaddition, (ii) coupling of a monomer representing either a 0 or a 1 (AB) by amidification and (iii) cleavage from the resin. **c** | Accelerated AB + CD method using, again, 3 different monomers, representing 0, 1 and a spacer. (i) Coupling of a monomer representing 0 or 1 (AB) by an anhydride–amine coupling, (ii) coupling of the spacer (CD, black circle) by a nitroxide radical reaction and (iii) cleavage from the resin. Adapted with permission from REF.¹²⁵, ACS.

was used to investigate the post-polymerization modification of the polymer by a Huisgen azide–alkyne cycloaddition reaction. Using this approach, it was possible to synthesize data-encoding polymers, which could be modified after polymerization¹²⁹. Later research improved on the phosphoramidite coupling method by using an orthogonal iterative approach, in which two different building blocks are linked without the need for protecting groups¹³⁰. Furthermore, the chosen building blocks simplified the readout by tandem mass spectrometry (MS/MS) (see below)¹³⁰. It is important to note that Lutz et al. co-workers have already reported the synthesis of a sequence-encoding polymer using automated phosphoramidite coupling¹³¹. By making some small adjustments to the original protocol — using a large excess of

monomer and applying capping steps — the synthesis and sequencing of polymers composed of more than 100 monomers could be achieved¹³¹.

An alternative solid-phase approach to achieve complete sequence control without the need for protecting groups is the AB + CD method, also developed by Lutz et al.^{124,132}. This approach makes use of two different building blocks, each containing orthogonally reactive functional groups AB (A = carboxylic acid and B = alkyne) and CD (C = amine and D = azide). A protecting group strategy is unnecessary, as A can react only with C (by amide formation), while B can react only with D through a copper-catalysed azide–alkyne cycloaddition. To encode data using this synthetic protocol, Lutz et al. chose two different AB building blocks,

representing 0 and 1, while the CD building block was used as a spacer¹³² (FIG. 4b). To simplify the AB + CD synthesis, four different AB dimers can be used, representing 00, 01, 10 and 11 (REF.¹³³). This reduces the number of coupling steps required to produce byte-encoded macromolecules (although it still remains a time-consuming process). An accelerated AB + CD protocol was also developed, allowing the coupling between monomers to proceed via consecutive anhydride–amine and nitroxide radical reactions (FIG. 4c). Repeating these steps enabled the synthesis of sequence-controlled polymers that were easy to read and easy to erase¹³⁴.

Another strategy to obtain sequence-controlled polymers was reported by Zydziak et al. Rather than rely on solid-supported synthesis to produce monodisperse products, the process couples six different monomers using Diels–Alder reactions and achieves selective coupling through a photochemical unmasking of the requisite diene¹³⁵. Each monomer contains a dienophile and a benzaldehyde moiety, with the latter converted to a diene upon irradiation before undergoing the Diels–Alder reaction¹³⁵.

Reading and rewriting. For the reading (sequencing) of biopolymers such as DNA, very fast and automated methods are available. Unfortunately, these methods are not applicable to synthetic polymers, and other more universal analysis procedures must be used, one of which is MS/MS^{136–138} (FIG. 5a). Here, the polymers to be sequenced are ionized and separated based on their mass-to-charge ratio, followed by further fragmentation, separation and detection. The resulting fragmentation pattern can subsequently be used to reconstruct the precursor ion and ultimately the polymer sequence. The obtained fragmentation pattern is strongly dependent on the nature of the polymer backbone, which gives synthetic polymers an advantage, as their molecular structures can be altered to favour an easy readout¹³⁹. The previously described phosphoramidite coupling, for instance, results in an easy fragmentation pattern, in which the phosphate bonds are easily ionized and dissociated in MS/MS. Introduction of alkoxyamine bonds, which have a lower dissociation energy, along the chain simplified the readout further by introducing two dissociation energies, that is, cleavage of the alkoxyamine bond, generating large fragments and cleavage of the phosphate bonds, generating smaller fragments^{139–141}. The alkoxyamine amide links in Lutz and co-workers' accelerated AB + CD synthesis described above (FIG. 4c) are relatively weak links and enable a fast readout for long chains^{125,134,142}, though this does come at some cost to the thermal stability of the polymers¹⁴³. Charles, Lutz and co-workers also described the MS/MS analysis of poly(triazole amide) sequences. Cleavage can occur at either the amide or the ether bonds, and thus, the pattern can be easily decoded¹⁴⁴. Besides using specific linkers between the monomers to simplify sequencing, post-polymerization modifications can also be used to simplify the readout; for example, the azide–alkyne cycloaddition can be used to specifically modify the side chains of the polymer^{129,145}. In addition to synthetic improvements, progress has been made in the

development of new software, enabling the sequence of polymers to be read in only a few milliseconds. Such rapid reading and decoding highlight one advantage of synthetic (over biological) digital polymers¹⁴⁶.

To elucidate the sequence of larger macromolecules, the electrical birefringence (Kerr effect) of a polymer solution in an electric field can be measured¹⁴⁷. The Kerr coefficient of a polymer depends on changes in the magnitude and/or orientation of the overall dipole moment with respect to its maximum polarizability, enabling the complete characterization of polymers. Although not extensively used for synthetic polymers, it has potential as an interesting technique in the future^{147,148}.

NMR can also be used to sequence a polymer. For some time, ¹³C NMR was one of the most widely used methods to identify short synthetic copolymers; however, as its sensitivity is limited, application to long polymers is challenging. Another NMR technique especially suitable for non-natural polymers is the tweezer technique^{149–151}. This method uses molecular reporters (tweezers) that can bind along the polymer chain through non-covalent interactions such as hydrogen bonding and π - π stacking. The tweezers can shift specific NMR signals, making the spectrum easier to interpret and quantify (FIG. 5b).

A new and promising sequencing technique for both natural and synthetic polymers is nanopore sequencing, which analyses the polymer structure by pulling it through a biological or synthetic pore (FIG. 5c). As the molecule moves through the pore, the channel current changes the current in a way that is characteristic of the polymer sequence and can thus be used to determine the primary structure of the polymer¹⁵². This technique was first introduced by Kasianowicz et al., who used a biological nanopore (α -haemolysin) to sequence a single-stranded DNA molecule¹⁵³. Later, modifications on the surface of the channel showed that the nanopore could be used to identify numerous features, for instance, the 3' and 5' ends of a DNA chain^{152,154,155}. Thus far, only a small number of studies using nanopore sequencing for synthetic polymers have been reported, including polyethylene glycol macromolecules, polystyrene sulfonate, dextran sulfate and poly(phosphodiester)s^{156–160}. More recently, theoretical studies have been performed on more complex polymers, such as branched polymers and heterogeneous copolymers with charged and uncharged blocks^{161,162}. These results show that nanopore sequencing has the potential to be a reliable method in the future but also that successful readout depends strongly on the charge, stiffness and conformation of the polymer chain¹²⁵.

An advantage of synthetic polymers over biopolymers is the ability to tune their properties to enable easy analysis and rewriting. The accelerated AB + CD method makes use of thermally labile links, which allows for easy sequencing, as mentioned above. Furthermore, these links can be easily broken by heating the polymer, which allows the digital information to be erased¹³⁴. This procedure will, however, break all linkers between the polymeric units, erasing all data, which means that the complete polymer has to be re-synthesized. To avoid the need for complete re-synthesis, Lutz et al. developed

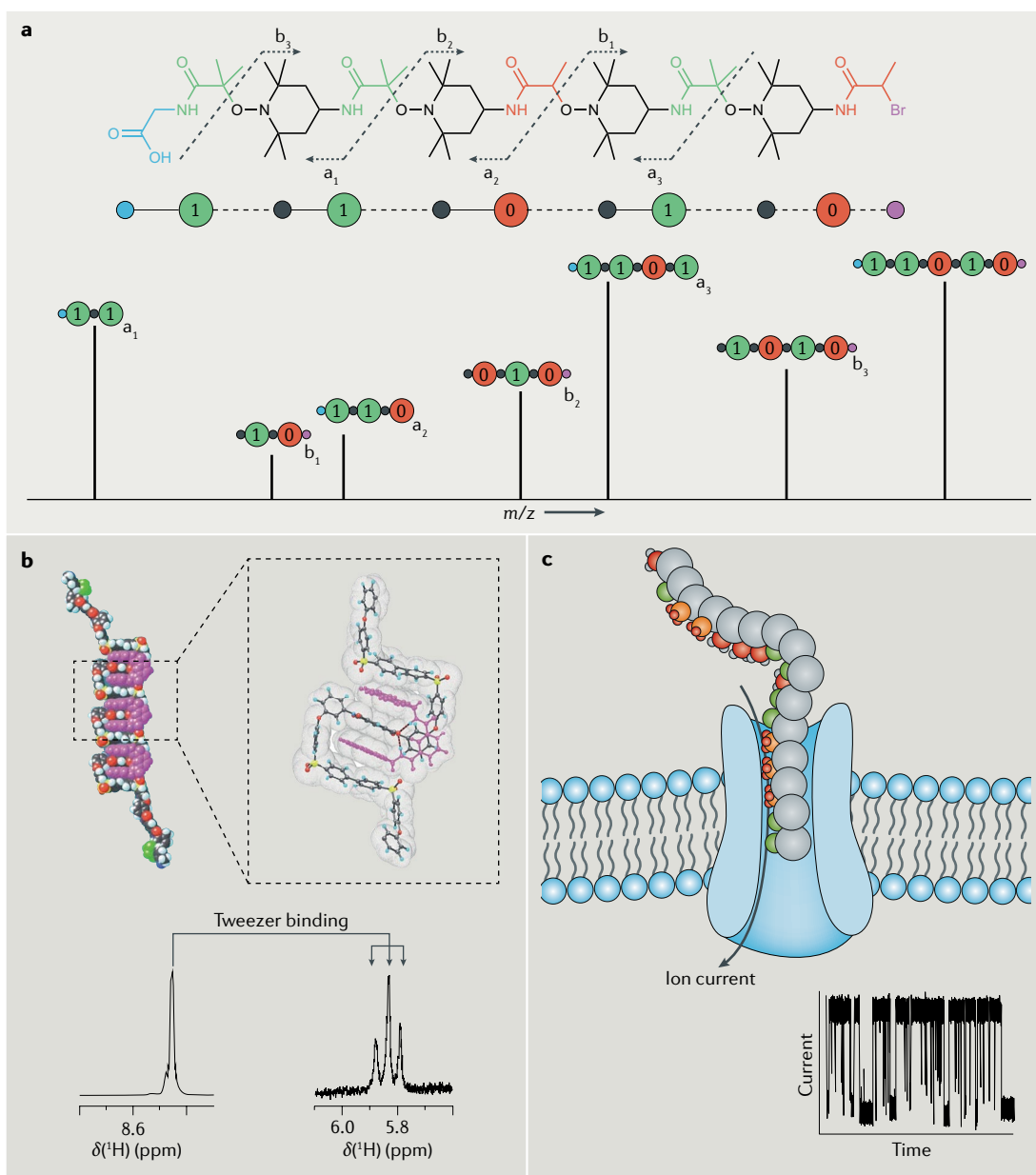


Fig. 5 | Reading of synthetic polymers. **a** | Example of tandem mass spectrometry (MS/MS) to read a polymer sequence. Monomers representing a binary 1 (green) or 0 (red) are separated by ‘weak’ alkoxyamine amide linkers (black circles and black chemical structures), which will break using relatively low energy, that is, before the dissociation of other bonds. Fragmentation will therefore result in the generation of straightforward fragments (a_1 – a_3 , b_1 – b_3), which can easily be interpreted from the MS/MS spectrum. **b** | NMR technique using tweezers (purple) that bind to a polymer chain by π – π stacking interactions. Upon binding, certain NMR peaks are separated and shifted, making the spectrum easier to read. **c** | Nanopore sequencing by moving a synthetic polymer (grey) through a biological or synthetic pore (blue). When the polymer moves through the pore, the functional groups (red, orange and green) change the ionic flow through the pore, resulting in characteristic changes in the current over time. Part **a** is adapted with permission from REF.¹²⁵, ACS. Part **b** is adapted with permission from REF.¹⁴⁹, Wiley-VCH. Part **c** is adapted with permission from REF.¹⁵⁷, PNAS.

a monomer that allowed modification after polymerization using a Huisgen azide–alkyne cycloaddition¹²⁹. Although complete rewriting is not possible, a number of simple changes could be made to the polymer while leaving the polymer chain largely unmodified¹²⁹. Another technique that could possibly be used for rewriting uses dynamic polymers, as described by Lehn et al.¹⁶³. These polymers are based upon a hydrazide and an aldehyde, which form an acylhydrazone bond through

a condensation reaction^{164,165}. Acylhydrazone formation is reversible under mildly acidic conditions, enabling data to be erased from the polymer¹⁶⁵. In the presence of other hydrazides or aldehydes, this reversibility could be exploited, at least in principle, to incorporate new acylhydrazines in the polymer chain. The incorporation of new monomers happens at random positions, leading to differently composed copolymers. Although this strategy allows data to be removed from the polymer chain,

selective rewriting at specific positions is not possible, as the insertion of new monomers is not spatially controlled. Thus, although rewriting of synthetic polymers remains a possibility, there have been very few reports of success, and tools to selectively change the data stored in synthetic polymers still need to be developed.

Writing by catalytic methods. Nature makes use of catalytic procedures to write information. Examples include the synthesis of proteins by the ribosome in which mRNA acts as a reading template and the copying of DNA by the DNA polymerase III enzyme¹⁶⁶. An important aspect of this writing is that it takes place in a processive fashion — the catalyst remains in contact with the polymeric substrate throughout. In this way, a large number of sequential writing events can take place, thus reducing the chance of errors (FIG. 6). In the alternative, known as distributive catalysis, the catalyst (enzyme) and substrate meet only once and are separated after reaction¹⁶⁷ (FIG. 6). Nolte and co-workers

developed a biomimetic catalytic system that can specifically and processively cleave DNA chains at AAA sites, which can be regarded as a first step towards catalytic writing^{168,169} (FIG. 6c). The catalyst is composed of the trimeric ring-shaped protein clamp (gp45) of the bacteriophage T4, which is usually associated with the replication polymerase (gp43). Nolte and co-workers replaced the replication polymerase by three manganese porphyrin complexes, which can cleave DNA in the presence of an oxidant. The modified clamp can bind to DNA and move along it unidirectionally while cleaving the AAA sites.

A second example from the same group involves a completely synthetic system that can ‘write’ epoxides on a high-molecular-mass polybutadiene chain with the help of a porphyrin cage catalyst and an oxidant. The catalyst threads onto the polymer chain, and while moving along it (in this case, in a hopping mode), it converts all double bonds into epoxide functions^{167,170,171} (FIG. 6d).

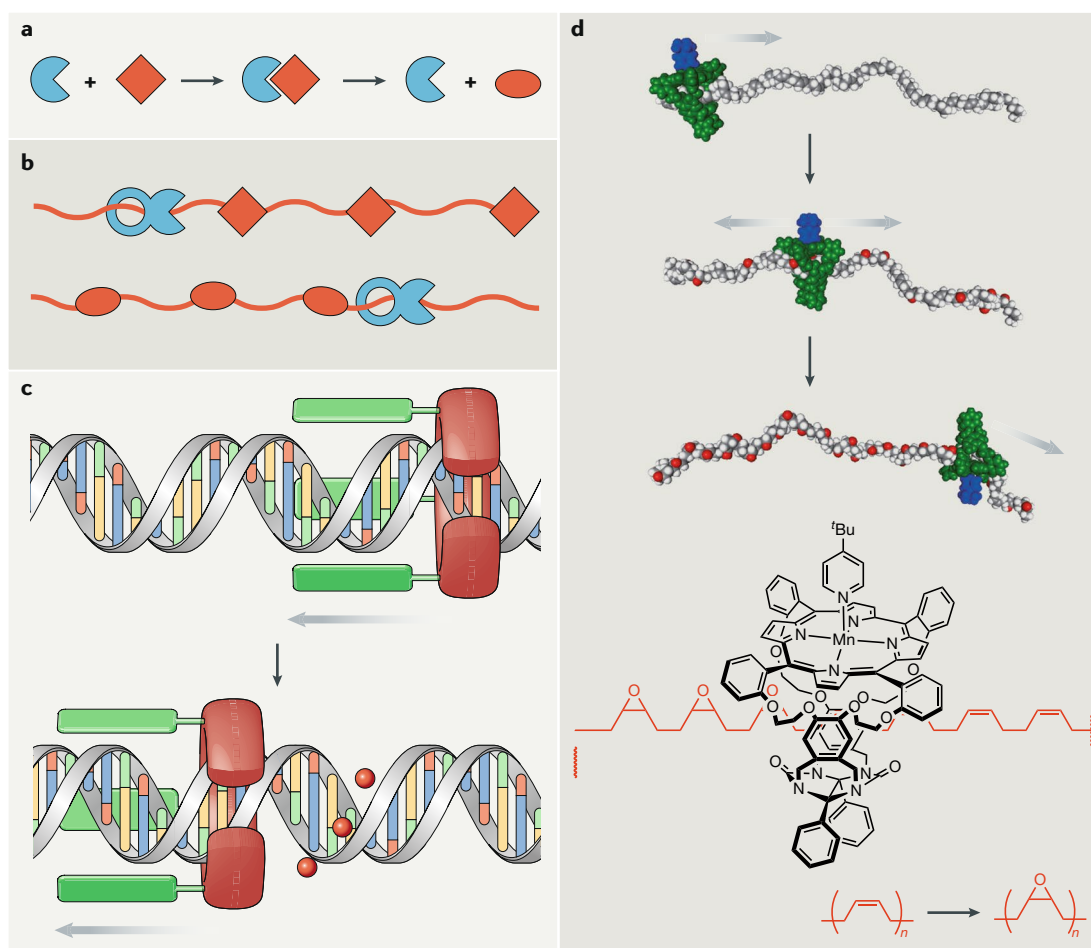


Fig. 6 | Catalytic writing. **a** | Distributive catalysis. A catalyst (blue) binds a substrate molecule (red diamond) and converts it into the product (red ellipse), after which the two separate. **b** | Processive catalysis. A catalyst (blue) binds to a (polymeric) substrate (red line with diamonds) and moves along it while performing multiple catalytic reactions without detaching¹⁶⁷. **c** | Biohybrid catalyst composed of a protein ring to which three manganese porphyrin catalysts have been attached. The catalyst cleaves DNA at AAA sites while moving along it. **d** | Synthetic catalyst constructed from a diphenylglycoluril cage compound and a manganese porphyrin complex. The catalyst threads onto polybutadiene and converts the double bonds of this polymer into epoxide functions while gliding along it. Adapted with permission from REF.¹⁶⁷, Wiley-VCH, and REF.¹⁶⁹, Springer Nature Limited.

Outlook

With the limits of silicon-based storage devices already in sight, attention has focused on alternative approaches to information storage. As we have described in this Review, DNA-based storage may become an interesting alternative for the current technologies, particularly in terms of storage density. Much progress has been made with regard to error protection mechanisms without much detriment to storage density. A major disadvantage of DNA compared with silicon is the much lower reading speed, which is especially problematic for use in random-access memory applications when only a small part of the data is desired. This means that for now, DNA is only applicable for archiving and long-term data storage purposes. A major problem still to be overcome is the current cost of DNA synthesis compared with the costs of silicon-based storage facilities. However, if we assume a decline in costs similar to that seen for silicon-based storage media (at least partly attributable to improvements in DNA technology), it seems likely that it will not be long before DNA-based storage is the standard for long-term data storage^{25,172,173}. This is certainly the case when the costs of maintenance and storage are taken into account, which are far smaller for DNA-based storage systems than for the silicon-based systems in current data centres. Further cost reductions might already be achieved quite rapidly by using quicker but less reliable synthesis protocols that require less time and reagents. Lower reliability would result in less valid DNA strands, but as the DNA fountain code has already shown, this can be compensated for by using robust and highly flexible coding strategies³⁴.

Future research will have to show whether DNA reading and rewritability can be improved, which will make DNA storage practical for data that are updated frequently. Data storage based on synthetic polymers — which can be prepared from a much larger set of monomers than biopolymers and which are more stable — might yet prove more useful for short-term applications. Furthermore, such synthetic systems do not require biological machineries and can be tuned for quick readout and rewritability. However, compared with DNA-based storage systems and computing, the field of synthetic encoded polymers is still in its infancy. It is expected that over time, the synthesis of long strands of synthetic encoded polymers will become easier and faster, while different aspects of the code can be easily changed, for example, in terms of the monomers. Of great fundamental interest are systems that encode information into biopolymers and synthetic polymers with the help of catalytic machines. This is the way nature stores and replicates information, and it is of interest to see whether we can effectively mimic this fascinating process. If processive catalytic systems based on clamp-shaped proteins and their attached enzyme writers, readers and erasers can be constructed, then other possibilities, such as the construction of biocomputers, come within reach.

We can conclude that DNA-based storage devices have clear potential to become good and reliable alternatives for long-term data storage. In addition, we believe that the use of natural and synthetic polymers to store (and process) data has the potential to completely reshape the global principle of data storage in the not-too-distant future.

Published online 30 October 2018

- Woods, C. (ed.) *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond* (Oriental Institute of the University of Chicago, 2010).
- Gantz, J. & Reinsel, D. The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. *Anal. Futur.* **2007**, 1–16 (2012).
- Cisco Systems, Inc. The zettabyte era: trends and analysis. *cisco* <https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/vni-hyperconnectivity-wp.pdf> (2017).
- Extance, A. How DNA could store all the world's data. *Nature* **537**, 22–24 (2016).
- Zhirnov, V., Zadeegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nat. Mater.* **15**, 366–370 (2016).
- Lunt, B. M. How long is long-term data storage? *Arch. Conf.* **2011**, 29–33 (2011).
- Shrivastava, S. & Badlani, R. Data storage in DNA. *Int. J. Electr. Energy* **2**, 119–124 (2014).
- Kumar, S. & Vijayaraghavan, R. Solid state drive (SSD) FAQ. *Dell* <https://www.dell.com/downloads/global/products/pvaul/en/solid-state-drive-faq-us.pdf> (2011).
- Greengard, S. Cracking the code on DNA storage. *Commun. ACM* **60**, 16–18 (2017).
- Greenberg, A., Hamilton, J., Maltz, D. A. & Patel, P. The cost of a cloud: research problems in data center networks. *SIGCOMM Comput. Commun. Rev.* **39**, 68–73 (2009).
- Bawden, T. Global warming: data centres to consume three times as much energy in next decade, experts warn. <https://www.independent.co.uk/environment/global-warming-data-centres-to-consume-three-times-as-much-energy-in-next-decade-experts-warn-a6830086.html> *Independent* (2016).
- Ritter, S. DNA to the rescue for data storage. *Chem. Eng. News Arch.* **93**, 40–41 (2015).
- Stikeman, A. Polymer memory. *Technol. Rev.* **105**, 31 (2002).
- Colquhoun, H. & Lutz, J.-F. Information-containing macromolecules. *Nat. Chem.* **6**, 455–456 (2014).
- Ogihara, M. & Ray, A. DNA computing on a chip. *Nature* **403**, 143 (2000).
- Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11**, 499–507 (2014).
- Shendure, J. et al. DNA sequencing at 40: past, present and future. *Nature* **550**, 345–353 (2017).
- Sanger, F. & Coulson, A. R. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–448 (1975).
- Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl Acad. Sci. USA* **74**, 560–564 (1977).
- Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155**, 335–350 (1987).
- Kelly, T. J. & Smith, H. O. A restriction enzyme from *Hemophilus influenzae*. II. Base sequence of the recognition site. *J. Mol. Biol.* **51**, 393–409 (1970).
- Smith, H. O. & Welcox, K. W. A restriction enzyme from *Hemophilus influenzae*. I. Purification and general properties. *J. Mol. Biol.* **51**, 379–391 (1970).
- Little, J. W., Zimmerman, S. B., Oshinsky, C. K. & Gellert, M. Enzymatic joining of DNA strands. II. An enzyme-adenylate intermediate in the dnp-dependent DNA ligase reaction. *Proc. Natl Acad. Sci. USA* **58**, 2004–2011 (1967).
- Zimmerman, S. B., Little, J. W., Oshinsky, C. K. & Gellert, M. Enzymatic joining of DNA strands: a novel reaction of diphosphopyridine nucleotide. *Proc. Natl Acad. Sci. USA* **57**, 1841–1848 (1967).
- O' Driscoll, A. & Sleator, R. D. Synthetic DNA: the next generation of big data storage. *Bioengineered* **4**, 123–125 (2013).
- Glanz, J. Power, pollution and the internet. *NYTimes* <https://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html?emc=eta1&r=0> (2012).
- De Silva, P. Y. & Ganegoda, G. U. New trends of digital data storage in DNA. *Biomed. Res. Int.* **2016**, 14 (2016).
- Adleman, L. M. Molecular computation of solutions to combinatorial problems. *Science* **266**, 1021–1024 (1994).
- Shilov, A. Western digital launches ultrastar DC HC530 14 TB PMR with TDMR HDD. *AnandTech* <https://www.anandtech.com/show/12665/western-digital-launches-ultrastar-dc-hc530-14-tb-pmr-with-tdmr-hdd> (2018).
- Mallory, M., Torabi, A. & Benakli, M. One terabit per square inch perpendicular recording conceptual design. *IEEE Trans. Magn.* **38**, 1719–1724 (2002).
- Milenkovic, O., Gabrys, R., Kiah, H. M. & Yazdi, S. M. H. T. Exabytes in a test tube: the case for DNA data storage. *IEEE Spectrum* <https://spectrum.ieee.org/semiconductors/devices/exabytes-in-a-test-tube-the-case-for-dna-data-storage> (2018).
- Castillo, M. From hard drives to flash drives to DNA drives. *Am. J. Neuroradiol.* **35**, 1–2 (2014).
- Carlson, R. The changing economics of DNA synthesis. *Nat. Biotechnol.* **27**, 1091 (2009).
- Erlach, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355**, 950–954 (2017).
- Bornholt, J. et al. A DNA-based archival storage system. *ACM SIGOPS Oper. Syst. Rev.* **50**, 637 (2016).
- Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P. & Barron, A. E. Landscape of next-generation sequencing technologies. *Anal. Chem.* **83**, 4327–4341 (2011).
- Shannon, C. E. A. Mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423 (1948).
- Cox, J. P. Long-term data storage in DNA. *Trends Biotechnol.* **19**, 247–250 (2001).
- Bancroft, C., Bowler, T., Bloom, B. & Clelland, C. T. Long-term storage of information in DNA. *Science* **293**, 1763–1765 (2001).

40. Bogard, C. M., Rouchka, E. C. & Arazi, B. DNA media storage. *Prog. Nat. Sci.* **18**, 603–609 (2008).
41. Carrillo, H. & Lipman, D. The multiple sequence alignment problem in biology. *SIAM J. Appl. Math.* **48**, 1073–1082 (1988).
42. Simpson, R. E. in *Introductory Electronics for Scientists and Engineers* (Addison-Wesley, 1987).
43. Reed, I. & Solomon, G. Polynomial codes over certain finite fields. *J. Soc. Ind. Appl. Math.* **8**, 300–304 (1960).
44. Moon, T. K. in *Error Correction Coding: Mathematical Methods and Algorithms* (Wiley-Interscience, 2005).
45. Blawat, M. et al. Forward error correction for DNA data storage. *Procedia Comput. Sci.* **80**, 1011–1022 (2016).
46. Doig, A. J. Improving the efficiency of the genetic code by varying the codon length—the perfect genetic code. *J. Theor. Biol.* **188**, 355–360 (1997).
47. Huffman, D. A. A. Method for the construction of minimum-redundancy codes. *Proc. IRE* **40**, 1098–1101 (1952).
48. Ailenberg, M. & Rotstein, O. An improved Huffman coding method for archiving text, images, and music characters in DNA. *Biotechniques* **47**, 747–754 (2009).
49. Smith, G. C., Fiddes, C. C., Hawkins, J. P. & Cox, J. P. L. Some possible codes for encrypting data in DNA. *Biotechnol. Lett.* **25**, 1125–1130 (2003).
50. Goldman, N. et al. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494**, 77–80 (2013).
51. Church, G. M., Gao, Y. & Kosuri, S. Next-generation digital information storage in DNA. *Science* **337**, 1628–1628 (2012).
52. Golomb, S. W. in *Mathematical Problems in the Biological Sciences (Proceedings of Symposia in Applied Mathematics)* Vol. 14 87–100 (American Mathematical Soc., 1962).
53. Davis, J. Microvenous. *Art J.* **55**, 70–74 (1996).
54. Clelland, C. T., Risca, V. & Bancroft, C. Hiding messages in DNA microdots. *Nature* **399**, 533–534 (1999).
55. Bourdenx, M. DNA as the next digital information storage support. *Mov. Disord.* **28**, 583 (2013).
56. MacKay, D. J. C. Fountain codes. *IEEE Proc.-Commun.* **152**, 1062–1068 (2005).
57. Micheloni, R. Solid-State Drive (SSD): a nonvolatile storage system. *Proc. IEEE* **105**, 583–588 (2017).
58. Friedland, A. E. et al. Synthetic gene networks that count. *Science* **324**, 1199–1202 (2009).
59. Bonnet, J., Subsoontorn, P. & Endy, D. Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc. Natl Acad. Sci. USA* **109**, 8884–8889 (2012).
60. Mayer, C., McInroy, G. R., Murat, P., Van Delft, P. & Balasubramanian, S. An epigenetics-inspired DNA-based data storage system. *Angew. Chem. Int. Ed.* **55**, 11144–11148 (2016).
61. Goldberg, A. D., Allis, C. D. & Bernstein, E. Epigenetics: a landscape takes shape. *Cell* **128**, 635–638 (2007).
62. Shapiro, R., Servis, R. E. & Welcher, M. Reactions of uracil and cytosine derivatives with sodium bisulfite. *J. Am. Chem. Soc.* **92**, 422–424 (1970).
63. Wang, R. Y.-H., Gehrke, C. W. & Ehrlich, M. Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues. *Nucleic Acids Res.* **8**, 4777–4790 (1980).
64. Tabatabaei Yazdi, S. M. H., Yuan, Y., Ma, J., Zhao, H. & Milenkovic, O. A. Rewritable, random-access DNA-based storage system. *Sci. Rep.* **5**, 14138 (2015).
65. Bryksin, A. V. & Matsumura, I. Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. *Biotechniques* **48**, 463–465 (2010).
66. Arita, M. in *Aspects of Molecular Computing* (eds Jonoska, N., Pa'un, G. & Rozenberg, G.) 23–35 (Springer Berlin Heidelberg, 2003).
67. Anchordoquy, T. J. & Molina, M. C. Preservation of DNA. *Cell Preserv. Technol.* **5**, 180–188 (2007).
68. Nicholson, W. L., Munakata, N., Horneck, G., Melosh, H. J. & Setlow, P. Resistance of *Bacillus* endospores to extreme terrestrial and extraterrestrial environments. *Microbiol. Mol. Biol. Rev.* **64**, 548–572 (2000).
69. Yachie, N., Sekiyama, K., Sugahara, J., Ohashi, Y. & Tomita, M. Alignment-based approach for durable data storage into living organisms. *Biotechnol. Prog.* **23**, 501–505 (2007).
70. Limbachiya, D. & Gupta, M. K. Natural Data Storage: a review on sending information from now to then via nature. Preprint at *arXiv* <http://arxiv.org/abs/1505.04890> (2015).
71. Shipman, S. L., Nivala, J., Macklis, J. D. & Church, G. M. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* **547**, 345 (2017).
72. Heaven, D. Video stored in live bacterial genome using CRISPR gene editing. *New Scientist* <https://www.newscientist.com/article/2140576-video-stored-in-live-bacterial-genome-using-crispr-gene-editing/> (2017).
73. Grass, R. N., Heckel, R., Puddu, M., Paunescu, D. & Stark, W. J. Robust chemical preservation of digital information on DNA in silica with error-correcting codes. *Angew. Chemie - Int. Ed.* **54**, 2552–2555 (2015).
74. Parker, J. Computing with DNA. *EMBO Rep.* **4**, 7–10 (2003).
75. Scudellari, M. Inner Workings: DNA for data storage and computing. *Proc. Natl Acad. Sci. USA* **112**, 15771–15772 (2015).
76. Cormen, T. H., Leiserson, C. E. & Rivest, R. L. *Introduction to Algorithms* (The MIT Press, 1990).
77. Braich, R. S., Chelyapov, N., Johnson, C., Rothmund, P. W. K. & Adleman, L. Solution of a 20-variable 3-SAT problem on a DNA computer. *Science* **296**, 499–502 (2002).
78. Benenson, Y. et al. Programmable and autonomous computing machine made of biomolecules. *Nature* **414**, 430–434 (2001).
79. Benenson, Y., Adar, R., Paz-Elizur, T., Livneh, Z. & Shapiro, E. DNA molecule provides a computing machine with both data and fuel. *Proc. Natl Acad. Sci. USA* **100**, 2191–2196 (2003).
80. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338 (2000).
81. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342 (2000).
82. Bonnet, J., Yin, P., Ortiz, M. E., Subsoontorn, P. & Endy, D. Amplifying genetic logic gates. *Science* **340**, 599–603 (2013).
83. Daniel, R., Rubens, J. R., Sarpeshkar, R. & Lu, T. K. Synthetic analog computation in living cells. *Nature* **497**, 619–623 (2013).
84. Farzadfar, F. & Lu, T. K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346**, 1256272 (2014).
85. Ratner, T., Piran, R., Jonoska, N. & Keinan, E. Biologically relevant molecular transducer with increased computing power and iterative abilities. *Chem. Biol.* **20**, 726–733 (2013).
86. Varghese, S., Elemans, J. A. A. W., Rowan, A. E. & Nolte, R. J. M. Molecular computing: paths to chemical Turing machines. *Chem. Sci.* **6**, 6050–6058 (2015).
87. Hirschberg, Y. Reversible formation and eradication of colors by irradiation at low temperatures. A photochemical memory model. *J. Am. Chem. Soc.* **78**, 2304–2312 (1956).
88. Adam, V. et al. Data storage based on photochromic and photoconvertible fluorescent proteins. *J. Biotechnol.* **149**, 289–298 (2010).
89. Ando, R., Hama, H., Yamamoto-Hino, M., Mizuno, H. & Miyawaki, A. An optical marker based on the UV-induced green-to-red photoconversion of a fluorescent protein. *Proc. Natl Acad. Sci. USA* **99**, 12651–12656 (2002).
90. Ando, R., Mizuno, H. & Miyawaki, A. Regulated fast nucleocytoplasmic shuttling observed by reversible protein highlighting. *Science* **306**, 1370–1373 (2004).
91. Adam, V. et al. Structural characterization of IrisFP, an optical highlighter undergoing multiple photo-induced transformations. *Proc. Natl Acad. Sci. USA* **105**, 18343–18348 (2008).
92. Mandzhikov, V. F., Murin, V. A. & Barachevskii, V. A. Nonlinear coloration of photochromic spiropyran solutions. *Sov. J. Quantum Electron.* **3**, 128–129 (1973).
93. Birge, R. R. Protein-based computers. *Sci. Am.* **272**, 90–95 (1995).
94. Renugopalakrishnan, V. et al. Retroengineering bacteriorhodopsins: design of smart proteins by bionanotechnology. *Int. J. Quantum Chem.* **95**, 627–631 (2003).
95. Renugopalakrishnan, R., Khizroev, K., Anand, A., Pingzuo, P. & Lindvold, L. Future memory storage technology: protein-based memory devices may facilitate surpassing Moore's law. *IEEE Trans. Magn.* **43**, 773–775 (2007).
96. Oesterheld, D., Brauchle, C. & Hampf, N. Bacteriorhodopsin: a biological material for information processing. *Q. Rev. Biophys.* **24**, 425–478 (1991).
97. Dawkins, R. *The Blind Watchmaker* (Longman, 1986).
98. Orgel, L. E. Molecular replication. *Nature* **358**, 203–209 (1992).
99. Sievers, D. & von Kiedrowski, G. Self-replication of complementary nucleotide-based oligomers. *Nature* **369**, 221–224 (1994).
100. Brudno, Y. & Liu, D. R. Recent progress toward the templated synthesis and directed evolution of sequence-defined synthetic polymers. *Chem. Biol.* **16**, 265–276 (2009).
101. Lutz, J.-F., Ouchi, M., Liu, D. R. & Sawamoto, M. Sequence-controlled polymers. *Science* **341**, 1238149 (2013).
102. Piccirilli, J. A., Benner, S. A., Krauch, T., Moroney, S. E. & Benner, S. A. Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature* **343**, 33–37 (1990).
103. Kool, E. T. Replacing the nucleobases in DNA with designer molecules. *Acc. Chem. Res.* **35**, 936–943 (2002).
104. Lewandowski, B. et al. Sequence-specific peptide synthesis by an artificial small-molecule machine. *Science* **339**, 189–193 (2013).
105. Niu, J., Hill, R. & Liu, D. R. Enzyme-free translation of DNA into sequence-defined synthetic polymers structurally unrelated to nucleic acids. *Nat. Chem.* **5**, 282–292 (2013).
106. Houshyar, S. et al. The scope for synthesis of macro-RAFT agents by sequential insertion of single monomer units. *Polym. Chem.* **3**, 1879–1889 (2012).
107. Tong, X., Guo, B. & Huang, Y. Toward the synthesis of sequence-controlled vinyl copolymers. *Chem. Commun.* **47**, 1455–1457 (2011).
108. Minoda, M., Sawamoto, M. & Higashimura, T. Sequence-regulated oligomers and polymers by living cationic polymerization. 2. Principle of sequence regulation and synthesis of sequence-regulated oligomers of functional vinyl ethers and styrene derivatives. *Macromolecules* **23**, 4889–4895 (1990).
109. Brooks, P. P. et al. Monomer sequencing in living anionic polymerization using kinetic control. *Macromol. Symp.* **323**, 42–50 (2013).
110. Rzaev, Z. M. O. Complex-radical alternating copolymerization. *Prog. Polym. Sci.* **25**, 163–217 (2000).
111. Pfeifer, S. & Lutz, J.-F. A. Facile procedure for controlling monomer sequence distribution in radical chain polymerizations. *J. Am. Chem. Soc.* **129**, 9542–9543 (2007).
112. Lutz, J.-F., Schmidt, B. V. K. J. & Pfeifer, S. Tailored polymer microstructures prepared by atom transfer radical copolymerization of styrene and N-substituted maleimides. *Macromol. Rapid Commun.* **32**, 127–135 (2011).
113. Chan-Seng, D., Zamfir, M. & Lutz, J.-F. Polymer-chain encoding: Synthesis of highly complex monomer sequence patterns by using automated protocols. *Angew. Chemie - Int. Ed.* **51**, 12254–12257 (2012).
114. Moatsou, D., Hansell, C. F. & O'Reilly, R. K. Precision polymers: a kinetic approach for functional poly(norbornenes). *Chem. Sci.* **5**, 2246–2250 (2014).
115. Gody, G., Zetterlund, P. B., Perrier, S. & Harrison, S. The limits of precision monomer placement in chain growth polymerization. *Nat. Commun.* **7**, 10514 (2016).
116. Gody, G., Maschmeyer, T., Zetterlund, P. B. & Perrier, S. Rapid and quantitative one-pot synthesis of sequence-controlled polymers by radical polymerization. *Nat. Commun.* **4**, 2505 (2013).
117. Engels, N. G. et al. Sequence-controlled methacrylic multiblock copolymers via sulfur-free RAFT emulsion polymerization. *Nat. Chem.* **9**, 171–178 (2017).
118. Ten Brummelhuis, N. Controlling monomer-sequence using supramolecular templates. *Polym. Chem.* **6**, 654–667 (2015).
119. Minoda, M., Sawamoto, M. & Higashimura, T. Sequence-regulated oligomers and polymers by living cationic polymerization. III. Synthesis and reactions of sequence-regulated oligomers with a polymerizable group. *J. Polym. Sci. Part A Polym. Chem.* **31**, 2789–2797 (1993).
120. Berthet, M.-A., Zarafshani, Z., Pfeifer, S. & Lutz, J.-F. Facile synthesis of functional periodic copolymers: a step toward polymer-based molecular arrays. *Macromolecules* **43**, 44–50 (2010).
121. Tsarevsky, N. V., Sumerlin, B. S. & Matyjaszewski, K. Step-growth “click” coupling of telechelic polymers prepared by atom transfer radical polymerization. *Macromolecules* **38**, 3558–3561 (2005).

122. Lutz, J.-F., Lehn, J.-M., Meijer, E. W. & Matyjaszewski, K. From precision polymers to complex materials and systems. *Nat. Rev. Mater.* **1**, 16024 (2016).
123. Badi, N. & Lutz, J.-F. Sequence control in polymer synthesis. *Chem. Soc. Rev.* **38**, 3383–3390 (2009).
124. Pfeifer, S., Zerafsani, Z., Badi, N. & Lutz, J.-F. Liquid-phase synthesis of block copolymers containing sequence-ordered segments. *J. Am. Chem. Soc.* **131**, 9195–9197 (2009).
125. Lutz, J.-F. Coding macromolecules: inputting information in polymers using monomer-based alphabets. *Macromolecules* **48**, 4759–4767 (2015).
126. Gunay, U. et al. Chemoselective synthesis of uniform sequence-coded polyurethanes and their use as molecular tags. *Chem* **1**, 114–126 (2016).
127. Beaucage, S. L. & Caruthers, M. H. Deoxynucleoside phosphoramidites—a new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* **22**, 1859–1862 (1981).
128. Beaucage, S. L. & Iyer, R. P. Advances in the synthesis of oligonucleotides by the phosphoramidite approach. *Tetrahedron* **48**, 2223–2311 (1992).
129. Al Ouahabi, A., Charles, L. & Lutz, J.-F. Synthesis of non-natural sequence-coded polymers using phosphoramidite chemistry. *J. Am. Chem. Soc.* **137**, 5629–5635 (2015).
130. Cavallo, G., Al Ouahabi, A., Oswald, L., Charles, L. & Lutz, J.-F. Orthogonal synthesis of “easy-to-read” information-containing polymers using phosphoramidite and radical coupling steps. *J. Am. Chem. Soc.* **138**, 9417–9420 (2016).
131. Al Ouahabi, A., Kotera, M., Charles, L. & Lutz, J.-F. Synthesis of monodisperse sequence-coded polymers with chain lengths above DP100. *ACS Macro Lett.* **4**, 1077–1080 (2015).
132. Trinh, T. T., Oswald, L., Chan-Seng, D. & Lutz, J. F. Synthesis of molecularly encoded oligomers using a chemoselective ‘AB+CD’ iterative approach. *Macromol. Rapid Commun.* **35**, 141–145 (2014).
133. Trinh, T. T., Oswald, L., Chan-Seng, D., Charles, L. & Lutz, J.-F. Preparation of information-containing macromolecules by ligation of dyad-encoded oligomers. *Chem. – A Eur. J.* **21**, 11961–11965 (2015).
134. Roy, R. K. et al. Design and synthesis of digitally encoded polymers that can be decoded and erased. *Nat. Commun.* **6**, 7237 (2015).
135. Zydziak, N. et al. Coding and decoding libraries of sequence-defined functional copolymers synthesized via photoligation. *Nat. Commun.* **7**, 13672 (2016).
136. Mutlu, H. & Lutz, J.-F. Reading polymers: sequencing of natural and synthetic macromolecules. *Angew. Chem. Int. Ed.* **53**, 13010–13019 (2014).
137. Gruendling, T., Weidner, S., Falkenhagen, J. & Barner-Kowollik, C. Mass spectrometry in polymer chemistry: a state-of-the-art up-date. *Polym. Chem.* **1**, 599–617 (2010).
138. Altuntas, E. & Schubert, U. S. “Polymeromics”: mass spectrometry based strategies in polymer science toward complete sequencing approaches: a review. *Anal. Chim. Acta* **808**, 56–69 (2014).
139. Charles, L. et al. MS/MS-assisted design of sequence-controlled synthetic polymers for improved reading of encoded information. *J. Am. Soc. Mass Spectrom.* **28**, 1149–1159 (2017).
140. Charles, L., Laure, C., Lutz, J.-F. & Roy, R. K. MS/MS sequencing of digitally encoded poly(alkoxyamine amide)s. *Macromolecules* **48**, 4319–4328 (2015).
141. Amalian, J.-A. et al. Controlling the structure of sequence-defined poly(phosphodiester)s for optimal MS/MS reading of digital information. *J. Mass Spectrom.* **52**, 788–798 (2017).
142. Al Ouahabi, A., Amalian, J.-A., Charles, L. & Lutz, J.-F. Mass spectrometry sequencing of long digital polymers facilitated by programmed inter-byte fragmentation. *Nat. Commun.* **8**, 967 (2017).
143. Nesvadba, P. N-alkoxyamines: synthesis, properties, and applications in polymer chemistry, organic synthesis, and materials science. *Chimia (Aarau)* **60**, 832–840 (2006).
144. Amalian, J.-A., Trinh, T. T., Lutz, J.-F. & Charles, L. MS/MS digital readout: analysis of binary information encoded in the monomer sequences of poly(triazole amide)s. *Anal. Chem.* **88**, 3715–3722 (2016).
145. König, N. F., Al Ouahabi, A., Poyer, S., Charles, L. & Lutz, J.-F. A simple post-polymerization modification method for controlling side-chain information in digital polymers. *Angew. Chemie – Int. Ed.* **56**, 7297–7301 (2017).
146. Burel, A., Carapito, C., Lutz, J.-F. & Charles, L. MS-DECODER: milliseconds sequencing of coded polymers. *Macromolecules* **50**, 8290–8296 (2017).
147. Tonelli, A. E. A. Case for characterizing polymers with the Kerr effect. *Macromolecules* **42**, 3830–3840 (2009).
148. Hardt, S. N. et al. Characterizing polymer macrostructures by identifying and locating microstructures along their chains with the Kerr effect. *J. Polym. Sci. Part B Polym. Phys.* **51**, 735–741 (2013).
149. Colquhoun, H. M. & Zhu, Z. Recognition of polyimide sequence information by a molecular tweezer. *Angew. Chem. Int. Ed.* **43**, 5040–5045 (2004).
150. Colquhoun, H. M., Zhu, Z., Cardin, C. J., Gan, Y. & Drew, M. G. B. Sterically controlled recognition of macromolecular sequence information by molecular tweezers. *J. Am. Chem. Soc.* **129**, 16163–16174 (2007).
151. Zhu, Z., Cardin, C. J., Gan, Y. & Colquhoun, H. M. Sequence-selective assembly of tweezer molecules on linear templates enables frameshift-reading of sequence information. *Nat. Chem.* **2**, 653–660 (2010).
152. Meller, A., Nivon, L., Brandin, E., Golovchenko, J. & Branton, D. Rapid nanopore discrimination between single polynucleotide molecules. *Proc. Natl Acad. Sci. USA* **97**, 1079–1084 (2000).
153. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl Acad. Sci. USA* **93**, 13770–13773 (1996).
154. Wanunu, M., Sutin, J., McNally, B., Chow, A. & Meller, A. DNA translocation governed by interactions with solid-state nanopores. *Biophys. J.* **95**, 4716–4725 (2008).
155. Wanunu, M. & Meller, A. Chemically modified solid-state nanopores. *Nano Lett.* **7**, 1580–1585 (2007).
156. Bezrukov, S. M., Vodyanov, I., Brutyan, R. A. & Kasianowicz, J. J. Dynamics and free energy of polymers partitioning into a nanoscale pore. *Macromolecules* **29**, 8517–8522 (1996).
157. Reiner, J. E., Kasianowicz, J. J., Nablo, B. J. & Robertson, J. W. F. Theory for polymer analysis using nanopore-based single-molecule mass spectrometry. *Proc. Natl Acad. Sci. USA* **107**, 12080–12085 (2010).
158. Movileanu, L. & Bayley, H. Partitioning of a polymer into a nanoscopic protein pore obeys a simple scaling law. *Proc. Natl Acad. Sci. USA* **98**, 10137–10141 (2001).
159. Gibrat, G. et al. Polyelectrolyte entry and transport through an asymmetric α -hemolysin channel. *J. Phys. Chem. B* **112**, 14687–14691 (2008).
160. Boukhet, M. et al. Translocation of precision polymers through biological nanopores. *Macromol. Rapid Commun.* **38**, 1700680 (2017).
161. Sakaue, T. & Brochard-Wyart, F. Nanopore-based characterization of branched polymers. *ACS Macro Lett.* **3**, 194–197 (2014).
162. Mirigian, S., Wang, Y. & Muthukumar, M. Translocation of a heterogeneous polymer. *J. Chem. Phys.* **137**, 64904 (2012).
163. Skene, W. G. & Lehn, J.-M. P. Dynamers: polyacrylhydrazones reversible covalent polymers, component exchange, and constitutional diversity. *Proc. Natl Acad. Sci. USA* **101**, 8270–8275 (2004).
164. Bunyapaiboonsri, T. et al. Dynamic deconvolution of a pre-equilibrated dynamic combinatorial library of acetylcholinesterase inhibitors. *ChemBioChem* **2**, 438–444 (2001).
165. Nguyen & Ivan Huc, R. Optimizing the reversibility of hydrazone formation for dynamic combinatorial chemistry. *Chem. Commun.* 942–943 (2003).
166. Clark, D. P. & Pazdernik, N. J. in *Biotechnology* 97–130 (Academic Cell, 2016).
167. van Dongen, S. F. M., Elemans, J. A. A. W., Rowan, A. E. & Nolte, R. J. M. Processive catalysis. *Angew. Chem. Int. Ed.* **53**, 11420–11428 (2014).
168. van Dongen, S. F. M. et al. A clamp-like biohybrid catalyst for DNA oxidation. *Nat. Chem.* **5**, 945 (2013).
169. Prins, L. J. & Scrimin, P. Processive catalysis: thread and cut. *Nat. Chem.* **5**, 899–900 (2013).
170. Thordarson, P., Bijsterveld, E. J. A., Rowan, A. E. & Nolte, R. J. M. Epoxidation of polybutadiene by a topologically linked catalyst. *Nature* **424**, 915–918 (2003).
171. Elemans, J. A. A. W., Bijsterveld, E. J. A., Rowan, A. E. & Nolte, R. J. M. Manganese porphyrin hosts as epoxidation catalysts – activity and stability control by axial ligand effects. *Eur. J. Org. Chem.* **2007**, 751–757 (2007).
172. Carr, P. A. & Church, G. M. Genome engineering. *Nat. Biotechnol.* **27**, 1151–1162 (2009).
173. Feher, T., Burland, V. & Posfai, G. In the fast lane: large-scale bacterial genome engineering. *J. Biotechnol.* **160**, 72–79 (2012).
174. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. in *Introduction to Algorithms* 3rd edn (The MIT Press, 2009).

Acknowledgements

R.J.M.N. acknowledges support from the European Research Council (ERC Advanced Grant ENCOPOLE-74092) and from the Dutch National Science Organization NWO (Gravitation program 024.001.035).

Author contributions

All authors contributed equally to all aspects of manuscript research, writing and revision.

Competing interests

The authors declare no competing interests.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Reviewer information

Nature Reviews Chemistry thanks S. Harrison and the other anonymous reviewer(s) for their contribution to the peer review of this work.