Procedia Computer Science

Volume 80, 2016, Pages 1011–1022

ICCS 2016. The International Conference on Computational Science

Forward Error Correction for DNA Data Storage

Meinolf Blawat[1], Klaus Gaedke[1], Ingo Hütter[1], Xiao-Ming Chen[1], Brian Turczyk[2], Samuel Inverso[2], Benjamin W. Pruitt[2], George M. Church[2]

[1]*Technicolor Research & Innovation, Hannover, Germany*

[2]*Harvard Medical School, Wyss-Institute, Boston, US.*

meinof.blawat@technicolor.com, klaus.gaedke@technicolor.com, ingo.huetter@technicolor.com, xiao-ming.chen@technicolor.com, brian.turczyk@wyss.harvard.edu, samuel.inverso@wyss.harvard.edu, benjamin.pruitt@wyss.harvard.edu, gmc@harvard.edu

Abstract

We report on a strong capacity boost in storing digital data in synthetic DNA. In principle, synthetic DNA is an ideal media to archive digital data for very long times because the achievable data density and longevity outperforms today's digital data storage media by far. On the other hand, neither the synthesis, nor the amplification and the sequencing of DNA strands can be performed error-free today and in the foreseeable future. In order to make synthetic DNA available as digital data storage media, specifically tailored forward error correction schemes have to be applied.

For the purpose of realizing a DNA data storage, we have developed an efficient and robust forward-error-correcting scheme adapted to the DNA channel. We based the design of the needed DNA channel model on data from a proof-of-concept conducted 2012 by a team from the Harvard Medical School [1]. Our forward error correction scheme is able to cope with all error types of today's DNA synthesis, amplification and sequencing processes, e.g. insertion, deletion, and swap errors.

In a successful experiment, we were able to store and retrieve error-free 22 MByte of digital data in synthetic DNA recently. The found residual error probability is already in the same order as it is in hard disk drives and can be easily improved further. This proves the feasibility to use synthetic DNA as long-term digital data storage media.

Keywords: Bio-technology, DNA, DNA synthesis and sequencing, digital data storage, data preservation, archiving

1 Introduction

Modern DNA synthesizers can concatenate the four DNA base nucleotides, namely Adenine (A), Cyanine (C), Guanine (G), and Thymine (T) to form masses of almost arbitrary chain molecules. Each nucleotide (nt) in such a chain called oligo can be considered to represent in principle 2 bit of digital data. Thus, digital data can be stored in oligos. Since the stored data can be recovered through reading the data representing oligos with common DNA sequencers the development of a non-volatile data storage technology based on DNA molecules is possible.

Already in 2012 George Church and his team have proven the concept of storing data in DNA molecules [1]. In this successfully conducted experiment a simple code was applied for data protection, i.e.,

Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.398

bit 0 was represented by either A or C, while bit 1 was represented by either G or T, meaning that 1 bit was stored per nucleotide. However, in order to store huge amount of data in DNA reliably over long periods of time, a more advanced error protection scheme is necessary for both reliability and high efficiency. A first approach to store error protected data in DNA was carried out by Nick Goldman and his team [2].

We are reporting on a strong capacity boost in storing data in DNA molecules that is made possible by a data encoding scheme we dedicatedly developed to perfectly match to the specific characteristics of modern bio-chemical and bio-technical machines and processes, like DNA synthesizers and sequencers. With the error protection scheme we are presenting, masses of valuable data can be stored in DNA molecules.

Astonishingly for laymen in the field of bio-chemistry, DNA molecules are very robust. In February 2015 the ETH Zürich has reported that DNA molecules are a well suited long-term storage medium [3]. Together with our dedicated error correction scheme, the DNA data storage technology is a very promising candidate to build an urgently needed archiving solution up on it.

In section 2, the working principle of the DNA data storage is introduced. The DNA channel characteristics are considered in section 3. We explain our channel adapted modulation concept in section 4. The details of our error protection scheme are presented in section 5. In autumn 2015 we successfully conducted an experiment in which we stored 22 MB of a MPEG compressed movie sequence as a first step towards a scheduled 1 GB experiment. Details of this experiment are given in section 6. Finally, a conclusion and outlook can be found section 7.

2 The DNA Data Storage Technology in a Nutshell

Modern synthesizers can produce oligos limited to about 250 nucleotides with an acceptable error rate. The limitation in length is determined because the error rates increase drastically the longer the oligos get. Consequently, in DNA data storage the data to be stored has to be divided up into small Data Blocks (DBs). Each of the DBs will then be assigned to correspondingly synthesized DNA oligos. In order to be able to recover the stored DBs in the correct order the block addresses have to be stored together with the DBs in the same oligo. In modern DNA synthesizers not only one copy per oligo is synthesized at the same time but many of them. The distribution of the amount of the copies of the individual oligos is called oligo coverage. Next in the data storing process, during several consecutive applied Polymerase Chain Reaction (PCR) steps the oligo coverage will be doubled in each step, or in other words, the oligos are being amplified. The totality of all oligos are finally “pored” together and stored. When recovering the stored data, the oligos have to be sequenced and transferred into digital data.

3 DNA Channel Characteristics

Biological, bio-chemical and bio-physical processes are prone to errors, while physical and chemical effects by itself cause errors in oligos over the course of the times. We have based our designed Forward Error Correction (FEC) scheme on a solely phenomenological error analysis of the experimental data Church and his team gathered [1].

A swap error occurred in an oligo, if a nucleotide has been replaced with an incorrect one, while the oligo length stays unchanged. An insertion or deletion error occurred in an oligo, if an additional nucleotide has been inserted or removed, respectively. Depending on the predominant error type the affected

oligo is correspondingly lengthened or shortened. In the experimental data of [1] we found that the swap error rate lies between $\sim 6.0 \cdot 10^{-4}$ and $\sim 1.4 \cdot 10^{-3}$, while insertion and deletion error rates are $\sim 1.0 \cdot 10^{-3}$ and $\sim 5.0 \cdot 10^{-3}$, respectively.

However, not only of swap, insertion and deletion errors have to be taken care of, but also of general bio-technological malfunctions. The major aspect in this regard is that the coverages of the synthesized and afterwards amplified oligos vary. We found in the experimental data [1] an approximately bell shaped distribution density function of the oligo coverage, while of some of the oligos we could not find any copy in the sequencer read data at all. Oligos that cannot be found in the read data are called missing oligos. Finally we can say that we did not see any evidence that the DNA storage system is not a memoryless data channel.

4 Modulation

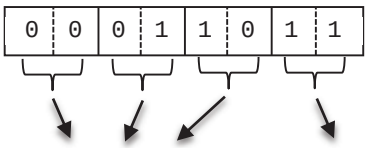
Modulation defines the mapping rule from information representing binary bits to nucleotide chains, while demodulation is the reverse process, i.e., to obtain the binary bits from sequenced nucleotide chains.

When developing the mapping of the binary information to the corresponding nucleotide sequence, the following four fundamental problems of the DNA channel had to be tackled.

- Error propagation: A single swap error, should alter bits after demodulation as little as possible. This allows efficient protection of the oligo addresses with a bit-correcting BCH code for a given redundancy.
- Run-length limitation: Because a sequence of more than 3 identical nucleotides may cause difficulties during sequencing of the oligos, the maximum run-length of identical nucleotides shall be limited to 3 [11], [12].
- Insertion and deletion errors: Since these two error types are quite common for today's synthesizers and sequencers, means for detecting and correcting insertion and deletion errors are mandatory.
- Self-reverse complementariness: Self-reverse complimentary sections within a DNA molecule, potentially causing amplification issues of the corresponding oligo, had to be avoided [2], [11], [12].

According to identified distortions due to the DNA channel, we have designed a modulation scheme to efficiently tackle these problems. Basically, our modulation maps 8 information bits to 5 nucleotides, where a 5-nucleotide tuple is referred to as a DNA symbol in the following. The mapping rule is illustrated in **Tables 1** and **2**. Specifically, bits 0 and 1 are mapped to the first nucleotide, bits 2 and 3 are mapped to the second nucleotide, and bit 4 and 5 are mapped to the fourth nucleotide of the DNA symbol according to the mapping rule in **Table 1**. That is, two-bit tuples 00, 01, 10, 11 are mapped to nucleotides A, C, G, T, respectively. Finally, bits 6 and 7 are mapped to the third and the fifth nucleotide of the DNA symbol according to **Table 2** that provides always 4 options for each pair of bits. All options that do not violate the above defined run-length limitation for identical nucleotides may be used. This design ensures that a single swap error affects not more than 2 bits during demodulation.

For example, the information byte 00011011 is mapped as follows: According to **Table 1**, the first two bits (00) are mapped to the first nucleotide A. Bits 2 and 3 (01) are mapped to the second nucleotide C and the bits 4 and 5 (10) are mapped to the fourth nucleotide G. According to **Table 2**, for the mapping of last two bits (11) to the third and the fifth nucleotides 4 options are possible: either AT, CA, GC, or TG. This leads to 4 different representations of the byte in DNA, namely, ACAGT, ACCGA, ACGGC, or ACTGG.



Value	Nucleotide
00	A
01	C
10	G
11	T

Table 1: Mapping of first 3 bit-tupel to 3 nucleotides

Value	Option 1	Option 2	Option 3	Option 4
00	AA	CC	GG	TT
01	AC	CG	GT	TA
10	AG	CT	GA	TC
11	AT	CA	GC	TG

Table 2: Mapping of last 2 bits to a pair of nucleotides

In order to limit the maximum run-length of identical nucleotides to 3 for any possible combinations of DNA symbols, we defined the following constraints resulting in valid DNA symbols:

- The first three nucleotides shall not be the same
- The two last nucleotides shall not be the same

Accordingly, for the above mentioned example, ACAGT, ACCGA, ACGGC are valid mappings, while ACTGG is invalid due to two same ending nucleotides, thus violating the second constraint.

With the above described constraints at least 2 valid DNA symbols can be found for every data byte. Furthermore, for 208 out of the possible 256 bytes exist three valid mappings, as shown in our example. That is, it is always possible to map binary data of an arbitrary length to DNA symbol sequences without violating the maximum defined run-length limitation. Because there are always at least two valid DNA symbols as representation for one byte available, it is possible to group the symbols into two complete clusters A and B and one incomplete cluster C. Therefore, the data can be mapped in DNA using symbols from cluster A and B in a specific manner, e.g. alternatingly. This kind of modulation supports the detection of insertion and deletion errors. Influences of insertion and deletion are shown by examples in **Figure 1**, where the alternating sequence of symbols from cluster A and B will be most likely interrupted at or at least near the position of such a kind of errors. In case of insertion (the upper half example in **Figure 1**), the original nucleotide sequence "ACAGT TCTGA CCGGA TGTAG" (the first and third symbols are from cluster A, and the second and fourth symbols from cluster B) is changed to "ACAGT **X**TCTG ACCGG ATGTA" with one inserted nucleotide denoted as "X". During demodulation, it is checked whether the sequenced nucleotide sequence matches to the alternating clusters A and B. For this example, it is observed that the third DNA symbol definitively does not match to the alternating scheme, indicating a potential insertion or deletion error. In the given example for a deletion, the third DNA symbol belongs to cluster C, which also violates the constraint that DNA symbols belong to the alternating clusters A or B.

Additionally, self-reverse complementary sections in the DNA strands can be eliminated. If such a section has been generated with DNA symbols from the alternating clusters A and B, symbols within the sequence can be replaced by their counterpart from a different cluster.

The theoretical limit of the storage capacity of the DNA molecules without any constraint such as run-length limitation or self-reverse complementariness is 2 bit per nucleotide. The proposed modulation reduces the storage capacity to 1.6 bit per nucleotide, but solves all of the above mentioned problems of the DNA channel.

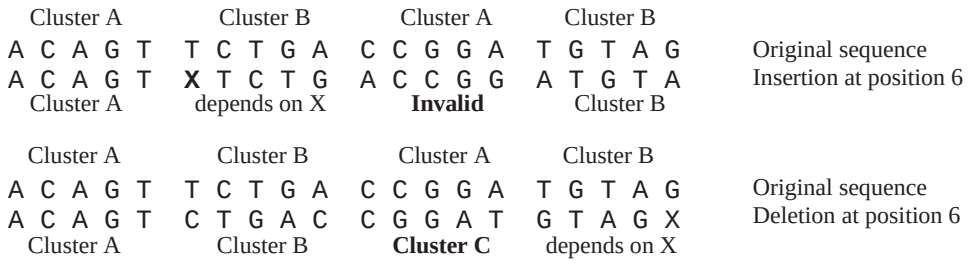


Figure 1: Insertion and deletion detection based on the use of alternating clusters for the DNA symbols

5 Forward Error Correction Scheme

We have drawn up a dedicatedly interleaved two-dimensional FEC scheme matching to the DNA channel. It is as effective as possible while it protects at the same time the stored data as strong as necessary. The FCC scheme takes especially care of the fact that in DNA storage the data to be stored is divided up into DBs. A block of data stored in consecutive oligos is protected in three succeeding ways:

- The addresses of the oligos are protected individually
- The data of consecutive oligo blocks are protected together
- To each oligo an Error Detection Code (EDC) is attached

The possibilities of errors in the oligos increase towards their ends. Consequently, we placed the address data at the beginning of the oligos, while the EDC data comes at their end. **Figure 2** illustrates the FEC scheme.

The interaction of the triple protection constituents supported by the modulation concept makes powerful decoding strategies possible. For instance, the detection and correction of insertion and deletion errors are enhanced based on the applied EDC.

5.1 Protection of Oligo Address

It is most crucial when recovering the stored data that the sequenced oligos can be aligned in the right order. Therefore, it is of particular importance to protect the address sections of the oligos strongly. According to our modulation design single nucleotide errors only lead to a maximum of two consecutive bit errors. Therefore, the Oligo Addresses (OAs) are effectively protected with a strong BCH code. In our experiment, a BCH (63, 39) code with a minimum Hamming distance of 9 was chosen to protect addresses, where 39 payload bits can be used for representing addresses [5]. The BCH code can detect at least 8 bit errors and correct up to 4 bit errors. Consequently, about 6.4 Tb data can be stored in DNA molecules, assuming that each oligo can store 150 bit of data.

5.2 Protection of Consecutive Oligo Blocks

It is important to effectively protect consecutive oligos in a manner so that completely missing oligos can be reconstructed efficiently. Reed Solomon (RS) block codes provide this functional feature very effectively as a maximum distance separable code [5], [6]. We used a RS code over a Galois field $GF(2^8)$ with a block size of 223 payload symbols to which 32 redundancy symbols are added, which in the common notation is written as RS(255, 223, 33) [5].

In terms of common forward error correction codes, missing oligos are erasure errors, which means that the positions where the errors occur are known. In random errors the positions are also unknown. If

S_P and S_E denote the number of erasures and random errors respectively, then the error correction capability of an RS(n, k, l) code is given by the following inequality (1) [5], [6].

$$n - k \geq S_P + 2 \times S_E \quad (1)$$

The effect of not completely corrected insertion and deletion errors is not limited to the position where a nucleotide was inserted or deleted which in fact means that essentially burst errors occur. Beneficially, the deployed RS code protects symbols of 8 bit, or in other words, is an efficient error protection to tackle any kind of burst errors.

5.3 Error Detection of Oligos

Modern oligo synthesizers from Agilent show a mean error rate of approximately $\frac{1}{300}$, while the occurrence probability obeys a Bernoulli distribution [11]. For instance, Agilent synthesizers are able to produce oligos of about 150 up to 250 nucleotides. Consequently, the probability is high that the produced oligos contain errors. However, the additional application of an effective EDC to the data stored in each oligo enables further possibilities to correct errors reliably, or even provide additional decoding possibilities, if completely unexpected failures occur.

For instance, the added EDC parity bits make it possible to compare two imperfect oligos and thus correct the errors while checking the resulting oligo against the EDC parity bits. Further on, when decoding the stored data, there is a small residual probability that the BCH (63, 39) error protection code of the OAs may fail. We therefore protected not only the payload bytes of an oligo, but also the OAs. In the unlikely case that the ECC of some OAs fails, the EDC parity bits prevent to sort oligos with wrong addresses. Also quite helpful is the fact that in interaction with our deployed modulation, the EDC is helpful to detect and correct insertion and deletion errors.

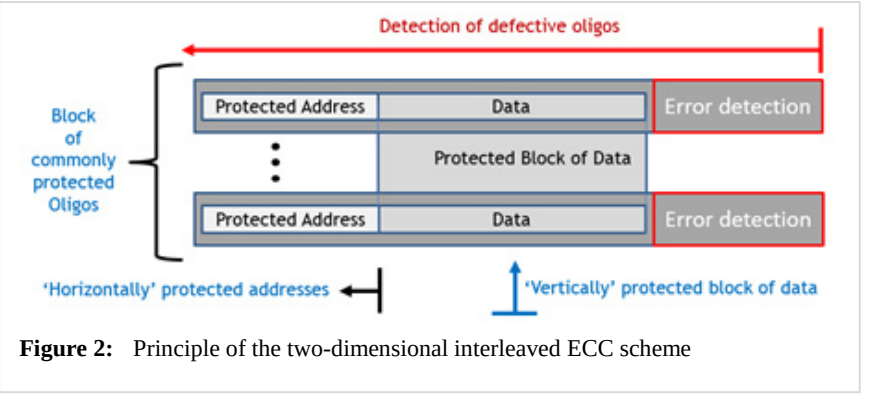


Figure 2: Principle of the two-dimensional interleaved ECC scheme

In this context we decided to use a CCITT 16 bit Cyclic Redundancy Check (CRC) as EDC for each oligo [7].

6 Successful 22 MB Experiment

6.1 Synthesizing, PCR Amplification and Sequencing

We synthesized 900 000 230 nt oligonucleotides on Agilent's Oligo Library Synthesis (OLS) microarray platform, divided into four libraries with 225 000 oligonucleotides each. DNA was eluted by Agilent to give a ~1 picomole pool of oligonucleotides in 100µL TE (10mM Tris-Cl pH 7.5, 0.1mM EDTA). Illumina-specific sequencing adaptors were introduced into the synthesized OLS pool in a two-stage serial PCR amplification using the SYBR Fast Master Mix (Kapa Biosystems). Reactions were performed using the following protocol on an Eppendorf Mastercycler Realplex4 Real-Time PCR machine by monitoring the SYBR Green channel signal:

1) 95˚C for 3 min, 2) 95˚C for 10 sec, 3) 60˚C for 30 sec, 4) Read Sybr Green Channel,
5) Goto Step 2 for a total of 9 cycles, 6) 68˚C for 30 sec, 7) Hold at 4˚C

Each reaction was harvested after 8 cycles of amplification (log phase) to avoid PCR bias in the resulting library. The resulting PCR products after each stage were purified using Agencourt AmpPureXP beads according to manufacturer's instructions.

We sequenced the amplified library by loading 1mL of 16pM library (16 fmoles; ~1:1000 of the amplified library) on 2 lanes of a Rapid Seq 300 cycle SBS kit on an Illumina HiSeq 2500 next-generation sequencer. We obtained 144,475,005 paired reads with 83.78% of the reads scoring >= Q30.

6.2 Primer Design

After data encoding into a payload sequence of 190nt was complete, the remaining 40 nt of each oligonucleotide were used to introduce a single pair of suitable flanking adaptors for PCR amplification post-synthesis. This function was carried out using a proprietary sequence identification algorithm that accounts for internal mispriming within the designed oligo library, as well as G/C content (melting temperature), homo- and hetero-dimerization, and internal hairpin formation that could reduce the efficiency of the resulting amplification.

6.3 Decoding

Oligos are sequenced in both forward and backward directions. Before decoding, merging of forwardly and backwardly sequenced oligo fragments is necessary. For this purpose, a specific algorithm has been developed that determines for each oligo the offset with the highest correlation between the forward and backward reads by first trying the expected offset and then varying the offset up to a defined

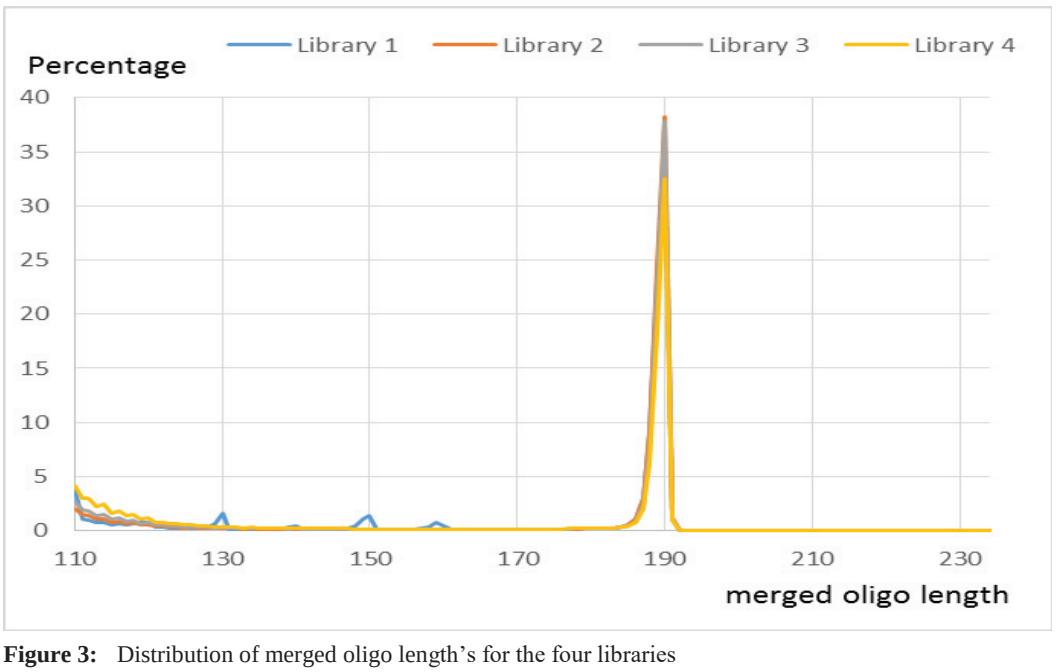


Figure 3: Distribution of merged oligo length's for the four libraries

limit. Then the merging is performed taking into account possibly different sequencer read quality values of single nucleotides: If the nucleotides are not identical, the nucleotide with the better quality value is chosen.

Figure 3 gives an overview of the oligo length distribution after merging. The percentage of merged oligos with the correct nucleotide length 190 lies between 32.5 % (library 4) and 38.3 % (library 2). The mean percentages of merged oligos with a length slightly different from 190 is included in **Table 3**.

As the length of the forward and reverse oligo fragments is 150 nt, the minimum length of the oligos after removing the primers of 20 nt length is 110 nt. The rising of the curves towards the minimum oligo length is caused by sequenced fragments, where no reasonable merging was possible. The peaks of the library 1 curve indicate an agglomeration of much too short thus defective oligos (see section 6.4).

The decoding was performed in four passes. In the first three passes all merged oligos were processed with different algorithms. When an oligo was found that fulfilled the criteria of the corresponding pass, it was stored in a key-value database while the oligo address served as key. During each pass it was controlled, whether already sufficient oligos were stored in the database in order to finalize the decoding. In the fourth pass missing oligos were recovered using a FEC code.

Oligo Length	Mean Percentage
187	2.6 %
188	8.0 %
189	22.0 %
191	1.1 %
192	0.02 %

Table 3: Mean percentage of oligos with a length slightly different from 190

Pass 1: In the first pass only oligos were selected, which had the correct length, the demodulation was performed without errors, the CRC checking succeeded, the BCH correction of the address didn't show any errors and the parity bit was correct.

Pass 2: In the second pass, remaining erroneous oligos were examined by means of majority voting and detection and correction of insertion and deletion errors.

For majority voting oligos with correct length but erroneous CRC were collected. All oligos with the same address were compared and a presumably correct oligo was reconstructed by selecting the nucleotides appearing most frequently.

If the length of an oligo differed from the correct length by one, it was assumed that an insertion or deletion error occurred. As described in section 4, the location of the defect was determined by examining the order of used modulation clusters. The decoder tried to repair the related code-word by deleting or inserting single nucleotides, respectively, until a valid code-word was generated, that leads to an oligo showing no errors.

	Library 1	Library 2	Library 3	Library 4
# synthesized oligos	225,000	225,000	225,000	225,000
# decoded oligos pass 1	213,452	224,998	224,998	224,574
# decoded oligos pass 2	75	0	1	180
# decoded oligos pass 2	3673	2	0	136
# decoded oligos pass 3	7217	0	1	95
# decoded oligos total (pass 1 -3)	224,417	225,000	225,000	224,985
Percentage decoded oligos total	99,74 %	100,0 %	100,0 %	99,99 %

Table 4: Overview of the number of decoded oligos in the single passes

Pass 3: In the third pass the decoder tried to reconstruct correct oligos from a plurality of oligos that were significantly to short, because a sequence of nucleotides anywhere in the middle of the oligos was missing. As the number of missing nucleotides and the position of this "hole" varied, it was possible to reconstruct correct oligos from these fragments. This was performed by gathering all oligos with an address, which could not be decoded successfully in passes 1 and 2. From these oligos the valid fragments at the beginning and the end of the oligos were determined by checking the correctness of the single code-words. Subsequently, the fragments were combined to a complete oligo by determining for each nucleotide position the nucleotide that occurs at this position in the fragments most often.

Table 4 gives an overview of the number of decoded oligos in the single passes.

Pass 4: After the passes 1 to 3 have been completed it might be possible that some oligos are finally missing. In addition, some incorrectly decoded oligos could be entered in the database also. Therefore, in pass 4 RS decoding is performed to recover the missing oligos and to correct wrongly entered oligos in the key-value database. In our experiment, a RS(255, 223, 33) was deployed, which can correct all possible combinations of erasure errors S_P and general errors S_E that solve Inequation (1).
Decoding results showed that the employed RS code was capable to successfully decode all RS blocks in all four libraries. The highest percentage of missing oligos occurred in library 1 (0.26 % and 583 oligos, respectively). Although the used RS code can reconstruct up to 12.5 % of missing oligos, due to the unequal distribution of missing oligos in library 1, a high error correction capability was actually required, as illustrated in **Figure 4** for the first 150 RS blocks. During RS decoding, each missing oligo was treated as an erasure.

For the remaining RS blocks of library 1, the number of missing oligos was always below 8 and the number of erroneous oligos below 3. In libraries 2 to 4 the maximum number of missing oligos per RS block was 1 and the maximum number of erroneous oligos was also 1. Thus, in library 1 the full correction capability of the RS code was used, while for the other libraries only a small part of the capability was necessary.

6.4 Systematic Oligo Distortions

In library 1 – especially at the beginning of this library – several incomplete oligos were detected. We found a correlation between the data stored in an oligo and the probability that no correct copy of this oligo was found in the sequenced read data. More specifically, if the data contain a long sequence of identical data bytes, the correspondingly formed oligos have nucleotide sequences with periodic repetitions of series of 10 nucleotides. The mentioned incompletely sequenced oligos contain only frag-

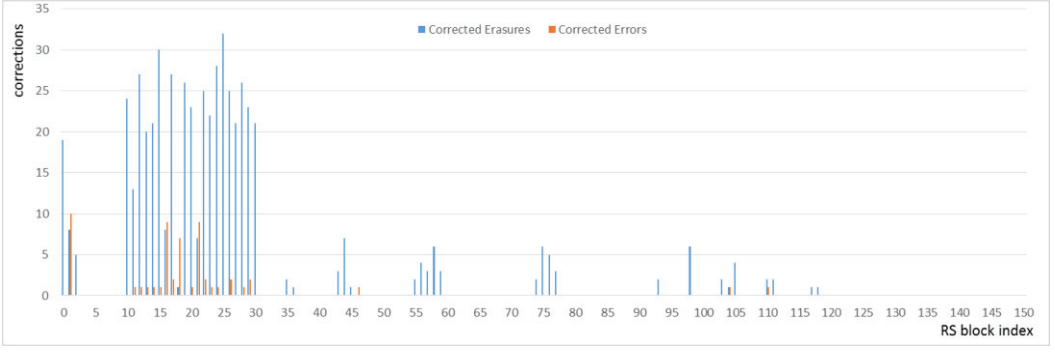


Figure 4: Number of corrected erasures (due to missing oligos) and corrected errors in the first 150 RS blocks of library 1

ments of the originally formed oligos, while nucleotide sequences with lengths of a multiple of 10 nucleotides are missing. This data dependent effect is visible in **Figure 3**. Only for library 1 the curve shows peaks at oligo lengths of 130, 140, 150 and 160 nucleotides.

6.5 Residual Error Probability

The practicability of a storage technology in general depends decisively on the total residual error probability. This statement holds especially for a long-term archiving solution. In order to show the suitability of the DNA data storage technology we approximately calculated the resulting residual error probability of our encoding scheme by estimating the probability of the determining residual error of the applied RS(255, 223, 33) code and compare the result with the corresponding values of common Hard Disk Drives (HDDs).

The residual error probability P_e for a RS(n,k,2s+1) code is [6], [7], [8]

$$P_e = P_{fr} + P_{fe},$$

where $[n, k, 2s + 1]$ denote code word length, payload word length, and the minimum distance of the Reed-Solomon (RS) code, respectively.

P_{fr} is the probability that decoding fails due to i random symbol errors with i>s, which can be evaluated as follows:

$$P_{fr} = 1 - \sum_{i=0}^{s} \binom{n}{i} \cdot e_{sy}^{i} \cdot (1 - e_{sy})^{(n-i)},$$

where e_{sy} is the symbol error rate.

P_{fe} is the probability that i≤s random errors can be corrected, but decoding fails due to more than 2(s-i) erasures. P_{fe} can be calculated as

$$P_{fe} = \sum_{i=0}^{s} \mathrm{P}_{sr}(i) \cdot P_{fe}(2(s - i)),$$

where $P_{sr}(i)$ is the probability that a codeword has i random symbol errors:

$$P_{sr}(i) = \binom{n}{i} \cdot e_{sy}^{i} \cdot (1 - e_{sy})^{(n-i)}$$

$P_{fe}(x)$ is the probability that there are more than x erasures in the received word and can be evaluated as

$$P_{fe}(x) = 1 - \sum_{i=0}^{x} \binom{n}{i} \cdot e_{er}^{i} \cdot (1 - e_{er})^{(n-i)},$$

Applying the preceding equations, we can show that already for the quite short code word length of the used RS (255,223,33) code over GF(2^8), the residual error probability is well below 10^{-15} for $e_{sy} = 10^{-3}$ and $e_{er} = 10^{-3}$. Applying a FEC code with longer code word length, e.g. a RS (4095,3583,513) code defined over GF(2^{16}), will improve the residual error probability further by several orders of magnitude, thus outperforming today's HDDs by fare. Typically, modern HDDs show read errors in the order of magnitude of 10^{-14} to 10^{-15} [13].

7 Conclusion and Outlook

Synthetic DNA is an excellent candidate to become the urgently needed long-term preservation medium for valuable digital data. DNA molecules, if stored dry, cold, and dark, are certainly robust enough to stay intact for at least several thousands of years, while no high data migration costs arise as with common mass technologies. Further strong arguments to use DNA molecules to archive digital data are the possible storage density in the order of 5 PB per gram, the simplicity to copy DNA molecules at any time and that there is in fact no format obsolescence risk.

As proven with our here described successful 22 Mbyte experiments, the problem of intrinsic errors of the synthesizer, PCR, sequencer, and storage chain can be solved efficiently with a robust FEC

scheme ensuring the integrity of the stored data. The achievable residual error is in any case certainly smaller than the residual error of traditional storage media.

In an already planned next development step we will increase the amount of stored data into the GB range. The presented FEC scheme is already designed for such and even much higher volumes of data. Nevertheless, incremental changes like a bigger RS block size will reduce the already achieved residual error further.

The remaining issue will be the cost for synthesizing and sequencing the DNA. Especially the synthesizing technology has to be considerably improved to make the archival of mass data in DNA cost-attractive. However, during the past years the development of bio-technology has outperformed Moore's Law by far [4]. We strongly believe, if this stormy progress goes on with the same pace, then an economical archiving solution based on DNA molecules could become possible in the coming 5 to 10 years.

8 Acknowledgements

We would like to thank Mark Crovella, Boston University (www.cs.bu.edu/~crovella/) for valuable discussions.

9 Reference

[1] Next-Generation Digital Information Storage in DNA
George M. Church, Yuan Gao, Sriram Kosuri
Science Vol. 337, 28 September 2012

[2] Towards practical, high-capacity, low-maintenance information storage in synthesized DNA
Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M. LeProust, Botond Sipos, Ewan Birney
Nature 11875, Nature, January 2013

[3] Data-storage for eternity
Angelika Jacobs
ETH Zürich, 13th of February 2015
https://www.ethz.ch/en/news-and-events/eth-news/news/2015/02/data-storage-for-eternity.html

[4] Regenesis: How Synthetic Biology Will Reinvent Nature and Ourselves
George M. Church, Ed Regis
Basic Books (1112), First Edition 2012

[5] Error Correction Coding: Mathematical Methods and Algorithms
Todd K. Moon
Wiley-Interscience, First Edition 2005

[6] Coding Theory
André Neubauer, Jürgen Freudenberger, Volker Kühn
John Wiley & Sons Ltd, 2007

[7] Error Control Coding
Lin Shu and Daniel J. Costello
Prentice Hall, Second Edition 2004

[8] Reed-Solomon Error Correction Codes (ECC) – Primer
AHA Application Note ANRS01_0404
AHA Products Group (AHA) of Comtech Telecommunications Corporation

[9] Next generation sequencing technology: Advances and applications
H.P.J. Buermans, J.T. den Dunnen
Leiden Genome Technology Center, Leiden University Medical Center, Postbus 9600, 2300 RC Leiden, The Netherlands
Biochimica et Biophysica Acta, Elsevier, 2014
[10] Next Generation Sequencing
Andy Vierstrade
Department of Biology, Ghent University, June 2012
[11] Aural Information by George Church
Harvard Medical School, Wyss Institute
http://arep.med.harvard.edu/gmc/
[12] Aural Information by Mark Crovella
Department of Computer Science, Boston University
http://www.cs.bu.edu/~crovella/
[13] Empirical Measurements of Disk Failure Rates and Error Rates
Jim Gray, Catharine van Ingen
Microsoft Research Technical Report MSR-TR-2005-166, December 2005