Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Multidimensional data organization and random access in large-scale DNA storage systems

Xin Song [a,b,c,*], Shalin Shah [a,c], John Reif [a,c,**]

[a] Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA
[b] Department of Biomedical Engineering, Duke University, Durham, NC 27708, USA
[c] Department of Computer Science, Duke University, Durham, NC 27708, USA

ARTICLE INFO

Article history:
Received 11 March 2021
Received in revised form 13 September 2021
Accepted 15 September 2021
Available online xxxx

Dataset link:
https://github.com/xinsong926/Multidimensional-DNA-Storage

Keywords:
DNA storage
Hierarchical memory
Data random access
Nested PCR
Amplification bias
PCR stochasticity

ABSTRACT

With impressive physical density and molecular-scale coding capacity, DNA is a promising substrate for building long-lasting data archival storage systems. To retrieve data from DNA storage, recent implementations typically rely on large libraries of meticulously designed orthogonal PCR primers, which fundamentally limit the capacity and scalability of practical DNA storage. This work combines nested and semi-nested PCR to enable multidimensional data organization and random access in large DNA storage. Our strategy effectively pushes the limit of DNA storage capacity and dramatically reduces the number of orthogonal primers needed for efficient PCR random access. Our design uses only $k * n$ primers to uniquely address n^k data-encoding oligos. The architecture inherently supports various well-defined PCR random-access patterns that can be tailored to organize and preserve the underlying DNA-encoded data structures and relations in simple database-like formats such as rows, columns, tables, and blocks of data entries. We design in silico PCR experiments of a four-dimensional DNA storage to illustrate the mechanisms of sixteen different random-access patterns each requiring no more than two PCR reactions to selectively amplify a target dataset of various sizes. To better approximate the physical system, we formulate mathematical models based on empirical distributions to analyze the effect of pipetting, PCR bias, and PCR stochasticity on the performance of multidimensional data queries from large DNA storage.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Advances in DNA synthesis and sequencing technologies have enabled the recent development of DNA-based data storage systems [1]. Typically, the digital data is segmented and encoded as a large pool of oligonucleotides (oligos) of fixed lengths that are appropriate for synthesis and sequencing. To achieve data retrieval and reconstruction, prior DNA archival storage systems leveraged PCR-based random access [2–5]. In these systems, each individual data-encoding oligo is appended with a short address sequence that acts as a unique primer target to selectively amply the oligo from the storage pool. For example, a prior work demonstrated individual file retrieval from a DNA storage encoding over 200 MB of data [2]. Oligos belonging to the same file are assigned with the same primer target and require an additional unique index for sequence alignment

* Corresponding author at: Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708, USA.
** Corresponding author.
E-mail addresses: xin.song@duke.edu (X. Song), reif@cs.duke.edu (J. Reif).

https://doi.org/10.1016/j.tcs.2021.09.021
0304-3975/© 2021 Elsevier B.V. All rights reserved.

Please cite this article as: X. Song, S. Shah and J. Reif, Multidimensional data organization and random access in large-scale DNA storage systems, Theoretical Computer Science, https://doi.org/10.1016/j.tcs.2021.09.021

and data reassembly. However, this type of addressing mechanism is unscalable since a vast number of orthogonal primers must be meticulously designed to ensure the specificity of PCR random access. Furthermore, it limits the way in which data can be organized and retrieved from a DNA storage. For example, retrieving a set of related files would require as many unique primers as needed for retrieving those individual files one by one. This is equivalent to carrying out large-scale multiplex PCR using numerous primers simultaneously, which can be costly, inefficient, and prone to spurious results due to the increased risk of mispriming and crosstalk.

In this work, we investigate the designs of multidimensional DNA storage systems and propose a scalable architecture that combines nested and semi-nested PCR – simple techniques known for improving selective amplifications of target oligos from large and potentially noisy backgrounds [6]. Our architecture features two important advantages: (1) it scales up the PCR-based random access by dramatically reducing the number of orthogonal primers needed, and (2) it supports multiple well-defined random-access patterns to efficiently organize and retrieve data from large DNA storage in simple database-like formats such as rows, columns, tables, and blocks. In theory, our architecture requires only $k * n$ orthogonal primers to organize and uniquely address n^k data-encoding oligos in large DNA storage, where k specifies the number of dimensions and n indicates the number of single data entries stored in each dimension. To better illustrate the mechanisms of multidimensional data queries, we present in silico PCR experiments of an n^4 DNA storage pool that supports arbitrary indexing and 16 different PCR random-access patterns using $4 * n$ primers. To estimate the performance of data queries from large-scale physical systems of DNA storage, we implement mathematical models to simulate the multidimensional PCR random access under the effect of pipetting, PCR bias, and PCR stochasticity.

2. Theory

Nested PCR is a well-established technique for improving the specificity and sensitivity of PCR reactions, where the amplicon from the first PCR reaction serves as the template for the second PCR amplification primed by an inner primer pair [6]. Nested PCR was previously studied in the context of building hierarchical DNA memories. Kashiwamura et al. [7] appended three address blocks on one end of the data-encoding oligos and used a common reverse primer target on the opposite end. Access to a single data oligo was achieved by specifying the sequential order of the address primers. The design concept was analogous to semi-nested PCR [8], where one of the outer primers used for the first PCR reaction is also used as an inner primer during the second PCR amplification. Semi-nested PCR was also used in a recent work by Tomek et al. [9] to improve the DNA storage capacity by combining a two-primer nested address system with physical separations for target file enrichment. Yamamoto et al. [10] expanded the nested primer hierarchy and concatenated multiple primer targets on both ends of the data oligos. Each primer pair was referred to as an address layer, and data retrieval operations proceed from priming the outer-most layer towards the inner layers. Their work experimentally demonstrated that by specifying a unique primer pair for each address layer, the target oligo could be amplified and extracted from an enormous address space after several nested PCR reactions. However, their primary focus was to scale up the nested primer hierarchy and to demonstrate the high specificity of single target oligo retrieval.

In contrast, our design goal is threefold: (1) providing a systematic architecture to efficiently organize the data-encoding oligos into a scalable and multidimensional address space, (2) minimizing the number of orthogonal primers needed for uniquely indexing oligos from a large storage pool, and (3) strategically combining the nested and semi-nested PCR (i.e., allowing the use of forward/reverse primers from the same or different address layers during data retrievals) to support multiple well-defined PCR random-access patterns in large DNA storage systems. Our architecture ensures that the amplicons from each PCR reaction always represent useful database-like relations in the form of rows, columns, tables, and blocks of data entries with respect to the overall data storage pool. This gives rise to multiple efficient data organization and retrieval patterns to assist the rational designs of DNA storage based on the underlying structures and relations of the data content. Moreover, our design strategy can be easily adjusted to allow different configurations of multidimensional data storage. In Subsection 3.1 we present two design variations. The first design supports 16 data random-access patterns (operating on the block, table, row, column, and single-entry levels), each requiring just one or two PCR reactions. The second variation eliminates the block-level organization to further simplify several particularly useful random-access patterns into a single PCR reaction. Next, in Subsection 3.2 we present in silico PCR experiments of a simple four-dimensional DNA storage to illustrate the mechanisms of sixteen different random-access patterns using pre-designed sequences of primers and data oligos. Then in Subsection 3.3 we present mathematical models to simulate and analyze the effect of PCR bias as well as the effect of PCR stochasticity on the performance of multidimensional data queries from large DNA storage. Finally, Section 4 gives a discussion and conclusions; Subsection 4.1 gives an estimate of theoretical storage capacity and physical density; Subsection 4.2 describes how to achieve efficient random access with high resolution and low error rate; Subsection 4.3 discusses potential applications and the future outlook.

3. Results and analysis

3.1. Designs of multidimensional DNA storage with multiple random-access patterns

In design A (Fig. 1), a multidimensional DNA storage is organized into blocks, tables, rows, and columns of data-encoding oligos by use of a hierarchical addressing mechanism. For a four-dimensional storage system, each oligo comprises several domains including a data payload block surrounded by three address blocks on both sides that function as the PCR

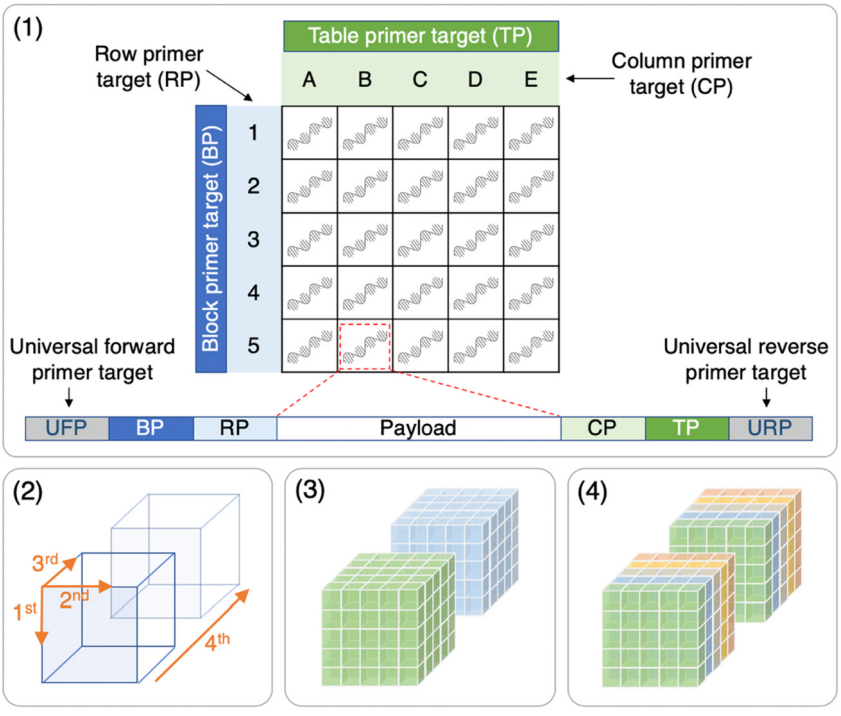


Fig. 1. Schematic illustration of a multidimensional DNA storage according to design A. (1) Design of data-encoding oligos and their organization as addressable entries in a 2-dimensional table. (2) Illustration of a four-dimensional addressable space. The 1st to 4th dimensions refers to rows, columns, tables, and blocks of entries, respectively. (3) Data blocks are distinguishable by the block primer target BP. (4) Data tables within each block are distinguishable by the table primer target TP. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

Table 1
Total number of orthogonal primers and uniquely addressable data entries in design A.

Number of orthogonal PCR primers needed (**4∗n**) = (# of rows in a table) + (# of columns in a table) + (# of tables in a block) + (# of blocks) + 2
Number of unique data entries addressable ($\mathbf{n^4}$) = (# of rows in a table) × (# of columns in a table) × (# of tables in a block) × (# of blocks)

primer binding sites (i.e., primer targets). These address blocks are termed and arranged according to their specific roles in forming the data organization levels and the random-access patterns. A four-dimensional DNA storage can contain multiple three-dimensional data blocks, and each block can contain multiple two-dimensional data tables each consisting of one-dimensional rows and columns of single data entries. Such a hierarchical architecture allows highly flexible and efficient random access by maximal reuse of only a small set of orthogonal primers. Specifically, blocks are distinguished by different block primer targets (BP) but reuse the same set of table primer targets (TP) within each block. In other words, tables within a block are distinguished by different TP but share the same BP. All single data entries in a table reuse the same BP/TP pair but are distinguished by different pairs of row primer target (RP) and column primer target (CP). Therefore, the same set of RP/CP pairs can be repeatedly reused by all tables in all blocks. Further, all oligos in the storage pool share the same pair of universal forward primer target (UFP) and universal reverse primer target (URP) as the outermost address layer. On one hand, this helps to support simultaneous data retrieval from all blocks by a single PCR reaction with UFP/URP. On the other hand, this facilitates numerous additional well-defined hierarchical data retrieval patterns (e.g., by pairing one of the inner address blocks with UFP or URP).

In the schematic illustrations, for simplicity, we depict DNA in single-stranded form without showing the complementary strand. Hence, the actual sequences of the reverse address blocks (CP, TP, URP) on the oligos are reverse complementary to the matching primer sequences. The total number of orthogonal primers needed to index arbitrary data entries in such a four-dimensional address space is calculated in Table 1. Approximately, this architecture uses $4 * n$ orthogonal primers to uniquely index n^4 data entries. Fig. 2 illustrates the 16 different data retrieval patterns enabled by this architecture. Depending on the retrieval pattern, the target data subset can be quickly enriched by just one or two PCR reactions. Specifically, if the data query needs to target two different addresses on the same side of payload, the random-access pattern would require two nested PCR reactions. If the query targets only one address on either side of the payload, the random-access pattern can be achieved by just one PCR reaction. For data queries that target the inner address layers, the resulting amplicons will lose their original outer address layers on the oligos after PCR. This is an inherent property of the hierarchical

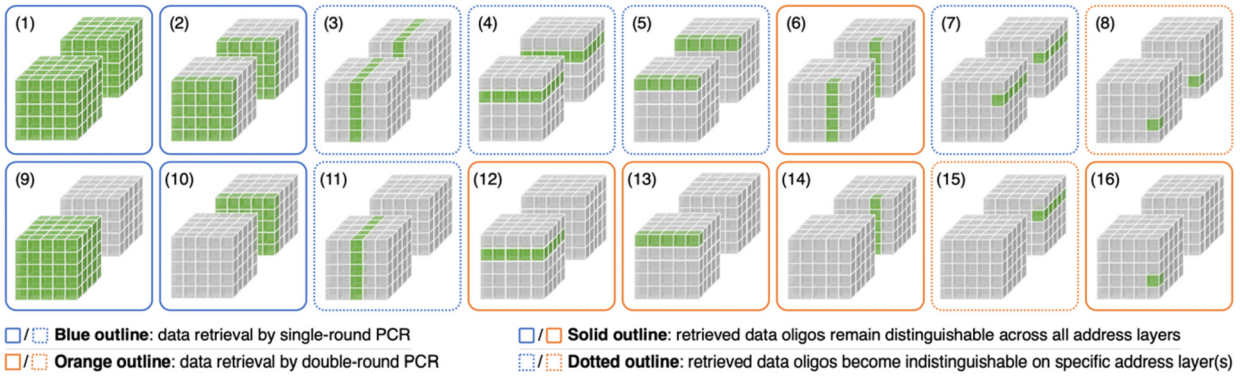


Fig. 2. Schematic illustration of data retrieval patterns supported by design A. The notation "ForwardPrimerTarget/ReversePrimerTarget" represents PCR reaction that operates on the indicated primer target pair. (1) UFP/URP: entire data storage (all blocks). (2) UFP/TP: specific table from all blocks. (3) UFP/CP: specific column from all tables in all blocks. (4) RP/URP: specific row from all tables in all blocks. (5) RP/TP: specific row from specific table in all blocks. (6) UFP/TP then UFP/CP: specific column from specific table in all blocks. (7) RP/CP: specific entry from all tables in all blocks. (8) UFP/TP then RP/CP: specific entry from specific table in all blocks. (9) BP/URP: specific block. (10) BP/TP: specific table from specific block. (11) BP/CP: specific column from all tables in specific block. (12) BP/URP then RP/URP: specific row from all tables in specific block. (13) BP/TP then RP/TP: specific row from specific table in specific block. (14) BP/TP then BP/CP: specific column from specific table in specific block. (15) BP/URP then RP/CP: specific entry from all tables in specific block. (16) BP/TP then RP/CP: specific entry from specific table in specific block. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

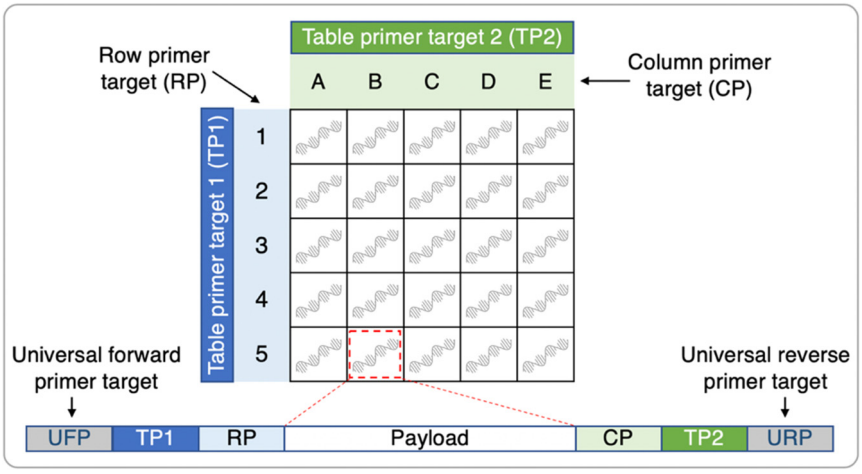


Fig. 3. Schematic illustration of a multidimensional DNA storage with the design of data-encoding oligos according to design B. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

data queries that needs to be considered when tailoring the random-access patterns to specific data storage contents. Notably, when target oligos from multiple addresses are simultaneously retrieved, several random-access patterns can maintain the oligos' mutual distinguishability on all address layers. In other words, the full addresses of the amplified oligos can be inferred from their remaining inner-layer address blocks after sequencing. This allows high-resolution data assembly and reconstruction down to single data entries. An example is pattern (14) of Fig. 2, where the exact indexing on BP, TP, and CP address layers can be completely inferred from the primers used during random access, and thus the retrieved oligos can be individually distinguished from each other according to the remaining RP address on the resulting amplicons. For data retrieval patterns that cannot circumvent the loss of certain address layer(s) after PCR random access, one may still take advantage of those patterns by associating them with storage contents that do not require high-resolution data reassembly. Pattern (15) of Fig. 2 is such an example where indistinguishability occurs on the TP address layer because the random access operates on an inner address layer (CP) without first amplifying the oligos using an outer address (TP) on the same side of payload.

To illustrate the flexibility of our multidimensional data storage design, we show that the design A can be slightly modified for DNA storage systems that do not benefit from the block-level organization but require rapid random access to arbitrary data rows/columns in arbitrary data tables by a single PCR reaction. In design B (Fig. 3), each entry in the table corresponds to a data-encoding oligo composed of several domains. Tables can be distinguished by either TP1 or TP2. All entries in a given table share the same TP1/TP2 pair. RP identifies the row, and CP identifies the column. The same set of RP/CP pairs are reused by all tables. The same UFP/URP pair is used as the outermost address layer on all data oligos to enable simultaneous retrieval of all tables or a specific row/column from all tables through the RP/URP or UFP/CP pair. The

Table 2
Total number of orthogonal primers and uniquely addressable data entries in design B.

Number of orthogonal PCR primers needed (**3∗n**) = (# of rows in a table) + (# of columns in a table) + (# of tables × 2) + 2
Number of unique data entries addressable ($\mathbf{n^3}$) = (# of rows in a table) × (# of columns in a table) × (# of tables)

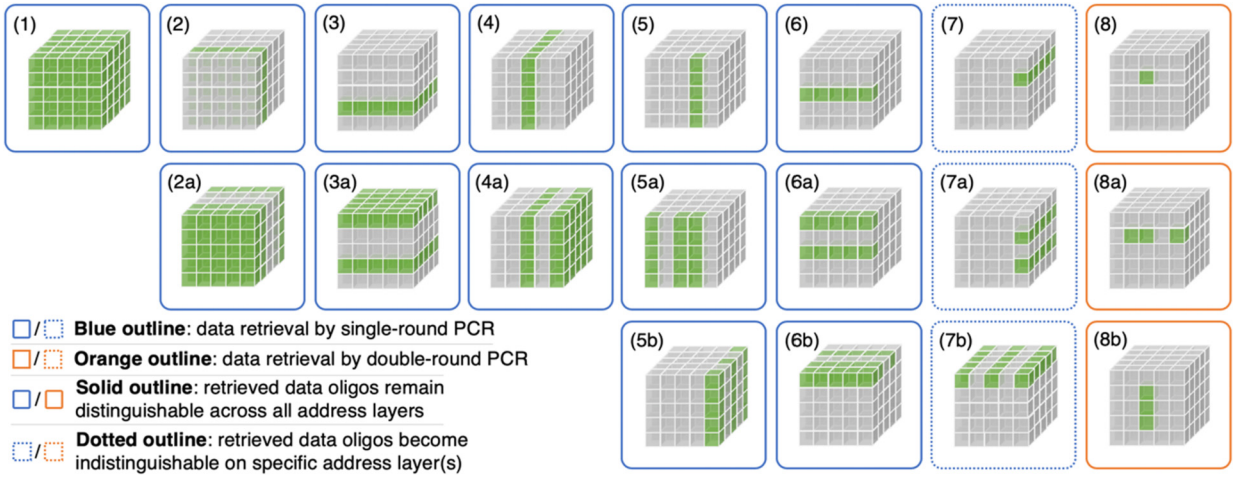


Fig. 4. Schematic illustration of data retrieval patterns supported by design B. The notation "ForwardPrimerTarget/ReversePrimerTarget" indicates PCR reaction that operates on the specified primer target pair. (1) UFP/URP: entire data storage. (2) TP1/TP2 (or TP1/URP or UFP/TP2): specific table. (2a) TP1(multiplex)/URP: multiple tables. (3) RP/URP: specific row from all tables. (3a) RP(multiplex)/URP: multiple rows from all tables. (4) UFP/CP: specific column from all tables. (4a) UFP/CP(multiplex): multiple columns from all tables. (5) TP1/CP: specific column from specific table. (5a) TP1/CP(multiplex): multiple columns from specific table. (5b) TP1(multiplex)/CP: specific column from multiple tables. (6) RP/TP2: specific row from specific table. (6a) RP(multiplex)/TP2: multiple rows from specific table. (6b) RP/TP2(multiplex): specific row from multiple tables. (7) RP/CP: specific entry from all tables. (7a) RP(multiplex)/CP: entry on specific column from multiple rows from all tables. (7b) RP/CP(multiplex): entry on specific row from multiple columns from all tables. (8) TP1/TP2 then RP/CP: specific entry from specific table. (8a) TP1/TP2 then RP/CP(multiplex): multiple entries on specific row from specific table. (8b) TP1/TP2 then RP(multiple)/CP: multiple entries on specific column from specific table. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

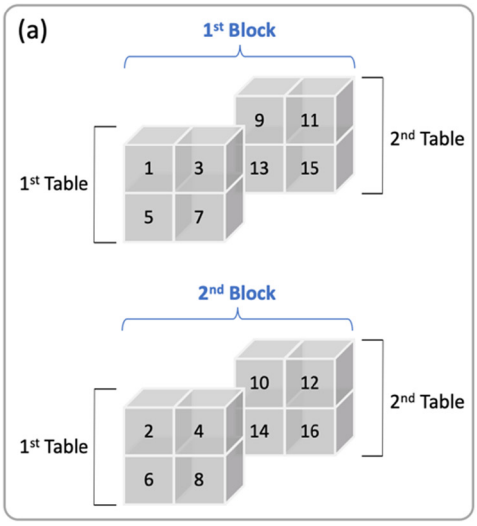
Table 3
List of orthogonal primers used in simulations of sixteen PCR random-access patterns.

Primer target	Primer sequence (5' -> 3')
UFP	AAGGCAAGTTGTTACCAGCA
URP	TGCGACCGTAATCAAACCAA
BP	AGCCGACAAGTTCAAACACA (BP1: 1st Block), GTTCAAATTGCGTGCGACAT (BP2: 2nd Block)
TP	ATTCGCGTCGCCTAATTTGT (TP1: 1st Table), AAACTGGAGGCGGCAAATTA (TP2: 2nd Table)
RP	TGGCTCATTTCACAATCGGT (RP1: 1st Row), ATAAATGACCTGCCGTGCAA (RP2: 2nd Row)
CP	TTCGTTCGTCGTTGATTGGT (CP1: 1st Column), AAACGGAGCCATGAGTTTGT (CP2: 2nd Column)

total number of orthogonal primers needed to index arbitrary entries in this three-dimensional address space is calculated in Table 2. Approximately, this architecture uses $3 * n$ orthogonal primers to uniquely index n^3 data entries. Fig. 4 illustrates 8 different data retrieval patterns inherently supported by this architecture along with 11 extended retrieval patterns (2a to 8a, 5b to 8b) by incorporating standard multiplex PCR. Except for random access of single data entries, all the other data retrieval patterns can be accomplished by just a single PCR reaction.

3.2. In silico PCR experiments of sixteen different random-access patterns

To better illustrate the mechanisms of multidimensional random access, we sought to conduct in silico PCR experiments to simulate all sixteen data retrieval patterns described in Fig. 2. However, most commercial PCR software are not designed to accommodate the large number of primers and target templates needed for practical (i.e., large scale) DNA storage simulations. The closest available software we found was FastPCR [11,12], which can accept multiple primers and template sequences as input to run multi-template PCR simulations that help to visually showcase the mechanism of our multidimensional data queries using pre-designed sequences of primers and data oligos. As shown in Fig. 5, we constructed a four-dimensional DNA storage consisting of two $2 \times 2 \times 2$ data blocks based on orthogonal primer sequences (Table 3) adapted from Organick et al. [2] The complete set of sequences of primers and data oligos used for simulating the hypothetical DNA storage is summarized in SI_primers_and_sequences.xlsx (Supporting Information). To simulate PCR random



(b)

Data entry #	Entry address (block, table, row, column)	Payload length (nt)	Total length (nt)
1	(1,1,1,1)	5	125
3	(1,1,1,2)	15	135
5	(1,1,2,1)	25	145
7	(1,1,2,2)	35	155
9	(1,2,1,1)	45	165
11	(1,2,1,2)	55	175
13	(1,2,2,1)	65	185
15	(1,2,2,2)	75	195
2	(2,1,1,1)	10	130
4	(2,1,1,2)	20	140
6	(2,1,2,1)	30	150
8	(2,1,2,2)	40	160
10	(2,2,1,1)	50	170
12	(2,2,1,2)	60	180
14	(2,2,2,1)	70	190
16	(2,2,2,2)	80	200

Fig. 5. Design of a hypothetical DNA storage for in silico PCR simulations of sixteen different random-access patterns. (a) Hierarchical organization of data entries in a four-dimensional address space. (b) Data-encoding oligos stored at different addresses are distinguished by different payload lengths in the simulations.

access that requires two (nested) PCR reactions, the amplicons from the 1st PCR reaction served as the input templates for the 2nd PCR reaction, and the resulting amplicons from the 2nd PCR reaction were compared against the target data entries for result verification. Details of the target data entries, input primer list, and amplicons predicted by FastPCR simulations are documented in SI_In silico PCR experiments.pdf (Supporting Information) for each random-access pattern described in Fig. 2. Despite its small size, this hypothetical storage contains enough addressable entries for us to demonstrate each of the sixteen random-access patterns enabled by the proposed multidimensional storage. In these in silico PCR experiments, oligos stored at different addresses in the hypothetical storage were assigned with data payloads of different lengths (represented by poly-A segments) to distinguish them from each other. Because the FastPCR software predicts a list of target amplicons (nucleotide sequences and lengths) after each round of simulated PCR reaction, the use of different payload lengths in data oligos allowed us to quickly distinguish the target data oligos from non-targets and verify the results of each PCR random-access pattern by simply inspecting the lengths of PCR amplicons. Furthermore, these small in-silico PCR examples serve as a proxy for how simple proof-of-concept PCR experiments can be designed and implemented in vitro to test out all 16 data-access patterns and interpret the data retrieval results simply by gel electrophoresis without sequencing. Because of the high specificity of nested PCR, spurious amplicons would have a much lower relative abundance than the target amplicons after hierarchical data queries. Therefore, practical implementations of multidimensional DNA storage can also distinguish the desired targets from background noises based on the relative abundance information obtained from sequencing the amplicon pool without explicit size-based filtering by gel electrophoresis.

It should be noted that the main limitation of these FastPCR simulations is that they do not account for the effects of PCR bias and stochasticity, which are important aspects of practical DNA storage systems. Because of this, we primarily used the FastPCR simulations as illustrative tools to showcase the mechanisms of the sixteen multidimensional random-access patterns rather than estimating the real-world performance of the proposed DNA storage architecture. For this latter task, we proceeded to construct mathematical models based on empirical distributions reported from prior experimental DNA storage systems to estimate the performance of multidimensional data queries from large DNA storage under the effect of pipetting, PCR bias, and PCR stochasticity.

3.3. Mathematical modeling of multidimensional PCR random access

Effect of PCR bias: In the ideal scenario of perfect PCR, every oligo in the storage pool is replicated with the same likelihood in each cycle of PCR, and we assume that the amplifications of oligos are stochastically independent events. Such a perfect PCR amplification can be modeled by an exponential function $n(j) = n_0 * 2^j$, where $n(j)$ is the total copy number of the oligo after the j^{th} PCR cycle, n_0 is the initial copy number of the oligo, and 2 is the maximal PCR efficiency corresponding to an amplification probability of 1. In mixed-template PCR reactions, different target templates may be amplified with different efficiencies due to factors such as sequence composition, secondary structures, GC content of the primer binding sites, etc. To study the effect of PCR bias on data random access, we adapt empirical distributions from prior work and model the PCR efficiency as a Gaussian random variable ($mean = 1.85,\ sd = 0.07$) which is template specific and remains constant between PCR cycles [13].

Effect of PCR stochasticity: In practice, PCR amplifications are imperfect, and each oligo may undergo replication with a probability less than one. Again, we assume that the amplifications of oligos are stochastically independent

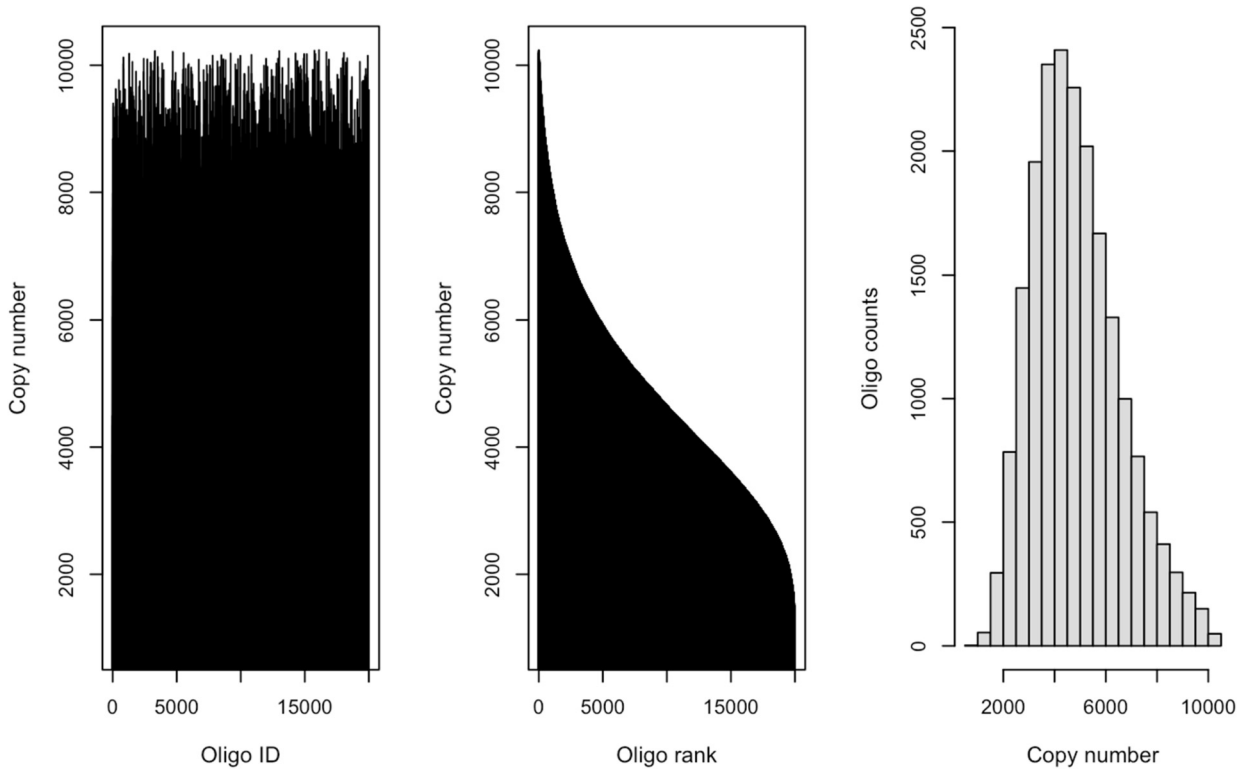


Fig. 6. Model simulation of PCR random access with the effect of sequence-specific PCR bias. Amplification target is entire pool of subset S (n = 20000, r = 10). Results plotted after 10 PCR cycles. Mean copy number: 4869 copies/oligo. Threshold: 487 copies/oligo. False negatives: 0 in 20000 (0%).

events. The behavior of PCR amplification with stochasticity can thus be described by the recurrence equation $n(j+1) = n(j) + B(n(j),\ P_{amp})$, where B is a binomially distributed random variable, $n(j)$ is the total copy number of the target oligo after the j^{th} PCR cycle, and P_{amp} is the oligo's amplification probability. To model the effect of PCR stochasticity without sequence-specific bias, we assume that all target oligos are amplified with the same probability during all cycles (e.g., $P_{target} = 0.85$), and all non-target oligos are amplified with the same spurious amplification probability during all cycles (e.g., $P_{spurious} = 0.15$). This is equivalent to modeling the PCR amplification as a Galton-Watson stochastic branching process [14–16].

Example of a DNA storage pool for PCR random access: Suppose we construct a large four-dimensional DNA storage system H consisting of $n = a*b*c*d$ uniquely addressable data oligos, where $a,\ b,\ c,\ d$ are the sizes of the four dimensions. We first synthesize the storage pool H with a physical redundancy of m and assume that the oligos are well mixed in the pool. Now we take a random subset of the storage pool by pipetting. This process can be modeled as repeated random sampling of H. The resulting subset of oligos can be represented by a multiset $|S| = r*n$, where each element of S is chosen independently and possibly redundantly from H, and r is a small number representing the relatively low physical redundancy of oligos in the resulting subset (i.e., $r < m$). Since H is of size mn and S is of size rn, the likelihood that any given oligo of the storage pool H is not in the multiset S is $\left(1-\frac{m}{mn}\right)^{rn} = \left(\left(1-\frac{1}{n}\right)^{n}\right)^{r} \approx e^{-r}$ for large n, where e is the Euler's number. Therefore, the likelihood that any given oligo of H occurs at least once in S is $1-e^{-r}$, and the fraction of complete loss of oligos (e.g., due to pipetting and other sources of fluid loss) can be estimated by $1-\left(1-e^{-r}\right) = e^{-r}$. Recall that r is the approximate physical redundancy of oligos in the subset S. In order to keep the oligo loss (dropouts) below 1%, a redundancy of at least $r = -\ln(0.01) \approx 4.61$ is needed. Practical DNA storage systems can incorporate data logical redundancy to tolerate potential oligo dropouts. For example, the Reed-Solomon encoding scheme used in a recent DNA storage implementation can tolerate up to $\frac{R}{100+R}\%$ oligo dropouts, where R is the percent logical redundancy [2]. Accordingly, in theory, a hypothetical 1% oligo loss due to pipetting can be mitigated by adding a small 2% logical redundancy (i.e., if $R = 2$, tolerance is $\frac{2}{100+2} > 1\%$).

To simplify the implementation and analysis of our PCR random-access models, we assume that the oligos in the subset S have a uniform low physical redundancy, and an appropriate level of logical redundancy is in place to tolerate the small fraction of oligo dropouts due to pipetting. To analyze the performance of our multidimensional PCR random-access patterns, we first simulated the amplification of the entire pool of subset S ($n = 20000,\ r = 10$) under the effect of PCR bias, stochasticity, and bias/stochasticity combined, respectively (Figs. 6–8). To calculate the data retrieval error rates, we set the copy number threshold equal to 10% of the average copy number of all oligos after amplification. Specifically, we count a

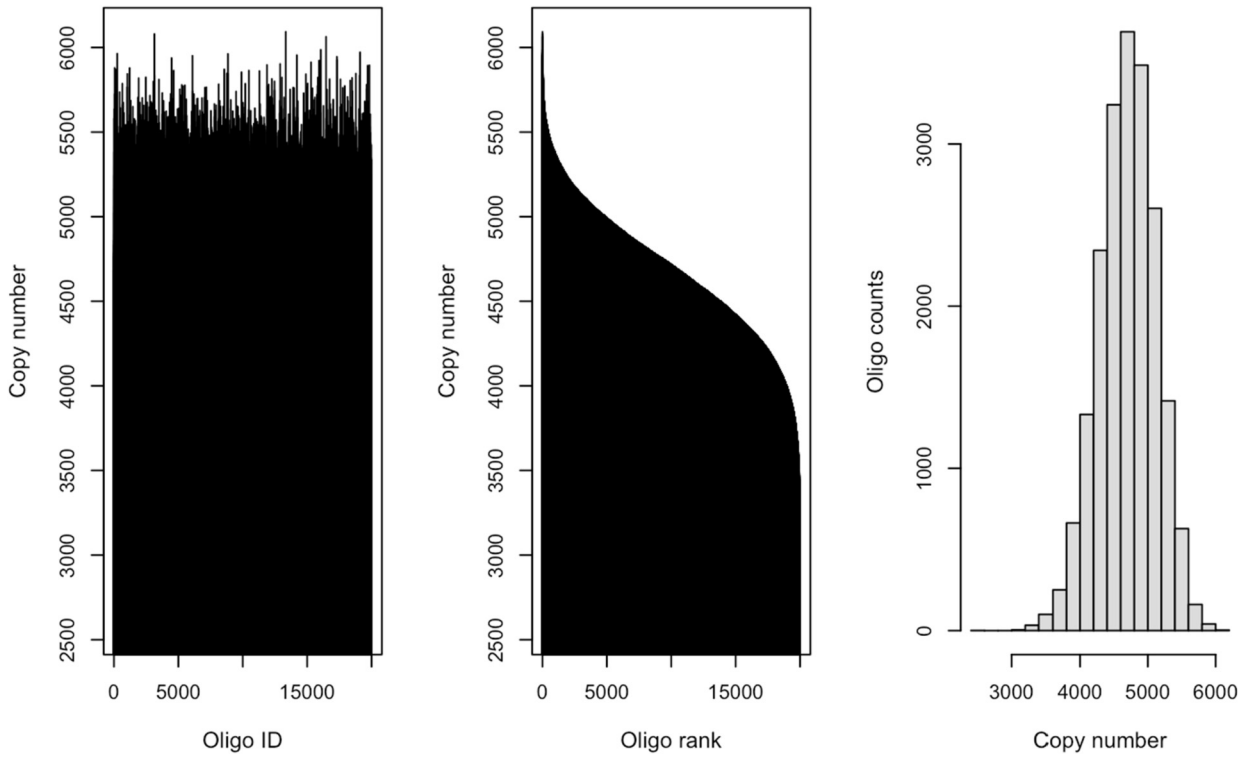


Fig. 7. Model simulation of PCR random access with the effect of PCR stochasticity. Amplification target is entire pool of subset S (n = 20000, r = 10). Results plotted after 10 PCR cycles. Mean copy number: 4699 copies/oligo. Threshold: 470 copies/oligo. False negatives: 0 in 20000 (0%).

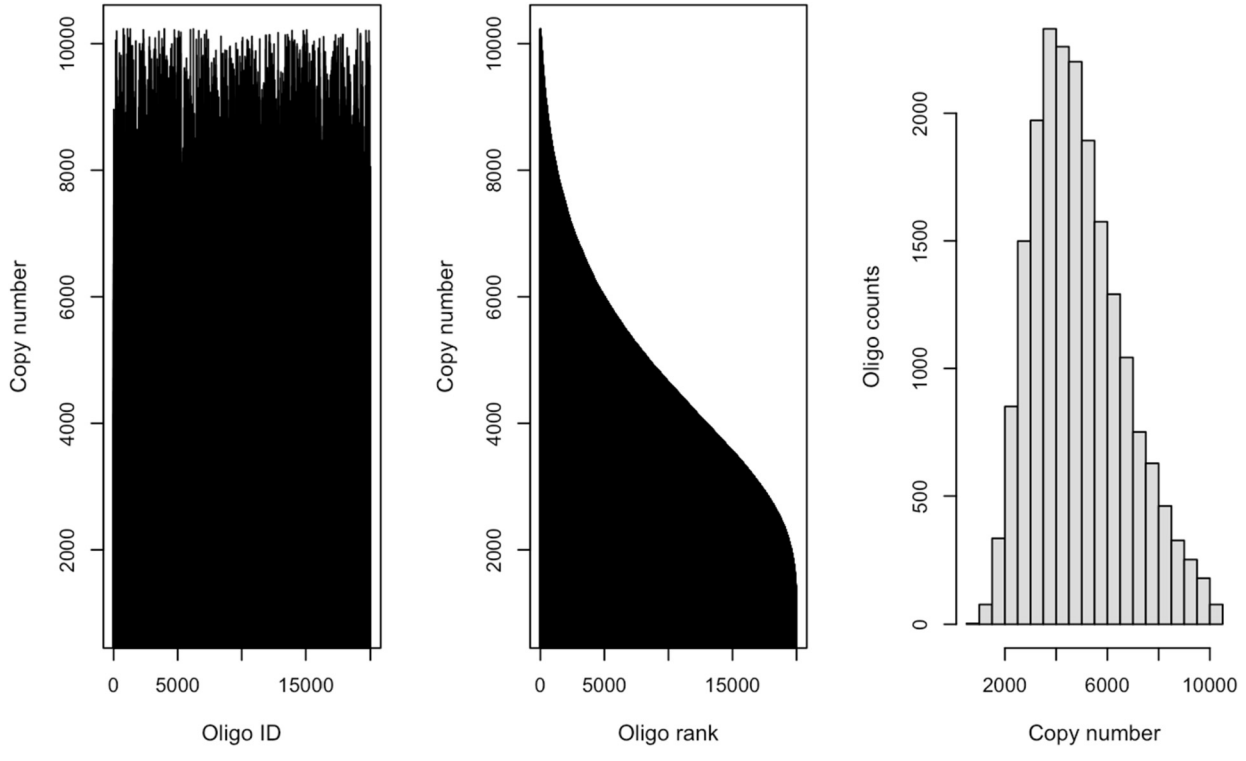


Fig. 8. Model simulation of PCR random access with combined effect of PCR bias and stochasticity. Amplification target is entire pool of subset S (n = 20000, r = 10). Results plotted after 10 PCR cycles. Mean copy number: 4892 copies/oligo. Threshold: 489 copies/oligo. False negatives: 0 in 20000 (0%).

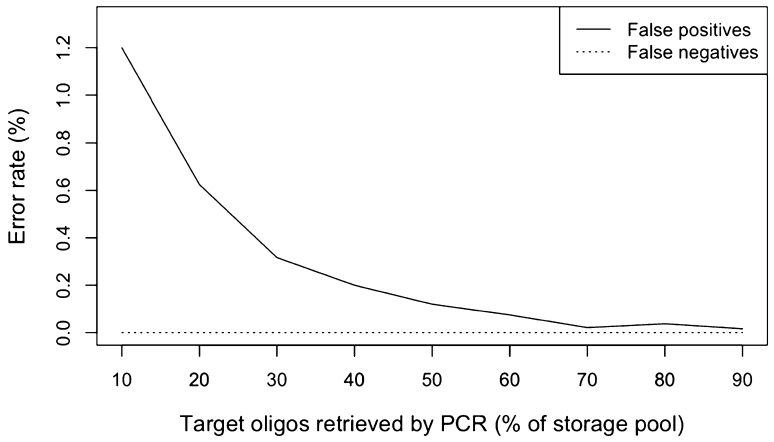


Fig. 9. Model simulation of PCR random-access patterns that require only a single PCR reaction to retrieve a data subset of varying sizes from the multi-dimensional storage pool. Results are plotted after 6 PCR cycles with an initial physical redundancy of 10 copies/oligo. Error rates are calculated based on the copy number threshold set to 10% of the average copy number of all oligos after amplification.

target oligo as a false negative if its copy number after PCR random access is less than the threshold. Similarly, we count a non-target oligo as a false positive if its copy number after PCR random access (due to spurious amplifications) is higher than the threshold. For modeling purposes, the value of the threshold can be adjusted or fitted to experimental data to qualitatively assess the influence of different model parameters on the trend of false positive/negative rates of PCR random access.

It is worth noting that some studies have shown that PCR bias is not a major source of error that skews sequence representation after the oligo pool amplification [13,14]. Therefore, we next focused on the model of PCR with stochasticity to analyze the performance of different PCR random-access patterns. Specifically, we varied the amount of target oligos from 10% to 90% (at 10% intervals) of the storage pool to simulate random-access patterns that require only a single PCR reaction to retrieve a data subset of various sizes. As shown in Fig. 9, the error rates remained very low in spite of the amount of target oligos retrieved. The performance of PCR random access depends on various parameters (e.g., PCR cycles, PCR efficiency, etc.) that can be adjusted in our mathematical models and R scripts (SI_Mathematical models.Rmd in Supporting Information) to better approximate physical systems. For example, Fig. 10 suggests that the errors can be significantly reduced by slightly increasing the number of PCR cycles or increasing the initial copy number of oligos in the storage pool. Fig. 11 shows that an improvement in PCR efficiency, which can be achieved by optimizing primer designs and other experimental conditions of PCR, can also help to reduce error rates in data queries from large DNA storage systems.

To model multidimensional data queries involving nested PCR reactions (e.g., the double-round patterns described in Fig. 2 and Fig. 4), we simulated nested PCR random-access in two stages. Specifically, the 1^{st} PCR reaction generates a pool of oligos containing both the desired amplicons (target oligos and their amplification products) and spurious amplicons (non-target oligos and their amplification products) predicted by our PCR stochasticity model. Then this entire oligo pool resulted from the 1^{st} PCR reaction is inputted back to the PCR stochasticity model to simulate the 2^{nd} (nested) PCR reaction. This modeling strategy does not assume any size-based filtering between nested PCR reactions and allows non-target oligos to be amplified with a small but nonzero probability. In fact, sometimes spurious amplicons from the 1^{st} PCR reaction may have the same size as the target amplicons and thus cannot be removed by size-based filtering method. For example, suppose we want to retrieve oligos from the data table 1 as the target, we may sometimes get spurious amplicons from table 2, 3, 4, etc. Because the 2^{nd} PCR reaction operates on an inner address layer, both the target and spurious oligos will be indistinguishably amplified by the 2^{nd} PCR reaction as long as they contain the same inner primer binding sites targeted by the 2^{nd} PCR. To account for these, the 2^{nd} (nested) PCR reaction of our model checks for spurious amplicons from three different sources: (1) desired amplicons from the 1^{st} PCR that are not targets of the 2^{nd} PCR, (2) spurious amplicons from the 1^{st} PCR that contain the inner primer binding sites targeted by the 2^{nd} PCR, and (3) spurious amplicons from the 1^{st} PCR without the inner primer binding sites targeted by the 2^{nd} PCR. All these three types of spurious amplicons are estimated by our PCR stochasticity model and combined to calculate the overall false positive rates of the nested PCR random access.

To assess the performance of nested PCR random-access reactions, we first varied the amount of target oligos from 10% to 90% (at 10% intervals) of the initial storage pool as input templates for the 1^{st} PCR reaction. We then took the entire pool of amplicons from the 1^{st} PCR reaction as input templates for the 2^{nd} PCR reaction and varied the amount of target oligos from 10% to 90% (at 10% intervals) of the input. As shown in Fig. 12, nested PCR random access reactions were simulated for a total of 12 PCR cycles with an initial physical redundancy of 10 copies/oligo. The error rates remained very low in spite of the amount of target oligos retrieved. Additional simulations (not shown) suggested that the errors can be significantly

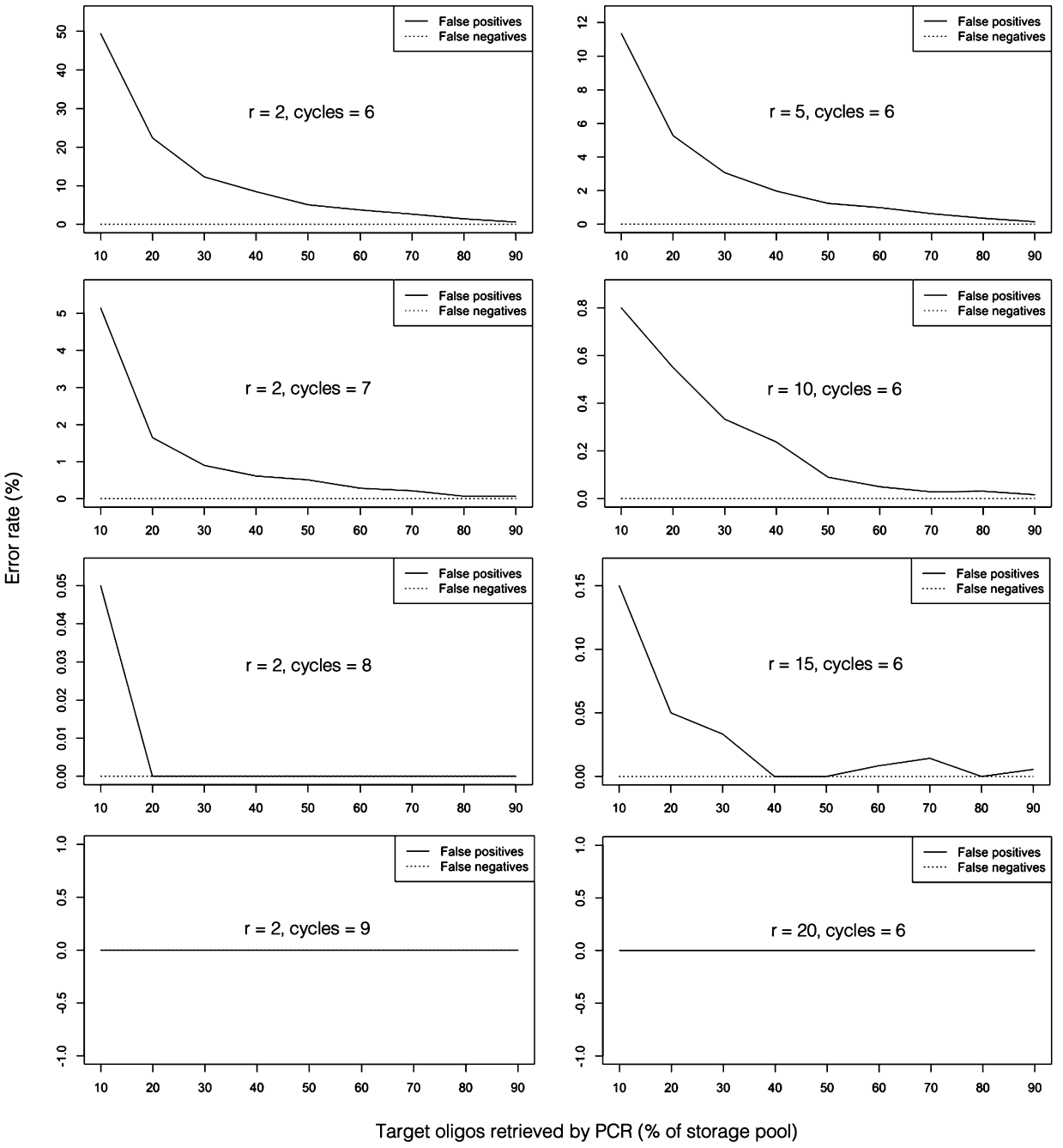


Fig. 10. Effect of PCR cycles and initial physical redundancy of oligos on the error rates of PCR random access. In each plot, 'r' indicates the average initial copy number of oligos in the storage pool, and 'cycles' indicates the total number of PCR cycles of a single PCR random-access reaction.

reduced by slightly increasing the number of PCR cycles or increasing the initial copy number of oligos in the storage pool, similar to our observations from single-reaction PCR random access models.

4. Discussion and conclusions

4.1. Theoretical storage capacity and physical density

A major advantage of our hierarchical, multidimensional DNA data storage architecture is that it significantly reduces the number of orthogonal primers needed to be meticulously designed to implement large DNA storage systems. A recent algorithm [2] was proposed to design up to 14000 pairs of orthogonal 20-mer primers to support PCR random access in

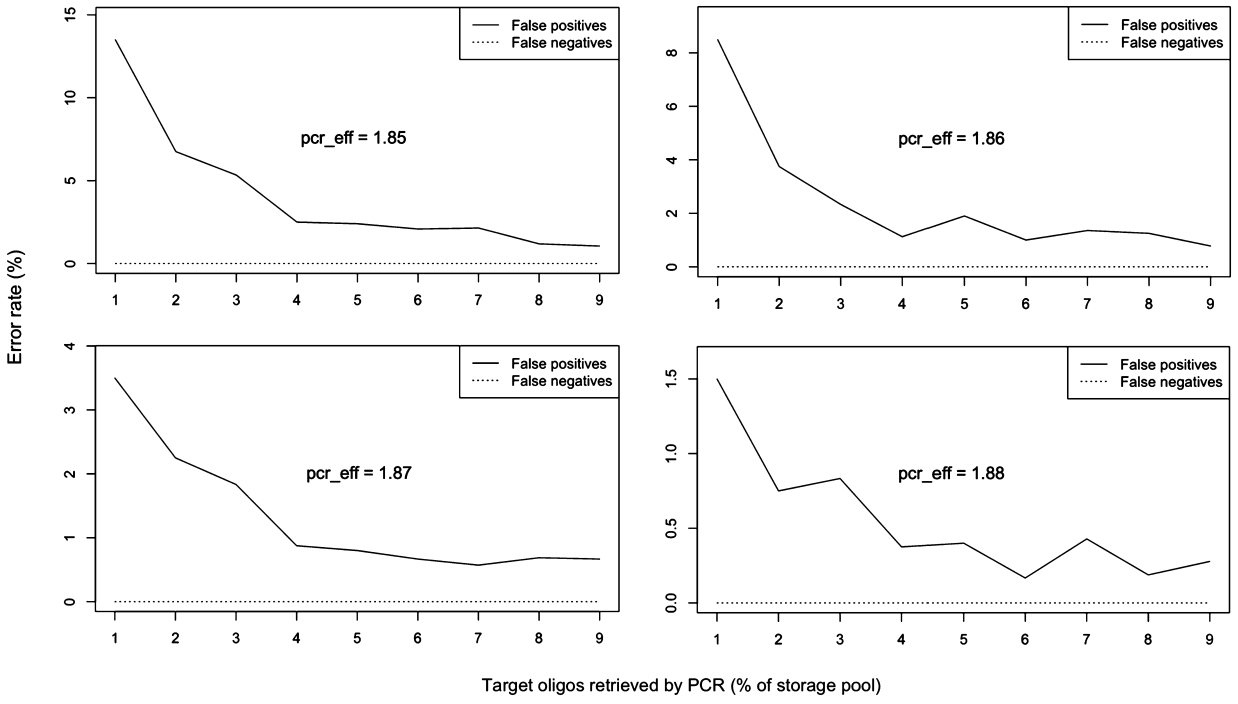


Fig. 11. Effect of PCR efficiency on the error rates of PCR random access. In this simulation, the amount of target oligos varied from 1% to 9% (at 1% intervals) of the storage pool. In each plot, "pcr_eff" represents the PCR amplification efficiency for target oligos. Non-target oligos are amplified with a default spurious amplification efficiency of 1.15. Results are plotted after 6 PCR cycles with an initial physical redundancy of 10 copies/oligo. Error rates are calculated based on the copy number threshold set to 10% of the average copy number of all oligos after amplification.

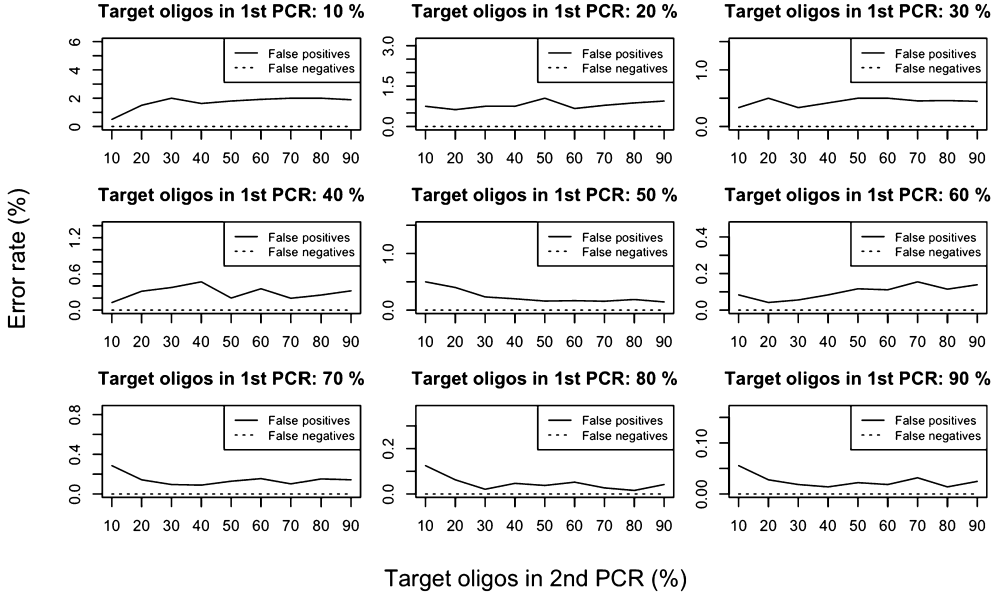


Fig. 12. Model simulation of PCR random-access patterns that require two (nested) PCR reactions to retrieve varying subsets of data from the multidimensional storage pool. Results are plotted after a total of 12 PCR cycles (i.e., 6 cycles in each PCR reaction) with an initial physical redundancy of 10 copies/oligo. Error rates are calculated based on the copy number threshold set to 10% of the average copy number of all oligos after amplification.

DNA storage of potential size up to a few terabytes. In contrast, a primer library of this size can be used much more efficiently by our storage architecture to dramatically increase the DNA storage capacity. Here we illustrate with an example based on the design from Fig. 1. Given a hypothetical library of 14000 orthogonal primer pairs, suppose we assign two primers for UFP and URP and then evenly distribute the remaining primers for use with BP, TP, RP, and CP. This would allow the simple construction of a very large DNA storage containing roughly 7000 blocks, each consisting of 7000 tables with

7000 rows and 7000 columns in each table ($n = 7000$, $k = 4$). Such an architecture could allow 7000^4 ($\approx$ 2401 trillion) data-encoding oligos to be hierarchically organized and uniquely addressed. If we assume that the length of each oligo is 200 nucleotide (nt) and the address blocks are 20-mers, every oligo in the DNA storage can carry a data payload of 80 nt ($200 - 20 \times 6$). While in theory each nucleobase may encode up to 2 bits, we assume a coding density of 1 bit/nt to account for coding redundancies needed for error correction. Accordingly, each oligo can encode 80 bits, which amounts to 24 petabytes (1 petabyte = 1000 terabytes) of data in the entire storage pool (Equation (1)). Based on the approximate molecular weight of 200 bp double-stranded oligos [17], the theoretical physical density of this DNA storage (assuming 100x synthesis redundancy) can reach $\sim$ 495 petabytes/gram (Equation (2)), which is comparable to estimations from prior work [18,19]. As DNA synthesis and sequencing technologies continue to improve and allow data encoding on longer synthetic oligos with higher coding densities, the storage capacity may further improve. In fact, several recent DNA storage systems have experimentally achieved higher than 1 bit/nt coding densities [1]. It is worth noting that the size of a single storage pool is ultimately limited by oligo interactions and diffusion kinetics, and therefore, practical implementations of large DNA storage may benefit from a shared multidimensional address space across multiple physically isolated storage pools [5].

$$\frac{7000^4 \text{ oligos}}{\text{storage}} \times \frac{80 \text{ nt}}{\text{oligo}} \times \frac{1 \text{ bit}}{\text{nt}} \approx \frac{24 \text{ petabytes}}{\text{storage}} \tag{1}$$

$$\frac{80 \text{ bits}}{\text{oligo}} \times \frac{1 \text{ oligo}}{100 \text{ strands}} \times \frac{6.022 \times 10^{23} \text{strands}}{\text{mol}} \times \frac{\text{mol}}{121638 \text{ gram}} \approx \frac{495 \text{ petabytes}}{\text{gram}} \tag{2}$$

4.2. Efficient random access with high resolution and low error rates

Our combined use of nested and semi-nested PCR lays the foundation for constructing large multidimensional DNA data storage while enabling numerous well-defined data retrieval patterns. Such a hierarchical storage can be leveraged to organize and preserve the underlying data structures and relations of DNA-encoded data. Notably, the architecture enables high-resolution data random access with remarkable efficiency and scalability. In the example scenario discussed above, every 80-bit data segment from the 24-petabyte multidimensional DNA storage is uniquely addressable and rapidly retrievable by just two PCR reactions based on the random-access pattern (16) illustrated in Fig. 2. The high specificity of nested PCR helps to reduce false positives in PCR random access. As discussed in our formulation of PCR models, several studies [13,14] reported that the impact of PCR bias on oligo distribution was not significant compared to the effect of PCR stochasticity. However, another study pointed out considerable variability between amplifications of different templates in complex pools [20]. Practically speaking, it is challenging to design large libraries of orthogonal primers without any differences in the primers' binding efficiencies. To mitigate potential impact of PCR bias on multidimensional random access, we would suggest ranking the orthogonal primers by their binding efficiencies such that primers with higher binding efficiencies are prioritized for use on outer address layers, whereas primers with lower binding efficiencies are mainly used on inner address layers. This simple strategy may help further improve the sensitivity (i.e., reducing false negatives) of data random access that require nested PCR amplifications.

4.3. Applications and outlook

The primary goal of this work is to use a small set of orthogonal primers to hierarchically structure large DNA data storage into a scalable and multidimensional storage space that inherently supports various efficient data retrieval schemes by simple PCR. In the example scenario discussed earlier, each data-encoding oligo serves as an 80-bit "data segment" in the storage, and a group of related data segments forming a "file" can be stored as a data table in the four-dimensional storage. Since a block can contain multiple tables, the block level can be used to categorize data into different "folders". It is even possible to assign multiple distinct data segments (distinguished by an additional index on the oligo) under the same address. However, with a fixed number of available primers, this would create a tradeoff between the random-access resolution and the storage capacity. To efficiently utilize the enormous address space established by our architecture, the dimensions of data tables and blocks should be carefully chosen according to the characteristics of the particular data items being stored. One caveat of having well-defined multidimensional data retrieval patterns is the potential waste of storage space allocated to unused indexable addresses. To mitigate this weakness, it may be useful to leverage data allocation techniques from computer file system designs. While we have only demonstrated designs for three- and four-dimensional data organization schemes, our proposed architecture for DNA storage could be easily scaled up with additional dimensions by adding new primer targets on the oligos. Although the combination of nested and semi-nested PCR allows very efficient reuse of primers on the inner address layers (i.e., oligos organized in different 5^{th}-dimension groups can use the same set of primers for their 4^{th}-dimension indexing, and so on), concatenating additional primer targets on the oligos would shorten their effective payload length and diminish the overall coding density of the DNA storage pool. Furthermore, random access to single addressable entries would likely require more than two PCR reactions when the storage is extended beyond four dimensions. This may negatively affect the data retrieval specificity and sensitivity for practical implementations.

On smaller scales, our designs may also find useful applications such as DNA-based lookup tables (e.g., by encoding numerical weights or Boolean values of useful functions) to potentially interface with DNA computing [1]. Alternatively, it may

be possible to design DNA reaction networks to compute a set of predicates, which subsequently activate the corresponding random-access primers (e.g., via strand displacement) to amplify a target set of data-encoding oligos from the DNA storage. Other variations of random access such as diagonal retrieval from a three-dimensional address space may be implemented by assigning special-purpose address targets on designated oligos. Another potential application is to hierarchically link several DNA databases together by encoding entries in one database as "pointers", which establish a relational network with the oligos in the second DNA database, and so forth. For example, oligos retrieved from one database may react in vitro to form a new set of primers to index another DNA database. The resulting amplicons may contain information that encodes specific operands or operators for subsequent in vitro computing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supporting information

Supporting information is available online at https://github.com/xinsong926/Multidimensional-DNA-Storage. This GitHub repository contains the following: (1) R Markdown file with the scripts for implementation and simulation of PCR random-access models (SI_Mathematical models.Rmd). (2) In silico PCR experiments illustrating sixteen different data random-access patterns (SI_In silico PCR experiments.pdf). (3) Sequences of primers and data oligos used for the in silico PCR experiments (SI_primers_and_sequences.xlsx).

Acknowledgements

This work was sponsored by NSF grant no. CCF 1617791, CCF 1813805, and CCF 1909848. X.S. acknowledges support from NSF grant no. DGE 1545220. We would like to thank Dr. Joshua Granek and Xiangyu Zhang for helpful discussions.

References

[1] X. Song, J. Reif, Nucleic acid databases and molecular-scale computing, ACS Nano 13 (2019) 6256–6268.
[2] L. Organick, S.D. Ang, Y.-J. Chen, R. Lopez, S. Yekhanin, K. Makarychev, M.Z. Racz, G. Kamath, P. Gopalan, B. Nguyen, et al., Random access in large-scale DNA data storage, Nat. Biotechnol. 36 (2018) 242–248.
[3] S.M.H.T. Yazdi, R. Gabrys, O. Milenkovic, Portable and error-free DNA-based data storage, Sci. Rep. 7 (2017) 5011.
[4] J. Bornholt, R. Lopez, D.M. Carmean, L. Ceze, G. Seelig, K. Strauss, A DNA-based archival storage system, Comput. Archit. News 44 (2016) 637–649.
[5] S. Newman, A.P. Stephenson, M. Willsey, B.H. Nguyen, C.N. Takahashi, K. Strauss, L. Ceze, High density DNA data storage library via dehydration with digital microfluidic retrieval, Nat. Commun. 10 (2019) 1706.
[6] M.R. Green, J. Sambrook, Nested polymerase chain reaction (PCR), Cold Spring Harb. Protoc. 2019 (2019) 175–178.
[7] S. Kashiwamura, M. Yamamoto, A. Kameda, T. Shiba, A. Ohuchi, Hierarchical DNA memory based on nested PCR, in: M. Hagiya, A. Ohuchi (Eds.), DNA Computing, DNA 2002, in: Lecture Notes in Computer Science, vol. 2568, Springer, Berlin, Heidelberg, 2003, pp. 112–123.
[8] X.Y. Zhang, M. Ehrlich, Detection and quantitation of low numbers of chromosomes containing bcl-2 oncogene translocations using semi-nested PCR, BioTechniques 16 (1994) 502–507.
[9] K.J. Tomek, K. Volkel, A. Simpson, A.G. Hass, E.W. Indermaur, J.M. Tuck, A.J. Keung, Driving the scalability of DNA-based information storage systems, ACS Synth. Biol. 8 (2019) 1241–1248.
[10] M. Yamamoto, S. Kashiwamura, A. Ohuchi, M. Furukawa, Large-scale DNA memory based on the nested PCR, Nat. Comput. 7 (2008) 335–346.
[11] R. Kalendar, B. Khassenov, Y. Ramankulov, O. Samuilova, K.I. Ivanov, FastPCR: an in silico tool for fast primer and probe design and advanced sequence analysis, Genomics 109 (2017) 312–319.
[12] R. Kalendar, D. Lee, A.H. Schulman, Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis, Genomics 98 (2011) 137–144.
[13] R. Heckel, G. Mikutis, R.N. Grass, A characterization of the DNA data storage channel, Sci. Rep. (9) (2019) 9663.
[14] J.M. Kebschull, A.M. Zador, Sources of PCR-induced distortions in high-throughput sequencing data sets, Nucleic Acids Res. 43 (2015) gkv717.
[15] B. Schierwater, D. Metzler, K. Krüger, B. Streit, The effects of nested primer binding sites on the reproducibility of PCR: mathematical modeling and computer simulation studies, J. Comput. Biol. 3 (1996) 235–251.
[16] Y.J. Chen, C.N. Takahashi, L. Organick, C. Bee, S.D. Ang, P. Weiss, B. Peck, G. Seelig, L. Ceze, K. Strauss, Quantifying molecular bias in DNA data storage, Nat. Commun. 11 (2020) 1–9.
[17] ThermoFisher Scientific, Nucleic acid molecular weight conversions, https://www.thermofisher.com/us/en/home/references/ambion-tech-support/rna-tools-and-calculators/dna-and-rna-molecular-weights-and-conversions.html. (Accessed 5 September 2019).
[18] G.M. Church, Y. Gao, S. Kosuri, Next-generation digital information storage in DNA, Science 337 (2012) 1628.
[19] Y. Erlich, D. Zielinski, DNA fountain enables a robust and efficient storage architecture, Science 355 (2017) 950–954.
[20] M.F. Polz, C.M. Cavanaugh, Bias in template-to-product ratios in multitemplate PCR, Appl. Environ. Microbiol. 64 (1998) 3724–3730.