Encoding information into polymers

Martin G.T.A. Rutten[a], Frits W. Vaandrager[b], Johannes A.A.W. Elemans[a], and Roeland J.M. Nolte[a,]*

[a]Radboud University, Institute for Molecules and Materials, Heyendaalseweg 135, 6525 AJ Nijmegen, The Netherlands.
[b]Radboud University, Institute for Computing and Information Sciences, Department of Software Science, Toernooiveld 212, 6525 EC Nijmegen, The Netherlands.
**email: R.Nolte@science.ru.nl*

Abstract| Polymers show great potential as a durable and high density alternative for data storage and for this purpose the natural polymer DNA has already attracted much interest from researchers. A DNA based storage system, which makes use of the four nucleotides to store binary codes, is more durable and can store information with a much higher density than conventional storage systems. Synthetic polymers have properties that make them even more suitable for data storage, at least in principle, if complete control over their composition, i.e. monomer sequence can be obtained. This review addresses the current status of data storage in DNA, proteins, and synthetic polymers, with the objective to overcome the problems of the current data storage technology.

Written records are crucial for our understanding of past civilizations. They are so important, that we commonly define "history" as the study of the past as it is described in written documents, and refer to earlier events as "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 about the Olmec civilization in Mexico through the Cascajal Block, a stone slab with 3,000-year-old writing made of serpentinite[1].

Digital data has completely changed the way we write, use, and access information nowadays and we live in what is commonly referred to as the 'digital world.' It is expected that the need for digital information will continue to grow, reaching the level of 44 trillion gigabytes in 2020[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 circa 10 years, while standard hard drives are susceptible to magnetic fields, high temperatures, and mechanical failures[6–8]. The decay of the storage media results in data loss, which is currently prevented by a constant shuffling of data between different devices and facilities. Due to the explosion of digital data, there is a constant need to migrate to new technologies that do not always support the old technologies[9]. Hence much of the information that we 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 here. Current storage technologies require significant space and enormous amounts of energy[10]. The world data centers currently consume annually ca. 420

terawatt hours of electricity, which is higher than the UK total energy consumption (300 terawatt hours)[11]. Hence, it is clear that other ways of writing and storing of information are needed.

As alternative for silicon-based devices, polymers show great potential for data storage, because they are stable, at least the synthetic ones, energy efficient, and have the possibilities of high storage densities[4,12–14]. Polymers are large macromolecules composed of many repeating units, i.e. 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 the biological processes in nature. At least theoretically, polymers offer the intriguing possibility to durably store all the data of the world in just a handful of material, which we could preserve safely in some cave or bunker on earth, or even on Mars.

In this paper the current status with regard to data storage in natural and synthetic polymers is reviewed. We will focus on fundamental aspects as the field has not developed yet to the extent that practical applications are possible. Current experiments, however, are promising and show a great potential for the near future. First, the units of information, bits and bytes will briefly be discussed, after which the basic principles and different strategies for DNA encoding will be outlined. In addition to DNA storage, also DNA computation, which has attracted much attention because of the possibility to perform parallel computation, will be examined[15]. Besides DNA also proteins can be used as storage systems, which will also briefly be reviewed. Finally, the last section of this review focusses on the most recent developments in alternative information storage, including especially synthetic polymers, both for data storage and computation.

General aspects

Data

For the purposes of this review, data is simply viewed as a sequence of bits, i.e. a row of 0's and 1's. We do not care whether this sequence represents a text file, an audio file, a movie, a tarball, or something else, and whether or not the data is 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 the information for the reproduction of a species in nature, namely in the form of a quaternary code, i.e. a specific sequence of 4 base pairs: A = adenine, G = guanine, C = cytosine, and T = thymine. DNA has several properties that makes it convenient for data storage. It is relatively robust and the tools to write and read information i.e. DNA synthesis and sequencing, are available. The synthesis of DNA is nowadays carried out using oligonucleotide arrays, which are able to synthesize large pools of DNA strands in parallel[16]. The reading of DNA (DNA sequencing) has seen tremendous developments during the past 40 years[17]. Since the mid-1970s, for a long time sequencing was achieved by methods developed by Sanger-Coulson and Maxam-Gilbert. Both approaches are based upon dividing a large DNA strand in different sections based on labeled base pairs[18,19]. The increase in demand for low-cost and rapid sequencing of large genomes urged the development of alternative approaches that began to take shape throughout the 1980s and 1990s, but 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 known, allow for a much faster nucleobase readout by analyzing in parallel large amounts of small DNA fragments immobilized on two-dimensional surfaces, using fluorescence-based detection and automated analysis. 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, higher sequencing speeds, and make use of smaller and often portable equipment. 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 recently.

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), like in a text document[20–25]. Additionally, a DNA-based storage system is a million times more energy efficient than the systems present in current computers, making it eco-friendly when compared to the energy consuming data centers[25–28]. In this connection it should be mentioned, however, that much of the energy consumed in the data centers is needed for writing, reading, copying, etc. and less for the data storage itself. The biggest advantage comes, however, when the density of data storage is considered, being significant higher than that of conventional methods, overcoming the problem of space to store all our data. Where currently the largest magnetic hard drive has a capacity of 14 terabytes[29], the maximum storage density for DNA is two bits per nucleotide or 455 exabytes per gram of single-stranded DNA. This means that the entire information produced in the world during one year can be stored in 4 grams of DNA[30].

Encoding data in DNA

To store data in DNA, it must be converted into a DNA sequence by a translational code. This code should be unambiguous and ideally also possess some kind of error correction. It is important to consider that every DNA strand, besides the data, also needs a forward and reverse primer sequence at the beginning and end of the strand. These primer sequences are 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 becomes cheaper nowadays, the synthesis of long strands of DNA is still relatively expensive[31]. To express the efficiency of a coding strategy the concept of Shannon information capacity, which gives a bound on how much information can be stored into one unit of the code, may be used[32]. Clearly, the information capacity of DNA is at most 2 bits per nucleotide, meaning that each nucleotide representing two bits of information (for instance: A = 00, C = 01, G = 10, T = 11)[33]. However, several factors limit this maximum capacity, one being the difference between A=T and G≡C base pairing. An A=T base pair has two hydrogen bonds, whereas C≡G has three hydrogen bonds, which means that the latter requires more energy to break. Different DNA strands will possess different melting temperatures, depending on their A=T/G≡C ratio, which makes the PCR amplification less efficient. Another difficulty, especially problematic for sequencing, is the occurrence of homopolymer runs (runs of two or

more identical bases), which are associated with higher error rates during sequencing[34]. Both these factors, A=T/G≡C ratio and homopolymers, limit the storage capacity, as not every nucleotide can be placed at every position. Even without homopolymer sequences present, DNA replication and sequencing are prone to errors, which will corrupt the data. To prevent this data corruption, multiple copies of the DNA strand are often included, which also reduces the 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 decoding the sequencing data. The channel is noisy due to various types of errors, as explained above. Information theory, as developed by Claude E. Shannon, defines the notion of capacity for a noisy channel and provides a mathematical model by which one can compute it[35,36]. This channel capacity provides a tight upper bound on the rate at which information can be reliably transmitted. Different from classical information theory, where noise is independently distributed, the error pattern in DNA heavily depends on the input sequence. Nevertheless, Erlich et al, after combining the expected dropout rates and barcoding demand, succeeded to derive an overall Shannon information capacity of ~1.83 bits per nucleotide for a range of practical architectures for DNA storage devices[32].

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. An aspect that makes this problem even more complex is the impossibility of synthesizing arbitrarily long DNA strands, making it impossible to create one strand containing all the data. Instead, the data needs to be divided in multiple smaller fragments that all encode a part of the entire sequence. Aligning the fragments allows one to retrieve all data, however, the problem is how to let the decoder know what the order of all fragments is. One option is to begin every DNA strand with a sequence that counts upwards, before the actual message starts; for instance, the first fragments has the binary code 00001, the second 00010, etc. Another way would be to use an encoding strategy in which every fragment encodes a part of the previous fragment with an extension. For instance, one fragment encoding the first sentence of a novel, a second fragment encoding the first and second sentence, the third fragment encoding the second and third sentence, etc. The direction becomes then clear by aligning the repeats[37]. Another method, developed by Bancroft et al., includes the use of two different DNA classes: one containing the data and another one containing the polyprimer key (PPK)[38]. In this strategy, every DNA strand, containing data, is composed of not only the data itself but also a unique sequencing primer (Figure 1). The PPK contains all sequencing primers in the correct order, holding the exact direction to align all DNA strands[38]. Using this method, Bancroft et al. were able to encode and decode the opening line of a novel[38].

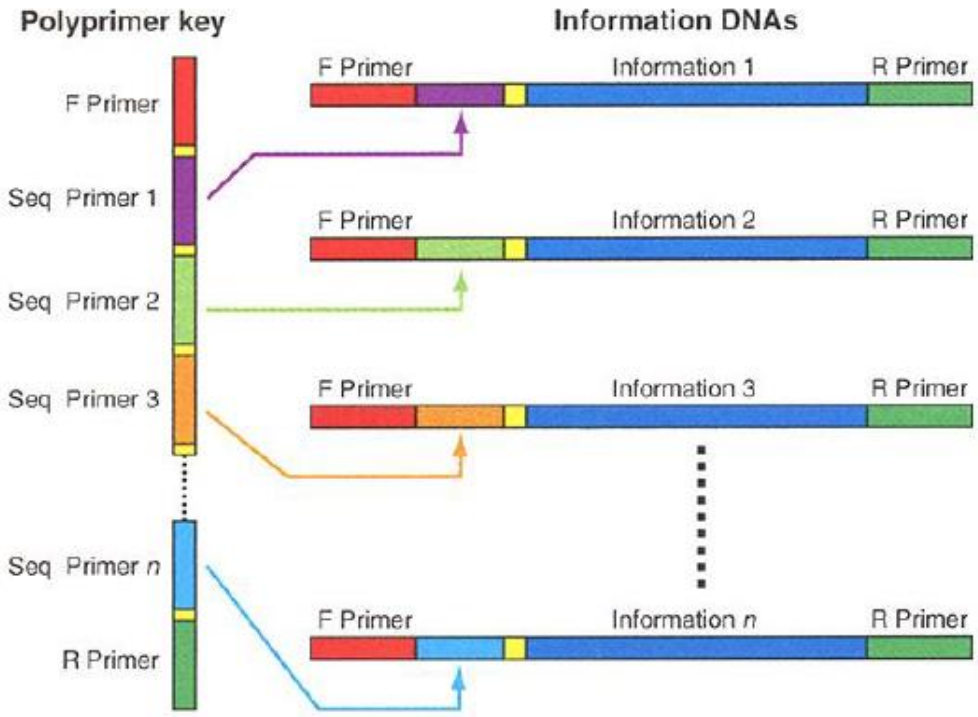

Figure 1| **Data storage in DNA.** Schematic presentation of the way DNA molecules are used for data storage by a method developed by Bancroft et al[38]. DNA strands that encode the data are composed of a forward and reverse primer, flanking the information and a sequencing primer. The polyprimer key (PPK) holds all sequence primers in the correct order, making that the fragments can be aligned correctly during sequencing. Adopted from Bancroft et al[38].

Error correction

Both DNA synthesis and sequencing are highly error-prone[16]. In addition, mutations may occur during storage. Error correction is therefore a key aspect in DNA storage as it would be unacceptable to lose data on a large scale. The simplest method for data correction is to include multiple copies of the same message, i.e. multiple DNA strands with the same sequence. This allows for correction of errors by comparing the DNA sequences by multiple sequence alignment[39]. The latter involves the correct retrieval of the sequences by using the conserved regions between the strands. In order to reduce the computational power needed to align all sequences, smart algorithms have been developed[39,40].

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 protection[41]. Two bit sequences, named A and B, may 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 upon 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 stands, to clarify from which strands the XOR was taken (Figure 2). This encoding strategy gives overall three stands and allows failure of one of these strands, as only two out of three strands are needed to reconstruct the third. This

encoding system allows for error correction but also for a denser information storage compared to multiple sequence alignment, which needs multiple copies of the same file.

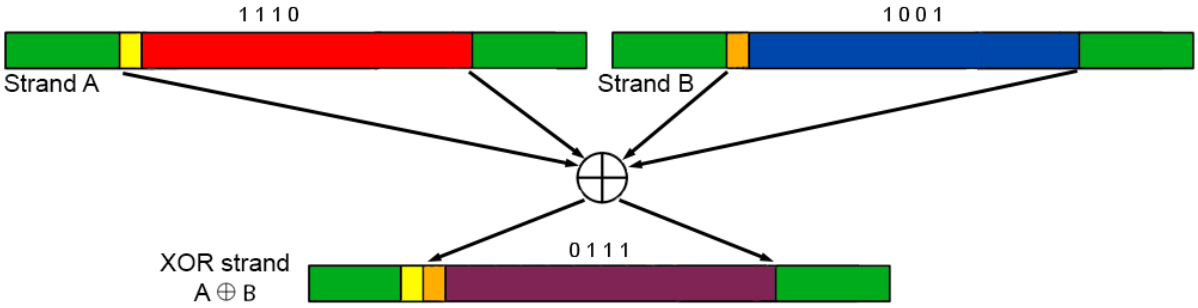

Figure 2| **XOR encoding.** Two strands are taken to compose a third one, which is the exclusive-or of both strands. Strands contain primers (green), unique address labels (yellow and orange) and data (red, blue, purple).

Another error correction method adapted from computer technology is the use of Reed–Solomon codes, which were introduced in 1960 and are applied in cd and dvd devices[42]. The exact mathematical basis of this correction method goes beyond the scope of this review. 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 is therefore divided in multiple pieces, e.g. 4 (Figure 3a). Every piece of data is given a coordinate point (*x*, *y*), defining the location (*x*-value) and the actual data (a row of 0's and 1's, *y*-value) (Figure 3a). A polynomial curve can be drawn through the created coordinate points and the polynomial function *P(x)* can be derived, necessary to create the parity symbols (Figure 3b). These parity symbols are extra data points along the line (DNA chain), calculated from the polynomial function, and are stored as parity besides the data (Figure 3b). When some of the original data is lost, the remaining data points and parity points 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 (Figure 3c)[43,44]. The abovementioned error correction methods require the use of extra nucleotides, which is often taken for granted. Some encoding strategies, however, also contain a form of error correction themselves, as will be outlined below.

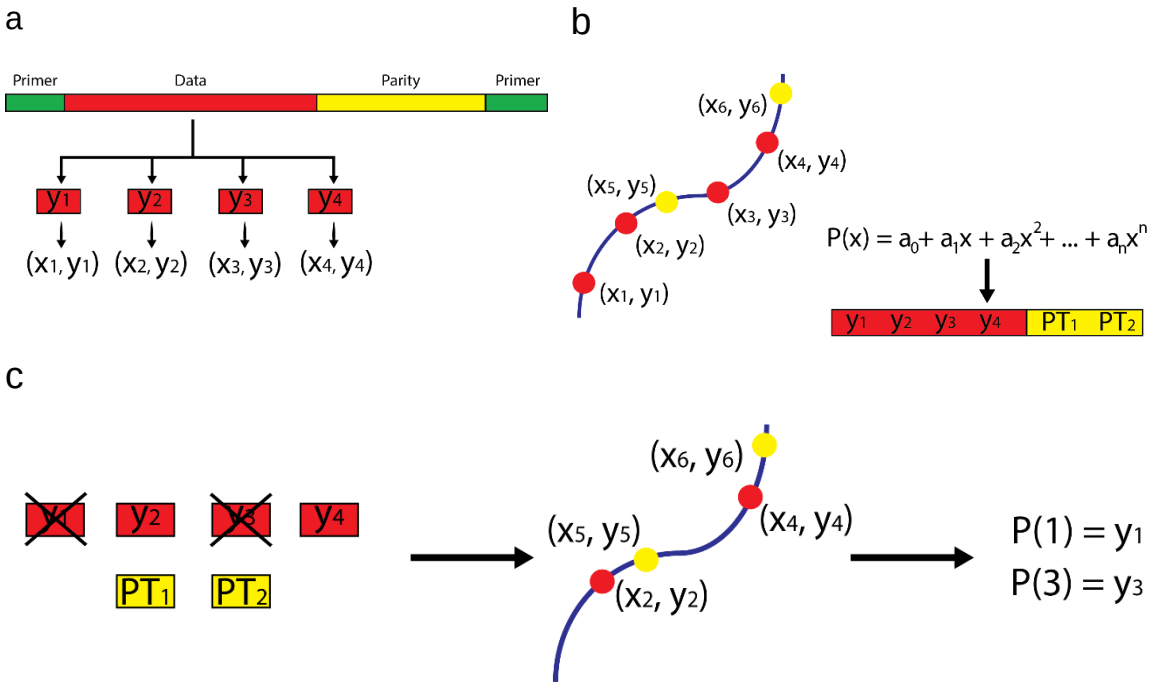

Figure 3| **Basic principle of Reed–Solomon correction codes. a**| representation of a DNA strand containing data (red), parity (yellow), and primers (green). To calculate the parity symbols, data is split into fragments and every fragment is given an x-coordinate. Coordinate points (x, y) are created, in which the y-coordinate represents the original data. **b**| A polynomial curve and function can be derived through the coordinate points of the data (red). Extra data points (yellow) are calculated as parity from the polynomial function. **c**| If two data fragments are lost (e.g. Y1 and Y3) the parity coordinate points can be used to reconstruct the original polynomial curve and function. The retrieved function can subsequently be utilized to recalculate the lost data points.

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 varied[45]. This strategy assigned a shorter codon length to more frequently occurring amino acids, whereas rare amino acids received a longer codon length[45]. A similar encoding strategy was later applied for the storage of text-files, through the use of Hoffmann encodings[46]. 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 one is that it is not possible to include any numbers, as the frequency of the numbers would be heavily text depended. This problem was solved by Ailenberg and Rotstein, who defined DNA codons for every character on the computer key board[47]. A second disadvantage is the absence of a clear pattern. This poses a problem mainly for long term storage, as the reader, not aware of the meaning, might confuse it with natural DNA and discard the message[48]. The problematic absence of a clear pattern was overcome via the introduction of primers along the DNA chain containing the messages, e.g. at every 500 nucleotides of data. This created a pattern comparable to the intron (primer) and exon (data) structure of DNA, which allowed for easy pattern identification. The presence of a clear pattern also made it possible to recognize mutations in the chain, which caused a shift in the reading frame[47].

Conversion of bits into nucleotides

The most obvious way to store binary data in DNA is by directly assigning two bits to every nucleotide (A = 00, C = 01, G = 10, T = 11), this creates four different states (0, 1, 2, 3) instead of two, achieving the maximal Shannon capacity, as mentioned earlier. By storing two bits per nucleotide, one makes optimal use of the four bases of DNA. This code, however, has no protection to errors[33]. A ternary code (0, 1, 2) can be used instead of a quaternary code to prevent the synthesis of homo-sequences, which cause a large amount of errors during sequencing[34]. In the case of 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 nucleotides[49]. Another variant to this code include the storage of 1 bit per nucleotide (for instance A, C = 0; T, G =1) used by Church et al[50]. This direct conversion offers, however, no protection to 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, which were specifically assigned to encode words in a text and are therefore almost not used anymore[48,51].

Storage in biological polymers

Storage in DNA in practice

In 1996 Davis was one of the first scientists to store a message in DNA by encoding a binary file, representing a single image[52]. The used encoding scheme, however, was inaccurate, as there was no distinction between a 0 and a 1. The four DNA bases were only used to determine how large a repeat of 0's and 1's was (C = 1, T = 2, A = 3, G = 4), for instance, 100111 was encoded as CTA. This encoding strategy leads to problems when the sequence is decoded, as every nucleotide can represent a repeat of 0's or 1's, meaning that CTA decodes to 100111, but also 011000[52]. Two years later the Genesis project was started by Eduardo Kac[27]. One sentence was encoded in two steps, first into a Morse code and subsequently into DNA (C = dot, T = dash, A = word space, G = letter space). The sentence was made as a synthetic gene and fused into bacteria. Ultraviolet light, however, was found to cause mutations and altered the message[27]. Another early attempt in 1999 included the storage of a 23 character long message, hidden in a DNA microdot[53]. Unique about this attempt was the use of two primers, flanking the DNA sequence, which enabled the use of the polymerase chain reaction (PCR) to amplify the message.

One of the remarkable early studies of large scale data storage was performed by Church et al., who encoded the draft version of an entire book, including 53.426 words, 11 images and one JavaScript program in DNA[50]. Instead of constructing one long strand, several smaller fragments were made, together encoding the entire binary file. All the data was first converted into bits, which were then translated to DNA nucleotides using a simple encoding strategy of 1 bit per base (A, C = 0; T, G = 1) (Figure 4). The entire sequence was split into non-overlapping strands. These strands included the data as well as a 19-nt 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 megabit, with a total of 10 errors[50].

Most of the errors encountered by Church et al. were caused by homopolymer runs and lack of coverage. To improve on Church's work, 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 file[49]. The encoding strategy used a tertiary code in combination with the Huffman encoding (*vide supra*), to compress the data. The original data was converted into base-3 digits (0,1,2) and every trit was converted into a single nucleotide, where the exact nucleotide depended on the trit and on the previous nucleotide, preventing identical consecutive nucleotides, and thus homopolymers. The obtained sequence was split into several DNA strands, containing data, indexing, and 1 nucleotide to indicate the orientation. In addition, also a parity check was included as another 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 of 75 nucleotides, meaning that every segment started with an offset of 25 nucleotides from the previous strand, which resulted in a fourfold redundancy (Figure 4)[49]. Of the five encoded files, four could be recovered without any errors. The fifth file contained two gaps of 25 nucleotides, where none of the four overlapping segments was sequenced. By taking the neighboring regions into account, the gaps could be manually filled with the missing nucleotides, after which the last file was also decoded successfully[49]. Altogether, a storage density of $2.2 \cdot 10^6$ GB·g^{-1} was achieved[49,54].

So far, the highest information density has been achieved by Erlich et al., who stored 17.1 megabit of information in DNA oligonucleotides with a density of 1.57 bits per nucleotide[32]. To realize this high density, an advanced erasure correcting encoding algorithm was used: a so-called fountain code with Luby-transform[55]. The used encoding strategy fragmented the binary sequence into non-overlapping segments, which were randomly combined to a single bit stream, called a droplet, by XOR-encoding. An identification tag was added in the form of a seed, to identify which segments were combined in the droplet (Figure 4). The droplets (the XOR code and the seed) were converted into a DNA sequence, by translating 00,01,10,11 to A, C, G, T, respectively. 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 to 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 (Figure 4). 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 proved the robustness of the encoding strategy[32].

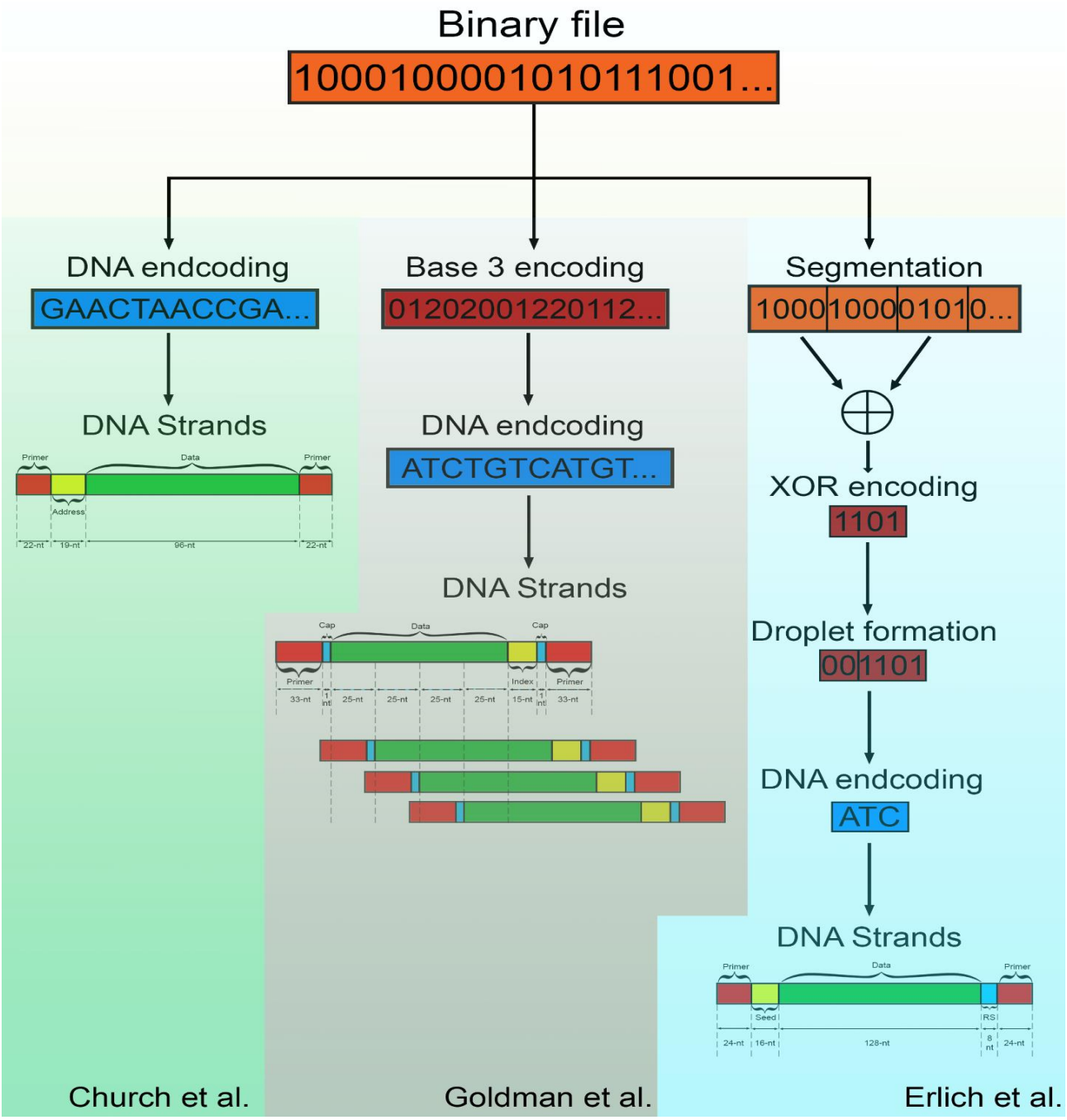

Figure 4| **Different encoding strategies for the storage of binary data in DNA.** Church et al. encoded the binary data using a 1 bit per nucleotide strategy. The data was stored into strands of 159 nucleotides, containing: 22 nucleotides as primers (red), 19 nucleotides as address labels (yellow) and 96 nucleotides as data (green)[50]. Goldman and et al. converted the binary file to base-3 digits, which were converted to DNA strands of 183 nucleotides, each containing 33 nucleotides as primers (red), 1 nucleotide as cap (blue), 15 nucleotides for indexing and parity (yellow), and 100 nucleotides containing data (green). Each strand started with an offset of 25 nucleotides, which resulted in a fourfold redundancy[49]. Erlich et al. divided the binary data into non-overlapping segments. These segments were randomly combined into droplets by XOR-encoding and an identification tag was added to identify which segments were combined. The droplets were converted into a DNA sequence, using a 2 bit per nucleotide strategy. The sequences were stored

into strands of 200 nucleotides, containing 24 nucleotides as primers (red), 16 nucleotides as seed (yellow), 8 nucleotides as Reed-Solomon codes (blue), and 128 nucleotides as data (green)[32].

These above examples are nice proofs of concepts in which commercial synthesis protocols and standard sequencing techniques are applied. There is much to win if one is able to develop faster writing and reading procedures. Also the materials aspects of encoding require further attention and need to be improved. Here lie challenging tasks for synthetic chemists and materials scientists.

Rewritable and random-access DNA storage

One of the major drawbacks of DNA storage is the time it takes to find and read the data, compared to silicon-based devices. If reading would occur on enzymatic speed (~100 nucleotides s^{-1}), the reading time would still be about seven orders of magnitude slower than that used by conventional hard drives (~10 GBits $s^{-1}s^{-1}$)[37,56]. Furthermore, in the previous described methods for DNA-based storage systems, one has to decode the entire sequence in order to find a specific set of bases. Besides this slow random access, another problem is the rewriting of the stored data. The methods discussed so far only represent the data in a read-only format, making it difficult to apply these systems to data storage that is subjected to change or needs regular updates. Researchers have tried to overcome the two major drawbacks of DNA data-storage, i.e. random access and rewritability. The first problem could be solved by using a barcode to store specific data in specific wells or pools (Figure 5)[33]. 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 needed, this strategy allows one to select the pool containing the desired data before decoding, limiting the amount of DNA strands that needs to be sequenced[33].

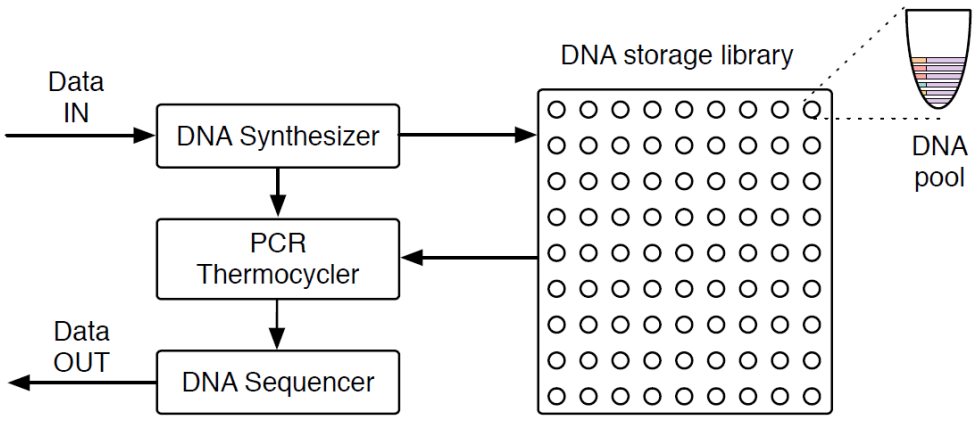

Figure 5| **DNA storage library.** Schematic drawing of a DNA storage system where DNA sequences are stored in pools. A specific piece of data, thus a specific sequence, can be chosen by selecting the right pool. Primers can be added to the pool, still containing multiple different strands, to selective amplify the desired strand. Adopted from Bornholt et al[33].

The second drawback, rewritability, can be solved by storing multiple copies of every strand, selecting one, and modify it while the other ones remain untouched for usage another time. Other methods include the use of specific enzymes that invert and restore specific DNA sequences[57,58] and the chemical transformation of DNA bases, e.g. the selective modification of cytosine to uracil[59–62].

An interesting DNA-based coding system that allowed random access and rewritability was developed by Yazdi et al[63]. This storage system only contained written text, which was kept in long strands of 1000 nucleotides, including specialized address strings that could be used for selective information access. The encoding strategy used codons of 21 nucleotides, where every codon corresponded to a single word. This fixed codon length was used to make rewriting as easy as possible and to prevent propagation errors. Rewriting was made possible by two DNA editing techniques: gBlock and Overlap Extension PCR (OE-PCR).[63,64] The gBlock method was used for short rewrites, where part of the new strand containing the edited part was synthesized by the gBlock methodology, while the remaining part of the old 1000 nucleotide strand was PCR amplified. The new and old strand contained an overlap of at least 30 base pairs, which allowed the two strands to be combined (Figure 6a)[63]. This gBlock method is very efficient but also uses long and therefore expensive primers. OE-PCR is more cost-efficient and was therefore used for the rewriting of longer blocks. Using OE-PCR, rewriting was performed in steps with short primers that contained the edited parts as overhang (Figure 6b). PCR was used to amplify all the parts of the strand, which could finally be combined by the overlap between the parts, introduced by the primers (Figure 6b).[63,64] If the rewriting segment was longer than 1000 base pairs, completely 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 successfully[63].

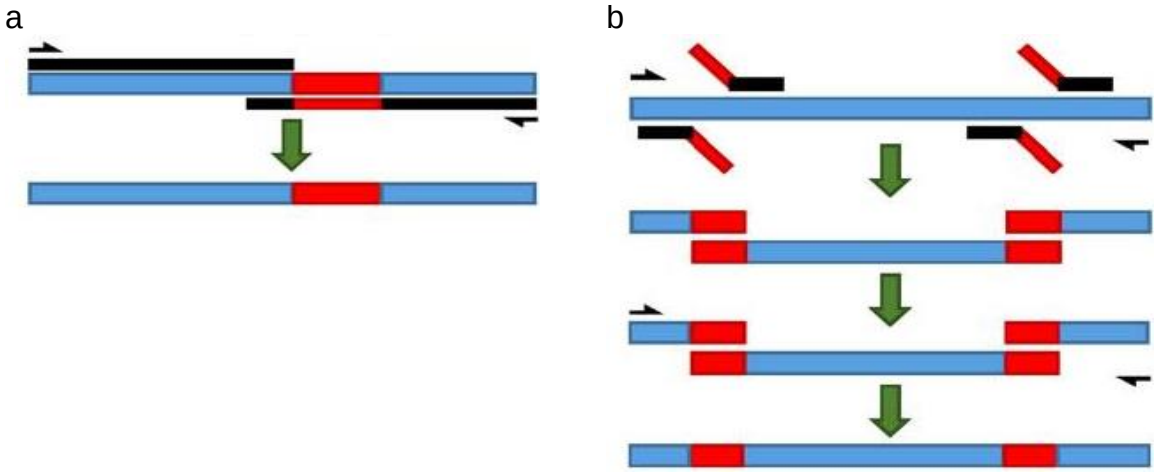

Figure 6| **The DNA editing methods gBlock and Overlap Extension PCR (OE-PCR)**. **a**| The gBlock methodology was used for short rewrites. A sequence containing the edited part of the fragment was synthesized via gBlock and the remaining part of the strand was PCR amplified. An overlap of at least 30 nucleotides was present between the two strands in order to combine both strands into one. **b**| By using 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. Adopted from Yazdi et al[63].

Storing of DNA

The way data encoded DNA strands will be stored largely depends on the purpose of the data system, nonetheless, there are some general methods. DNA can be stored on a solid support, where one end of the double stranded DNA is immobilized, which reduces the risk of unwanted aggregation of strands[65]. Instead of a solid approach, a solution approach is also possible. The latter allows for easier and faster replication and sequencing, as the molecules are more flexible and easier accessible. Furthermore, it allows for autonomous information processing and the possibility to store encoded DNA in micro-organisms[27].

DNA in solution at 4^{o}C decays within weeks and in the solid state at -80^{o}C it is stable for 3-5 years[66]. Hence, for long term storage other options need to be utilized. One possibility is the use of micro-organisms, as they can withstand extreme circumstances and can be retrieved after a long time[37,67]. Data meant for long term storage, i.e. next generations, can even be stored in different micro-organisms, to secure the highest possible chance of data recovery. The first who tried to use bacteria to store data where Yachi et al., who stored the formula $E=MC^2$ in the genomic DNA of B. subtillius[68]. In later research E. coli was used as a storage device, which resulted in a storage capacity of 1 kilobyte per cell[69]. More recently, Church et al. stored a digital movie in bacteria using the CRISPR–Cas technique and allowed correct retrieval[70]. Although mutations occur in the genome of bacteria, the rate and amount should be low enough to allow correct data retrieval[71]. In addition, data should always be stored in colonies of bacteria, providing many bacteria containing data and many data strands within all bacteria.

Grass et al. explored the use of synthetic silica matrixes to store DNA[72]. 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 aging experiments revealed that data could be recovered after treating the DNA in silica at 70°C for one week, equivalent to 2000 years in central Europe, or over 2 million years at the Global Seed Vault (-18°C) [72].

DNA computation

Besides storage, DNA has also the potential to build synthetic biological circuits, comparable to electric circuits. Biological circuits can be used to solve computational problems with the help of molecular biology. To use biological circuits, the computational problem needs to be translated into biological terms, i.e. DNA. The easy modification, amplification, and stability of DNA molecules makes them suitable for engineering circuits. Furthermore, DNA computation is energy efficient and allows parallel computations in the form of chemical reactions to be performed[73,74]. Already in 1994, Adleman used DNA as a tool to solve a Hamiltonian path problem: The Traveling Salesman Problem[28]. This mathematical problem is about a salesman, who has to travel between several cities, e.g. 7. Starting in a random city, the question of the traveling salesman is: what is the shortest route that visits each city only once? (Figure 7a). When a computer would try to solve this problem, it would essentially have to consider all possible paths that visit all 7 cities (eliminating invalid paths that visit the same city multiple times). This clearly is a very time consuming process for a sequential machine. As an alternative, Adleman used DNA computation to solve this problem. Each city was represented by a unique oligomeric strand of 20 nucleotides. Paths between the cities were also represented by 20 nucleotides: the last 10 nucleotides of the starting city and the first 10 of the ending city (Figure 7b). When paths and cities were mixed, the desire of DNA

to form double stranded helixes 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, assuming there is an excess of city and path strands to make the combinations[28]. Adleman was able to generate all solutions in a few hours, after which the elimination of non-valid paths could begin. A valid solution should contain seven cities, which meant that longer or shorter strands could immediately be eliminated. Strands with duplicated cities could also be eliminated, since each city may only be visited once. This need to eliminate invalid combinations immediately shows the major drawback of this DNA-based method: the elimination of all invalid paths took Adleman seven days[28].

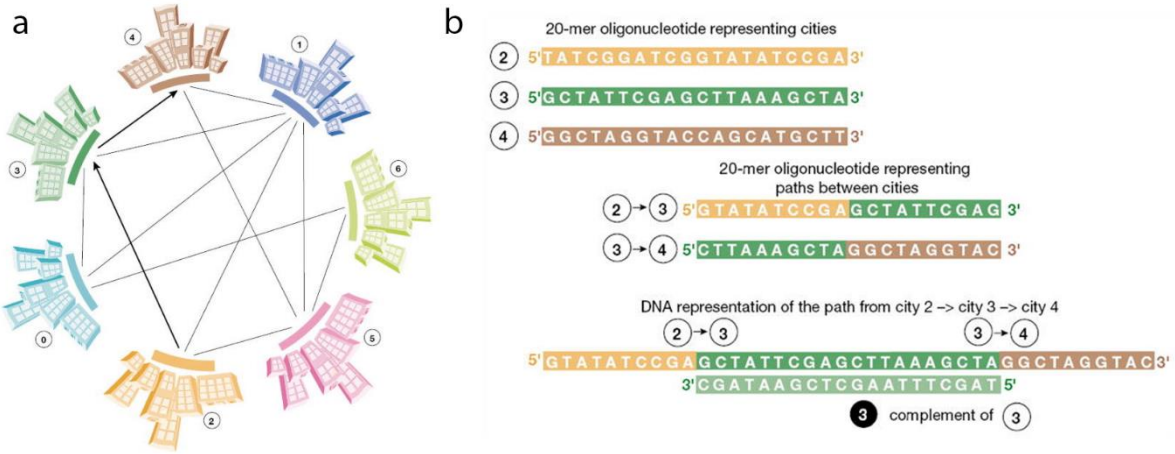

Figure 7| **DNA computing. a**| Graphical representation of the Hamiltonian path problem: The Traveling Salesman. The objective is to find the shortest route for a salesman who has to visit seven cities and visit them only once. **b**| DNA representation of The Traveling Salesman problem. Cities are 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 helixes led to all nucleotide possible combinations, from which the correct solution can be derived. Adopted from Parker[73].

Although Adleman showed that DNA can be used to solve computational problems, his approach cannot not compete with conventional silicon-based computers. Nonetheless, researchers saw the potential of parallel computation and continued research and development in this area. This led in 2002 to the development of a DNA computer, able to solve a complex computational issue: a Boolean satisfiability (SAT) problem with 20 variables, i.e. a computational problem for which the fastest known algorithms require exponential time to solve.[75] The SAT problem gives an expression in the form of a Boolean formula, consisting of AND, OR, and NOT operators, and variables, which can either be true (1) or false (0). The question is, can all variables be set to true or false to make the entire expression true. The expression for the DNA computer consisted of 24 classes, each consisting of 3 variables separated by an OR operator, e.g. (X_3 = false OR X_{16} = false or X_{18} = true) AND (X_5 = true OR X_{12} = true OR X_9 = false), etc[76]. This leads in total to more than 1 million (2^{20}) possibilities that had to be checked. Therefore, 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 another one containing one class 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 class would be captured and non-satisfying sequences moved through. Captured sequences from the first class went through the same process again, however, now with the second class of the expression, etc. Eventually, this resulted in the retrieval of the correct answer, satisfying all the classes of the expression[76].

Other DNA-based computers have been developed by Shapiro et al., who used DNA and enzymes to solve computational problems autonomously[77,78]. In these computers the hardware consisted of a restriction nuclease and ligases, and the software and input were encoded by double-stranded DNA. The automation process was based upon processing the input molecule via a cascade of restriction, hybridization, and ligation cycles, producing an output molecule encoding the computational result[77,78].

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[79,80]and DNA-based transistors[81]. Analog 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 cell[82]. 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 stimulus[83]. These DNA strands were targeted to the genome, thereby converting cellular signals into DNA-encoded memory[83]. Keinan et al. developed a more complex DNA computer, capable of iterative computation, i.e. using the output of one computation for a secondary computation process, etc. 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. Besides the possibility of iteratively computation, this DNA computer also produced biological relevant results, opening ways to regulate and change biomolecular processes[84,85].

The above-mentioned developments show that DNA has not only potential as a data storing device but also as a computer. The main drawback of DNA computation, however, lies in the extraction of the data, which still takes a huge amount of time compared to silicon-based devices.

Data storage with 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 have also the potential to act as storage devices. For use in data storage, the main focus has been on photo-switchable proteins, where 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 color transformation, triggered by absorption of a photon[86]. New possibilities were opened with the discovery of photo-convertible fluorescence proteins (PCFPs) and photo-switchable fluorescent proteins (RSFPs), which included Kaede, Dronpa, and EosFP, where the bits 0 and 1 were represented by the colors green and red, respectively[87–89]. Using IrisFP (a mutant of EosFP) color switching between red and green could be combined with switching between a dark and bright state[90,91]. Another protein used for data storage is Bacteriorhodopsin (bR), a light-activated protein from the membrane of the microbe Halobacterium salinarum. Upon

irradiation, the light is converted into chemical energy, which sets the molecule into an intermediate state for a maximum of a few days[92]. For data storage, the protein was modified, such that it could remain in this intermediate state for a few years[93]. Binary values 0 and 1 were represented by the bright state and the darks state of the protein, respectively. Encoding was performed by using a laser with a specific wavelength to set the protein into a shape representing a 0 (Figure 8a). A laser of another wavelength was used to convert the protein into a shape representing a 1 (Figure 8b). For reading, a low power laser beam was used to detect the conformation of the protein without disturbing the conformation itself (Figure 8c, d). The ability of bR to shift between different states also allows for rewritable data storage[94,95].

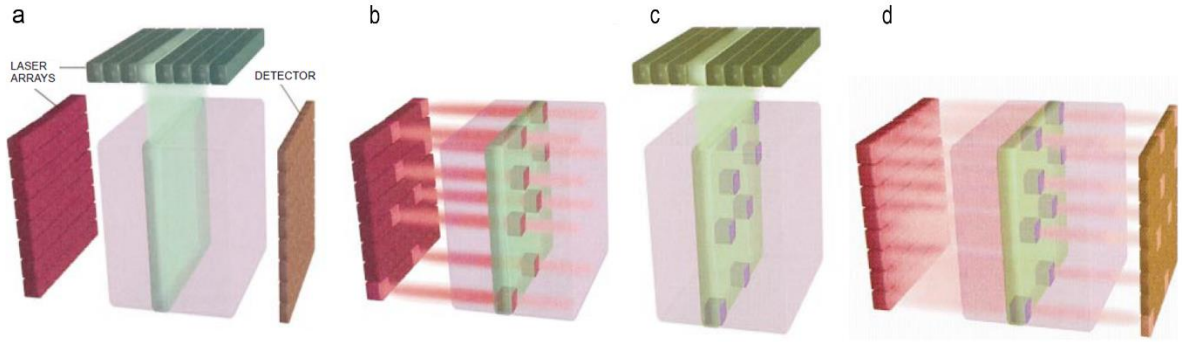

Figure 8| **Data storage with proteins. a**| Writing of data into cubes of bacteriorhodopsin (purple). First a green laser irradiates a plane of the cube, activating the photocycle. **b**| A second laser, specifically irradiates parts of the cube that should be converted to a binary state 1, while the remaining parts represent a binary 0. **c**| Reading starts by selecting a plane of the cube and activating it. **d**| A laser with a low intensity is used to shine through the cube, where parts representing a binary 0 absorb the light and parts representing a binary 1 let the light go through. This results in a bright and dark pattern on the detector. Adopted and modified from Birge [92].

Storage in synthetic polymers

Synthesis

In addition to DNA and proteins, synthetic polymers are also suitable for data storage, at least in principle. Already in 1986 Richard Dawkins mentioned that in theory every polymer could be used to store data, as long as it would be composed of at least two different monomers[96]. Although it is possible to synthesize polymers with more than two monomers in a controlled fashion, which would be more economical for data storage, most data-encoding polymers exhibit only two different monomers (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 these synthetic data-encoding copolymers, it is essential to achieve perfect control over the monomer sequence, which can be achieved by different methods, for instance biological ones. DNA can be used as a template to which free nucleotides, including non-natural ones, can associate after which the associated monomers are polymerized

chemically or by using a polymerase[97–102]. Drawbacks are the low efficiency of the process and the difficulties of removing the synthesized polymer from the template. Recently, molecular machines have been developed mimicking the biological polymerization reactions. An example includes the artificial small-molecule machine designed by Leigh and co-workers[103]. The machine is based upon a rotaxane, i.e. a molecular ring interlocked on a molecular axis, to which amino acids are attached. Upon activation the molecular ring moves along the axis and accepts the amino acids from it, leading to the synthesis of a gradually growing chain. Besides reported drawbacks in yield and kinetics, these kind of molecular machines are also limited to the synthesis of natural polymers, i.e. polypeptides[103]. To work around the natural boundaries of biological polymerization techniques, Liu et al. designed a DNA translation system to synthesize sequence-controlled polymers not based on natural monomers[104]. 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. The DNA base pairs have in this system a very similar function as tRNA, i.e. they bring the desired building blocks in the correct order to the template. Subsequent cleavage of the linker results in the release of the synthetic polymer[104].

Complete chemical polymerization has the advantage that a much wider range of building blocks are available, but achieving perfect sequence control remains a challenge. The classical chain- and step-growth polymerizations, do not allow perfect sequence control. In chain-polymerization this problem can be overcome by applying living chain polymerization methods, in which the polymer chains grow in a uniform way, as the initiation is much faster than the propagation. This leads to well-defined polymers with controlled chain lengths, while side reactions are suppressed. For instance, perfect sequence control has been obtained by living cationic and anionic polymerizations[105,106] and by radical polymerizations[107,108]. In the latter case, specific co-monomer pairs can be used to achieve an alternating pattern[109]. Lutz et al. improved on this alternating method by tuning the sequence via a time-controlled addition of the monomers[110,111]. In this strategy, the donor monomer was present in excess and polymerized by a radical reaction, while the acceptor monomer was added in small amounts. The favored donor-acceptor interaction between the monomers caused the acceptor monomers to be incorporated into small regions of the polymer backbone[110,111].

Besides chain polymerization, step growth polymerization techniques can also 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 easy and straightforward, they do not allow for perfect sequence control. New step growth polymerization techniques using radical polymerization[112,113] or click chemistry[114], however, allow for such a sequence-controlled polymerization. The latter can also be achieved by applying multi-step-growth synthesis, which involves the stepwise chemical attachment of monomers attached to a support[115]. This procedure results in very monodisperse polymers, i.e. polymer chains with the same length. One of these methods is solid-phase iterative synthesis, which is very similar to the well-known solid-phase peptide synthesis methodology. It uses an insoluble support to which the polymers are ‘grown’ by the stepwise addition of monomers[116]. This method is very efficient but also very time consuming and, furthermore, the efficiency of the coupling steps makes that this method can only be used for the synthesis of short polymers. Despite its disadvantages, solid-support

synthesis is still the most frequently and most reliable method for the synthesis of sequence-controlled polymers[115]. An alternative is the use of a soluble polymer chain as support[117], which makes the process more efficient, but the synthesis of long sequences is still not possible[100]. The group of Lutz has investigated numerous strategies to exploit this multi-step-growth methodology for the production of sequence-controlled polymers, containing data[118]. The previously mentioned step growth synthesis of polyurethanes could for instance be improved by applying a multi-step-growth approach, as was demonstrated by Lutz et al[119]. The used strategy relied on two chemo-selective steps, i.e. the reaction of an alcohol with an N-hydroxysuccinimide (NHS) moiety and the reaction of an amine with NHS. Data could be encoded using different amino alcohol monomers (serving as 0 and 1) while *N,N′*-disuccinimidyl carbonate, containing two NHS moieties was used as linker[119]. Another method developed by Lutz et al., is based upon the phosphoramidite coupling technique, which has already been used for oligonucleotide synthesis[120].The synthesis makes use of a solid support and the monomers are coupled one by one, in three steps (Figure 9a). 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 minute[121]. 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)[122]. In addition, another monomer, containing a 2,2-dipropargyl-propyl group (representing 1'), 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 polymerizatio[122]. A follow-up research improved on the phosphoramidite coupling method by using an orthogonal iterative approach, where two different building blocks are linked without the need of protecting group[123]. Furthermore, the chosen building blocks simplified the read-out by tandem mass spectrometry (MS/MS), as will be discussed in the next section[123]. Important to mention is the fact that Lutz et al. already synthesized a sequence-coding polymer using automated phosphoramidite coupling[124]. By making some small adjustments to the original protocol, i.e. by using a large excess of monomer, by applying capping steps, the synthesis and sequencing of polymers composed of more than 100 monomers could be achieved[124].

An alternative solid-phase approach to achieve complete sequence control without the need of protecting groups is the ''AB + CD'' method, also developed by Lutz et al.[117]. It makes use of two different building blocks, each containing two different functional groups AB (A = carboxylic acid, B = alkyne) and CD (C = amine, D = azide). Protective groups are not necessary as A can only react with C by amidification and B can only react with D by a copper-catalyzed azide-alkyne cycloaddition. To use this synthesis protocol for data encoding, Lutz et al. chose two different AB building blocks, representing 0 and 1, while the CD building block was used as a spacer (Figure 9b)[125]. To simplify the ''AB+CD'' synthesis, four different AB dimers can be used, representing 00, 01, 10, and 1[126]. This reduces the amount of coupling steps needed 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 the monomers to proceed via consecutive anhydride-amine and nitroxide radical reactions (Figure 9c). Repeating these steps allowed the synthesis of sequence-controlled polymers that were easy to read and easy to eras[127].

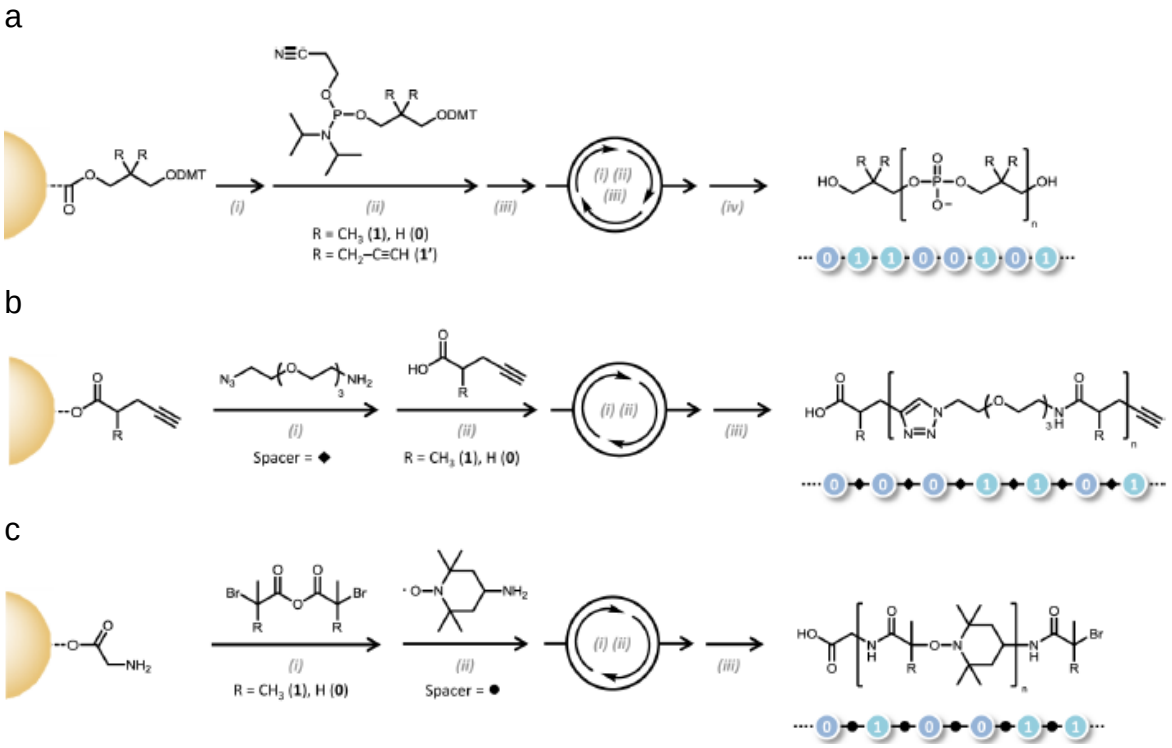

Figure 9| **Different strategies for the synthesis of information-containing macromolecules**. **a**| Phosphoramidite coupling, which was used in three steps: (*i*) deprotection with N,N-dimethyltryptamine (DMT), (*ii*) coupling of the next monomer, representing 1, 0, or 1', (*iii*) oxidation of the phosphite bond to a phosphate, (*iv*) cleavage from the resin. **b**| The ''AB + CD'' method, involving three different monomers, representing 0,1, and a spacer. (*i*) Coupling of the spacer (CD) by an azide-alkyne copper-catalyzed cycloaddition, (*ii*) coupling of a monomer representing either a 0 or a 1 (AB) by amidification, (*iii*) cleavage from the resin. **c**| Accelerated ''AB + CD'' method using, again, three 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) by a nitroxide radical reaction, (*iii*) cleavage from the resin. Adopted from Lutz[118].

Another strategy to obtain sequence-controlled polymers was reported by Zydziak et al, which was not based upon solid-support synthesis but on photoligation of six different monomers[128]. Each monomer contained a dienophile and a benzaldehyde, where the latter one could be converted to a diene upon irradiation. This reactive diene could than react with a dienophile moiety of another monomer by a Diels-Alder reaction. This photochemical concept yielded monodisperse compounds and instead of a chemical deprotection step, light was used to obtain the reactive moieties[128].

Reading and rewriting

For the reading (sequencing) of biopolymers, i.e. DNA, very fast and automated methods are available. Unfortunately, these methods are not applicable to synthetic polymers and more universal analysis procedures must be used, one of which is tandem mass spectrometry (MS/MS)[129–131]. In the latter method the polymers to be sequenced are ionized and separated based on their mass-to-charge ratio, after which the ions are fragmented,

separated, and detected. The obtain fragments can subsequently be used to reconstruct the precursor ion and ultimately the polymer sequence. The obtained fragmentation pattern depends strongly on the nature of the backbone, which gives synthetic polymers an advantage, as their molecular structures can be altered to favor an easy read out[132]. The previous mentioned phosphoramidite coupling results, for instance, in an easy fragmentation pattern, in which the phosphate bonds are easily ionized and dissociated in MS/MS. Introduction of alkoxyamine bonds, with a lower dissociation energy, along the chain even simplified the readout by introducing two dissociation energies, i.e. cleavage of the alkoxyamine bond, generating large fragments, and cleavage of the phosphate bonds, generating smaller fragments[132–134]. The accelerated ‘‘AB + CD’’ synthesis, mentioned above also employs poly(alkoxyamine amides) with ‘weak links’ between the monomers (AB and CD) allowing a fast readout for even long chains[118,127,135]. Charles et al. also showed that poly(triazole amide)s containing a specific sequence could be easily decoded by MS/MS as it generates two products upon cleavage, i.e. the amide bond and the ether bond[136]. Specific software is nowadays available, allowing one to decipher the sequence of these polymers in a few milliseconds, showing the great advantage of synthetic digital polymers over biological ones[137].

NMR can also be used to sequence a polymer. For a long time ^{13}C NMR was one of the most used methods to identify short synthetic copolymers, however, as its sensitivity is limited, the usage for long polymers is problematic. To elucidate the sequence of larger macromolecules, the electrical birefringence (Kerr effect) of a polymer solution in an electric field can be measured[138]. 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 holds potential as an interesting technique in the future[138,139]. Another NMR technique especially suitable for non-natural polymers is the tweezer technique[140–142]. This method uses molecular reporters (tweezers) that can bind along the polymer chain by 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 to quantify.

A new promising sequencing technique for both natural and synthetic polymers is nano-pore sequencing, which analyzes the polymer structure by pulling it through a biological or synthetic pore. When the molecule moves through the pore, it changes the current through the channel in a way that is characteristic for the molecule. The specific changes of the current can be used to identify the primary structure of polymers[143]. This technique was first introduced by Kasianowicz et al., who used a biological nanopore (α-hemolysin) to sequence a single-stranded DNA molecule[144]. Later, modifications on the surface of the channel showed that the nanopore could identify numerous features, for instance, the 3’ and 5’ ends of a DNA chain[145,146]. So far, only a small number of studies using nanopore sequencing for synthetic polymers has been reported, including PEG macromolecules, polystyrene sulfonate, dextran sulphate, and poly(phosphodiesters)[147–151]. More recently, some theoretical studies have been performed on more complex polymers, such as branched polymers and heterogeneous copolymers with charged and uncharged blocks[152,153]. These results show that nanopore sequencing might become a good and reliable

method in the future, but also that the successful readout will depend strongly on the charge, stiffness, and the conformation of the polymer chain[118].

An advantage of synthetic polymers when compared to biopolymers is the ability to tune them for properties such as degradation and rewriting of data. 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 erased[127]. This procedure will, however, break all linkers between the polymeric units, erasing all data, which means that the complete polymer has to be resynthesized. To prevent complete re-synthesis, Lutz et al. developed a monomer, which allowed modification after polymerization, by using a Huisgen azide-alkyne cycloaddition. In this way changes could be made to the code, while keeping the original polymer intact[122]. Another technique for rewriting data uses dynamic polymers, as designed by Lehn et al[154]. These polymers are based upon a hydrazide and an aldehyde, which form an acylhydrazone bond by a condensation reaction[155,156]. Acylhydrazone formation is, however, reversible under mild acidic conditions[156]. In the presence of other hydrazides or aldehydes, this reversibility could be exploited to create new acylhydrazines and thus rewrite the data. Although rewriting of synthetic polymers is possible, so far practically nothing has been done in rewriting data on synthetic polymers. Hence, tools to selectively change data on synthetic polymers still have to be developed.

Writing by catalytic methods

Nature makes use of catalytic procedures to write information, as is the case for the synthesis of proteins on ribosomes, in which m-RNA acts as the reading template. Also the copying of DNA by e.g. the DNA polymerase III enzyme system is an example of catalytic writing in nature[157]. An important aspect of this writing is that it takes place in a processive fashion, meaning that the catalyst remains in contact with the polymeric substrate without detaching. In this way a large number of sequential writing events can take place, which reduces the chance of errors. Processive catalysis is the opposite of distributive catalysis, in which the catalyst (enzyme) and substrate only meet once and after reaction separat[158]. The group of Nolte has developed a biomimetic catalytic system that can specifically cleave DNA chains at AAA sites, which is a first step in the direction of writing (Figure 10a)[159]. The catalyst is composed of the trimeric ring-shaped protein clamp (gp45) of the bacteriophage T4, which is associated with the replication polymerase (gp43). The group has replaced the replication polymerase by three manganese porphyrin complexes, which in the presence of an oxidant act as cleaving catalysts. The authors show that the modified clamp can bind to DNA and move along it uni-directionally, while cleaving the AAA sites.

Another example from the same group involves a completely synthetic system that can write epoxides on a high molecular weight polybutadiene chain with the help of a porphyrin cage catalyst and an oxidant. The catalysts 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 (Figure 10b)[158,160,161].

a b

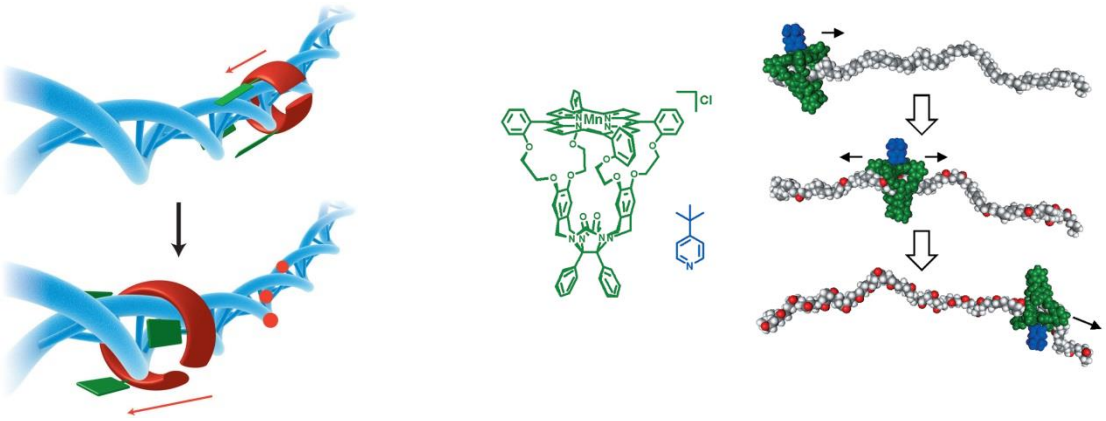

Figure 10| **Catalytic writing. a**| Bio-hybrid catalyst composed of a protein ring to which three manganese porphyrin catalysts have been attached. The catalyst cleaves DNA at AAA sites while moving along it. **b**| Synthetic catalyst constructed from a diphenylglycoluril cage compound and a manganese porphyrin complex. The catalyst threads onto polybutadiene and converts the double bonds of this polymer into epoxide functions while gliding along it. Adopted from van Dongen et al.[158] and Prins et al.[164]

Outlook

With the boundaries of silicon-based storage devices already in sight, attention has to be given to alternative solutions for storing information. As has been shown in this review DNA-based storage may become an interesting alternative for the current storage technology, especially in terms of storage density. Over the years much progress has been made, especially with regard to error protection mechanisms without giving up too much on storage density. One major disadvantage of DNA compared to silicon is the much lower reading speed, which is problematic, especially for use as random access memories, when only a small part of the data is desired. This makes that for now, DNA is only applicable for archiving and long term data storage. A major problem, still to be overcome, is the current cost of DNA synthesis compared to the costs of silicon-based storage facilities. However, assuming a similar decline in costs as was the case for the silicon-based storage media and considering the fact that the DNA technology can be expected to improve further, it is likely that it will not take long before DNA-based storage is the standard for long term data storage[25,162,163]. This certainly is the case when also the costs of maintenance and storage are taken into account. These are significant smaller for DNA-based storage systems than for the silicon-based systems in the current data centers. Furthermore, cost reduction could already be achieved quite rapidly and easily by using quicker but less reliable synthesis protocols, which require less time and reagents. Lower reliability will result in less valid DNA strands, but as the DNA fountain code already showed, this can be compensated by using robust and high-flexible coding strategies[32].

Future research will have to show whether DNA reading and rewritability can be improved, which will make DNA storage practical for use of data that changes on a more daily basis. More interesting for short-term applications, however, might be data storage systems based on synthetic polymers, which can be prepared from a much larger set of

monomers than biopolymers and are more stable. Furthermore, such synthetic systems do not require biological machineries and can be tuned for quick read-out and rewritability. However, when compared to DNA-based storage systems, the field of synthetic encoded polymers is still in its infancy. It can be 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, e.g. in terms of the monomers. Of great fundamental interest are the systems that encode information into bio- 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 make a step forward in mimicking this fascinating process. If processive catalytic systems based on clamp-shaped proteins and attached enzyme writers, readers, and erases, as known from histones, the chief protein components of chromatin, can be constructed, also other possibilities come within reach, e.g. the construction of bio-computers.

Altogether we may conclude that DNA-based storage devices have a clear potential to become a good and reliable alternative for long term data storage. 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.

References

1. Woods, C. *Visible Language: Inventions of Writing in the Ancient Middle East and Beyond*. *Oriental Institute Museum Publications* (Oriental Institute of the University of Chicago, 2010).
2. Gantz, J. & Reinsel, D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east. *IDC iView IDC Anal. Futur.* **2007,** 1–16 (2012).
3. Trends and Analysis Cisco Company. *White Pap.* June (2017).
4. Extance, A. How DNA could store all the world's data. *Nature* **537,** 22–24 (2016).
5. Zhirnov, V., Zadegan, R. M., Sandhu, G. S., Church, G. M. & Hughes, W. L. Nucleic acid memory. *Nat. Mater.* **15,** 366–370 (2016).
6. Lunt, B. M. How Long Is Long-Term Data Storage? *Arch. Conf.* **2011,** 29–33 (2011).
7. Shrivastava, S. & Badlani, R. Data Storage in DNA. *Int. J. Electr. Energy* **2,** 119–124 (2014).
8. Kumar, S. & Vijayaraghavan, R. Solid State Drive (SSD) FAQ. *Dell* (2011). Available at: https://www.dell.com/downloads/global/products/pvaul/en/solid-state-drive-faq-us.pdf. (Accessed: 8th May 2018)
9. Greengard, S. Cracking the Code on DNA Storage. *Commun. ACM* **60,** 16–18 (2017).
10. Greenberg, A., Hamilton, J., Maltz, D. A. & Patel, P. The cost of a cloud: research problems in data center networks. *ACM SIGCOMM Comput. Commun. Rev.* **39,** 68–73 (2008).
11. Bawden, T. Global warming: Data centres to consume three times as much energy in next decade, experts warn. *The Independent* (2016).

12. Ritter, S. DNA To The Rescue For Data Storage. *Chem. Eng. News Arch.* **93,** 40–41 (2015).

13. Stikeman, A. Polymer Memory. *Technol. Rev.* **105,** 31 (2002).

14. Colquhoun, H. & Lutz, J.-F. Information-containing macromolecules. *Nat. Chem.* **6,** 455 (2014).

15. Ogihara, M. & Ray, A. DNA computing on a chip. *Nature* **403,** 143 (2000).

16. Kosuri, S. & Church, G. M. Large-scale de novo DNA synthesis: technologies and applications. *Nat. Methods* **11,** 499–507 (2014).

17. Shendure, J. *et al.* DNA sequencing at 40: past, present and future. *Nature* **550,** 345–353 (2017).

18. 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).

19. Maxam, A. M. & Gilbert, W. A new method for sequencing DNA. *Proc. Natl. Acad. Sci.* **74,** 560–564 (1977).

20. Mullis, K. B. & Faloona, F. A. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods Enzymol.* **155,** 335–350 (1987).

21. Kelly, T. J. J. & Smith, H. O. A restriction enzyme from Hemophilus influenzae. II. *J. Mol. Biol.* **51,** 393–409 (1970).

22. Smith, H. O. & Wilcox, K. W. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. *J. Mol. Biol.* **51,** 379–391 (1970).

23. Little, J. W., Zimmerman, S. B., Oshinsky, C. K. & Gellert, M. Enzymatic joining of DNA strands, II. An enzyme-adenylate intermediate in the dpn-dependent DNA ligase reaction. *Proc. Natl. Acad. Sci. U. S. A.* **58,** 2004–2011 (1967).

24. 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. U. S. A.* **57,** 1841–1848 (1967).

25. O' Driscoll, A. & Sleator, R. D. Synthetic DNA: the next generation of big data storage. *Bioengineered* **4,** 123–125 (2013).

26. Glanz, J. The Cloud Factories: Power, Pollution and the Internet. *New York Times* (2012).

27. De Silva, P. Y. & Ganegoda, G. U. New Trends of Digital Data Storage in DNA. *Biomed Res. Int.* 14 (2016). doi:http://dx.doi.org/10.1155/2016/8072463

28. Adleman, L. M. Molecular computation of solutions to combinatorial problems. *Science* **266,** 1021–1024 (1994).

29. Shilov, A. Western Digital Launches Ultrastar DC HC530 14 TB PMR with TDMR HDD. *Anadtech.com* (2018). Available at: https://www.anandtech.com/show/12665/western-digital-launches-ultrastar-dc-hc530-14-tb-pmr-with-tdmr-hdd. (Accessed: 30th April 2018)

30. Castillo, M. From Hard Drives to Flash Drives to DNA Drives. *Am. J. Neuroradiol.* **35,** 1–2 (2014).

31. Carlson, R. The changing economics of DNA synthesis. *Nat. Biotechnol.* **27,** 1091 (2009).

32. Erlich, Y. & Zielinski, D. DNA Fountain enables a robust and efficient storage architecture. *Science* **355,** 950–954 (2017).

33. Bornholt, J, et al. A DNA-based archival storage system. *ACM SIGOPS Oper. Syst. Rev.* **50,** 637 (2016).

34. 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).

35. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27,** 379–423 (1948).

36. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27,** 623–656 (1948).

37. Cox, J. P. Long-term data storage in DNA. *Trends Biotechnol.* **19,** 247–250 (2001).

38. Bancroft, C., Bowler, T., Bloom, B. & Clelland, C. T. Long-Term Storage of Information in DNA. *Science* **293,** 1763–1765 (2001).

39. Bogard, C. M., Rouchka, E. C. & Arazi, B. DNA media storage. *Prog. Nat. Sci.* **18,** 603–609 (2008).

40. Carrillo, H. & Lipman, D. The Multiple Sequence Alignment Problem in Biology. *SIAM J. Appl. Math.* **48,** 1073–1082 (1988).

41. Simpson, R. E. *Introductory Electronics for Scientists and Engineers*. (Addison-Wesley, 1987).

42. Reed, I. & Solomon, G. Polynomial Codes Over Certain Finite Fields. *J. Soc. Ind. Appl. Math.* **8,** 300–304 (1960).

43. Moon, T. K. *Error correction coding: Mathematical methods and algorithms* . (Wiley-Interscience, 2005).

44. Blawat, M. *et al.* Forward Error Correction for DNA Data Storage. *Procedia Comput. Sci.* **80,** 1011–1022 (2016).

45. 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).

46. Huffman, D. A. A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE* **40,** 1098–1101 (1952).

47. Ailenberg, M. & Rotstein, O. An improved Huffman coding method for archiving text, images, and music characters in DNA. *Biotechniques* **47,** 747–754 (2009).

48. 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).

49. Goldman, N. *et al.* Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature* **494,** 77–80 (2013).

50. Church, G. M., Gao, Y. & Kosuri, S. Next-Generation Digital Information Storage in

DNA. *Science* **337,** 1628–1628 (2012).

51. Golomb, S. Efficient coding for the desoxyribonucleic channel. *Proc. Fifteenth Symp. Appl. Math.* **14,** 87–100 (1962).

52. Davis, J. Microvenus. *Art J.* **55,** 70–74 (1996).

53. Clelland, C. T., Risca, V. & Bancroft, C. Hiding messages in DNA microdots. *Nature* **399,** 533–534 (1999).

54. Bourdenx, M. DNA as the next digital information storage support. *Mov. Disord.* **28,** 583 (2013).

55. MacKay, D. J. C. Fountain codes. *IEE Proceedings-Communications* **152,** 1062–1068 (2005).

56. Micheloni, R. Solid-State Drive (SSD): A Nonvolatile Storage System. *Proc. IEEE* **105,** 583–588 (2017).

57. Friedland, A. E. *et al.* Synthetic Gene Networks That Count. *Science* **324,** 1199–1202 (2009).

58. Bonnet, J., Subsoontorn, P. & Endy, D. Rewritable digital data storage in live cells via engineered control of recombination directionality. *Proc. Natl. Acad. Sci.* **109,** 8884–8889 (2012).

59. Mayer, C., McInroy, G. R., Murat, P., Van Delft, P. & Balasubramanian, S. An Epigenetics-Inspired DNA-Based Data Storage System. *Angew. Chemie Int. Ed.* **55,** 11144–11148 (2016).

60. Goldberg, A. D., Allis, C. D. & Bernstein, E. Epigenetics: A Landscape Takes Shape. *Cell* **128,** 635–638 (2007).

61. Shapiro, R., Servis, R. E. & Welcher, M. Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite. *J. Am. Chem. Soc.* **92,** 422–424 (1970).

62. 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).

63. 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).

64. Bryksin, A. V & Matsumura, I. Overlap extension PCR cloning: a simple and reliable way to create recombinant plasmids. *Biotechniques* **48,** 463–465 (2010).

65. Arita, M. in *Aspects of Molecular Computing* (eds. Jonoska, N., Păun, G. & Rozenberg, G.) 23–35 (Springer Berlin Heidelberg, 2004). doi:10.1007/978-3-540-24635-0_2

66. Anchordoquy, T. J. & Molina, M. C. Preservation of DNA. *Cell Preserv. Technol.* **5,** 180–188 (2007).

67. 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).

68. 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).

69. Limbachiya, D. & Gupta, M. K. Natural Data Storage: A Review on sending Information from now to then via Nature. *arXiv* (2015). Available at: http://arxiv.org/abs/1505.04890.

70. 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).

71. Heaven, D. Video stored in live bacterial genome using CRISPR gene editing. *New scientist* (2017). Available at: https://www.newscientist.com/article/2140576-video-stored-in-live-bacterial-genome-using-crispr-gene-editing/. (Accessed: 6th May 2018)

72. 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).

73. Parker, J. Computing with DNA. *EMBO Rep.* **4,** 7–10 (2003).

74. Scudellari, M. Inner Workings: DNA for data storage and computing. *Proc. Natl. Acad. Sci.* **112,** 15771–15772 (2015).

75. Cornen, T. H., Leiserson, R. L. & Rivest, R. L. *Introduction to Algorithms*. (The MIT Press, 1990).

76. Braich, R. S., Chelyapov, N., Johnson, C., Rothemund, P. W. K. & Adleman, L. Solution of a 20-Variable 3-SAT Problem on a DNA Computer. *Science* **296,** 499–502 (2002).

77. Benenson, Y. *et al.* Programmable and autonomous computing machine made of biomolecules. *Nature* **414,** 430–434 (2001).

78. 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.* **100,** 2191–2196 (2003).

79. Elowitz, M. B. & Leibler, S. A synthetic oscillatory network of transcriptional regulators. *Nature* **403,** 335–338 (2000).

80. Gardner, T. S., Cantor, C. R. & Collins, J. J. Construction of a genetic toggle switch in Escherichia coli. *Nature* **403,** 339–342 (2000).

81. Bonnet, J., Yin, P., Ortiz, M. E., Subsoontorn, P. & Endy, D. Amplifying Genetic Logic Gates. *Science* **340,** 599 LP-603 (2013).

82. Daniel, R., Rubens, J. R., Sarpeshkar, R. & Lu, T. K. Synthetic analog computation in living cells. *Nature* **497,** 619–623 (2013).

83. Farzadfard, F. & Lu, T. K. Genomically encoded analog memory with precise in vivo DNA writing in living cell populations. *Science* **346,** (2014).

84. 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).

85. 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).

86. Hirshberg, Y. Reversible Formation and Eradication of Colors by Irradiation at Low Temperatures. A Photochemical Memory Model. *J. Am. Chem. Soc.* **78,** 2304–2312 (1956).

87. Adam, V. *et al.* Data storage based on photochromic and photoconvertible fluorescent proteins. *J. Biotechnol.* **149,** 289–298 (2010).

88. 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. U. S. A.* **99,** 12651–12656 (2002).

89. Ando, R., Mizuno, H. & Miyawaki, A. Regulated fast nucleocytoplasmic shuttling observed by reversible protein highlighting. *Science* **306,** 1370–1373 (2004).

90. Adam, V. *et al.* Structural characterization of IrisFP, an optical highlighter undergoing multiple photo-induced transformations. *Proc. Natl. Acad. Sci.* **105,** 18343–18348 (2008).

91. Barachevskii, V. F. M. and V. A. M. and V. A. Nonlinear coloration of photochromic spiropyran solutions. *Sov. J. Quantum Electron.* **3,** 128 (1973).

92. Birge, R. R. Protein-Based Computers. *Sci. Am.* **272,** 90–95 (1995).

93. Renugopalakrishnan, V. *et al.* Retroengineering bacteriorhodopsins: Design of smart proteins by bionanotechnology. *Int. J. Quantum Chem.* **95,** 627–631 (2003).

94. 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).

95. Oesterhelt, D., Brauchle, C. & Hampp, N. Bacteriorhodopsin: a biological material for information processing. *Q. Rev. Biophys.* **24,** 425–478 (1991).

96. Dawkins, R. *The blind watchmaker*. (Longman, 1986).

97. Orgel, L. E. Molecular replication. *Nature* **358,** 203–209 (1992).

98. Sievers, D. & von Kiedrowski, G. Self-replication of complementary nucleotide-based oligomers. *Nature* **369,** 221–224 (1994).

99. 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).

100. Lutz, J.-F., Ouchi, M., Liu, D. R. & Sawamoto, M. Sequence-controlled polymers. *Science* **341,** 1238149 (2013).

101. 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).

102. Kool, E. T. Replacing the Nucleobases in DNA with Designer Molecules. *Acc. Chem. Res.* **35,** 936–943 (2002).

103. Lewandowski, B. *et al.* Sequence-Specific Peptide Synthesis by an Artificial Small-

Molecule Machine. *Science* **339,** 189 LP-193 (2013).

104. Niu, J., Hili, 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).

105. 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).

106. Hadjichristidis, N., Pitsikalis, M., Pispas, S. & Iatrou, H. Polymers with Complex Architecture by Living Anionic Polymerization. *Chem. Rev.* **101,** 3747–3792 (2001).

107. 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).

108. Tong, X., Guo, B. & Huang, Y. Toward the synthesis of sequence-controlled vinyl copolymers. *Chem. Commun.* **47,** 1455–1457 (2011).

109. Rzaev, Z. M. O. Complex-radical alternating copolymerization. *Prog. Polym. Sci.* **25,** 163–217 (2000).

110. 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).

111. 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).

112. 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).

113. 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).

114. 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).

115. 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).

116. Badi, N. & Lutz, J.-F. Sequence control in polymer synthesis. *Chem. Soc. Rev.* **38,** 3383–3390 (2009).

117. Pfeifer, S., Zarafshani, Z., Badi, N. & Lutz, J.-F. Liquid-Phase Synthesis of Block Copolymers Containing Sequence-Ordered Segments. *J. Am. Chem. Soc.* **131,** 9195–9197 (2009).

118. Lutz, J.-F. Coding Macromolecules: Inputting Information in Polymers Using Monomer-Based Alphabets. *Macromolecules* **48,** 4759–4767 (2015).

119. Gunay, U. *et al.* Chemoselective Synthesis of Uniform Sequence-Coded Polyurethanes and Their Use as Molecular Tags. *Chem* **1,** 114–126 (2016).

120. Beaucage, S. L. & Caruthers, M. H. Deoxynucleoside phosphoramidites—A new class of key intermediates for deoxypolynucleotide synthesis. *Tetrahedron Lett.* **22,** 1859–1862 (1981).

121. Beaucage, S. L. & Iyer, R. P. Advances in the Synthesis of Oligonucleotides by the Phosphoramidite Approach. *Tetrahedron* **48,** 2223–2311 (1992).

122. Al Ouahabi, A., Charles, L. & Lutz, J.-F. Synthesis of Non-Natural Sequence-Encoded Polymers Using Phosphoramidite Chemistry. *J. Am. Chem. Soc.* **137,** 5629–5635 (2015).

123. 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).

124. 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).

125. 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).

126. 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).

127. Roy, R. K. *et al.* Design and synthesis of digitally encoded polymers that can be decoded and erased. *Nat. Commun.* **6,** 7237 (2015).

128. Zydziak, N. *et al.* Coding and decoding libraries of sequence-defined functional copolymers synthesized via photoligation. *Nat. Commun.* **7,** 13672 (2016).

129. Mutlu, H. & Lutz, J.-F. Reading Polymers: Sequencing of Natural and Synthetic Macromolecules. *Angew. Chemie Int. Ed.* **53,** 13010–13019 (2014).

130. 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).

131. Altuntaş, 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).

132. 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).

133. 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).

134. J.- A., 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).

135. 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).

136. 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).

137. Burel, A., Carapito, C., Lutz, J.-F. & Charles, L. MS-DECODER: Milliseconds Sequencing of Coded Polymers. *Macromolecules* **50,** 8290–8296 (2017).

138. Tonelli, A. E. A Case for Characterizing Polymers with the Kerr Effect. *Macromolecules* **42,** 3830–3840 (2009).

139. Hardrict, 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).

140. Colquhoun, H. M. & Zhu, Z. Recognition of Polyimide Sequence Information by a Molecular Tweezer. *Angew. Chemie Int. Ed.* **43,** 5040–5045 (2004).

141. 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).

142. 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).

143. Meller, A., Nivon, L., Brandin, E., Golovchenko, J. & Branton, D. Rapid nanopore discrimination between single polynucleotide molecules. *Proc. Natl. Acad. Sci.* **97,** 1079–1084 (2000).

144. Kasianowicz, J. J., Brandin, E., Branton, D. & Deamer, D. W. Characterization of individual polynucleotide molecules using a membrane channel. *Proc. Natl. Acad. Sci.* **93,** 13770–13773 (1996).

145. 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).

146. Wanunu, M. & Meller, A. Chemically Modified Solid-State Nanopores. *Nano Lett.* **7,** 1580–1585 (2007).

147. Bezrukov, S. M., Vodyanoy, I., Brutyan, R. A. & Kasianowicz, J. J. Dynamics and Free Energy of Polymers Partitioning into a Nanoscale Pore. *Macromolecules* **29,** 8517–8522 (1996).

148. 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.* **107,** 12080–12085 (2010).

149. Movileanu, L. & Bayley, H. Partitioning of a polymer into a nanoscopic protein pore obeys a simple scaling law. *Proc. Natl. Acad. Sci.* **98,** 10137–10141 (2001).

150. Gibrat, G. *et al.* Polyelectrolyte Entry and Transport through an Asymmetric α-Hemolysin Channel. *J. Phys. Chem. B* **112,** 14687–14691 (2008).

151. Mordjane, B. *et al.* Translocation of Precision Polymers through Biological Nanopores. *Macromol. Rapid Commun.* **38,** 1700680 (2017).

152. Sakaue, T. & Brochard-Wyart, F. Nanopore-Based Characterization of Branched Polymers. *ACS Macro Lett.* **3,** 194–197 (2014).

153. Mirigian, S., Wang, Y. & Muthukumar, M. Translocation of a heterogeneous polymer. *J. Chem. Phys.* **137,** 64904 (2012).

154. Skene, W. G. & Lehn, J.-M. P. Dynamers: Polyacylhydrazone reversible covalent polymers, component exchange, and constitutional diversity. *Proc. Natl. Acad. Sci. United States Am.* **101,** 8270–8275 (2004).

155. Bunyapaiboonsri, T. *et al.* Dynamic Deconvolution of a Pre-Equilibrated Dynamic Combinatorial Library of Acetylcholinesterase Inhibitors. *ChemBioChem* **2,** 438–444 (2001).

156. Nguyen and Ivan Huc, R. Optimizing the reversibility of hydrazone formation for dynamic combinatorial chemistry. *Chem. Commun.* 942–943 (2003).

157. Clark, D. P. & Pazdernik, N. J. in *Biotechnology* 97–130 (Academic Cell, 2016).

158. van Dongen, S. F. M., Elemans, J. A. A. W., Rowan, A. E. & Nolte, R. J. M. Processive catalysis. *Angew. Chem. Int. Ed. Engl.* **53,** 11420–11428 (2014).

159. van Dongen, S. F. M. *et al.* A clamp-like biohybrid catalyst for DNA oxidation. *Nat. Chem.* **5,** 945 (2013).

160. 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).

161. 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. *European J. Org. Chem.* **2007,** 751–757 (2007).

162. Carr, P. A. & Church, G. M. Genome engineering. *Nat. Biotechnol.* **27,** 1151–1162 (2009).

163. Feher, T., Burland, V. & Posfai, G. In the fast lane: large-scale bacterial genome engineering. *J. Biotechnol.* **160,** 72–79 (2012).

164. Prins, L.J., & Scrimin, P. Processive catalysis: Thread and cut. *Nature Chem.* **5**, 899 (2013).

Acknowledgements

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