Primer design for DNA storage random access

Jérémy Mateos[1,3], Dominique Lavenier[2], Melpomeni Dimopoulou[3], Anthony Genot[4], Marc Antonini[1]
[1] I3S Laboratory, Côte d'Azur University, CNRS, UMR 7271, Sophia Antipolis, France
[2] University of Rennes, Inria, CNRS, IRISA, Campus de Beaulieu, Rennes, France
[3] Pearcode, Sophia-Antipolis, France
[4] LIMMS (IRL2820)/CNRS-IIS, University of Tokyo, Tokyo, Japan
Email: mateos@i3s.unice.fr, dominique.lavenier@irisa.fr, melpomeni@pearcode.io, genot@iis.u-tokyo.ac.jp, am@i3s.unice.fr

Abstract

DNA is a promising candidate for data storage due to its high density and long-term stability. However, accessing specific data, known as random access, is challenging. This process uses primers, short DNA segments that act as identifiers. Efficient random access depends on high-quality primers, constrained by DNA structure. This paper introduces a method to generate primers that meet strict biochemical criteria, avoiding sequences that form problematic shapes. The proposed tool uses computational models to predict primer binding affinity and specificity, allowing users to adjust parameters for lab protocols, enhancing data retrieval efficiency and optimization.

Keywords

DNA data storage, PCR-based random access

1 Introduction

Unlike traditional storage media which face problems of longevity, integrity, ecology and energy consumption, DNA offers an incredible data density and stability over long periods, lasting thousands of years if stored under optimal conditions.

The DNA data storage workflow consists of six steps:

1. **Encoding:** Digital data is converted into sequences of A, C, G, T, cut into short chunks called oligonucleotides, and formatted with addressing fields,
2. **DNA synthesis:** Synthetic oligonucleotides are created based on the encoded information
3. **Storage:**DNA molecules are stored in controlled environments to ensure stability and longevity,
4. **Data retrieval:** Specific segments of stored DNA data are accessed randomly, mixing multiple DNA molecules in the same container,
5. **Sequencing:** DNA sequencers read and translate DNA molecules into A, C, G, T sequences,
6. **Decoding:** Retrieved sequences are reorganized and corrected to recover the original data.

This paper examines the efficiency of the data recovery phase in DNA-based storage according to the quality of the addressing primers (small adressing DNA sequence). Millions of files are stored in DNA molecules within the same space. Random access involves selecting DNA molecules corresponding to a file through PCR (Polymerase Chain Reaction) [1] [2]. PCR allows selective duplication of these molecules using primers that identify the sequences associated to one file. This technique, adapted for DNA data storage, ensures specific file retrieval, similar to accessing digital archives.

PCR-based random access [3] is an innovative modification of the classical PCR, commonly used in molecular biology, but requires precise primer design. Primers must attach only to short segments at the extremities of the oligonucleotides and should not form undesired shapes or loops to avoid non specific retrieval or data loss. To our knowledge, no software exists specifically for designing primers for DNA data storage, as existing tools are tailored for genomics and unsuitable for this purpose [4]. High-quality primer design is crucial for accurate file extraction, given the complexity of managing millions of sequences in the same location.

2 Molecular random access

As introduced previously, DNA data storage involves mixing millions of DNA oligonucleotides in a single container. Files are represented by DNA oligonucleotides with specific primer pairs at their extremities that barcode the files. Those files are present with multiple copies of each oligonucleotides. Accessing files uses a "PCR-based random access" method to amplify and retrieve specific DNA oligos among the others.

Unlike traditional PCR, DNA data storage primers are designed specifically for files and not derived from existing sequences. Each file has a unique molecular address provided by a primer, which must bind, called hybridization, efficiently to prevent incorrect amplification leading to data loss.

The PCR is a mix containing DNA, primers, DNA polymerase, cations (Na+ and Mg2+), and nucleotides. It involves three main steps, forming a cycle, that are repeated multiple times to replicate the DNA:

1. **Denaturation:** The reaction mix is heated to around 94-98°C to initiate replication.
2. **Annealing:** The temperature is lowered to allow the primers to bind to their complementary sequence on the oligo. This temperature is based on the **melting temperature (Tm)** depending on the primers being used, which will be explained in the following paragraph.
3. **Extension:** The temperature is raised to 72°C, the temperature to activate the DNA polymerase. DNA

polymerase synthesizes a new DNA strand by adding nucleotides to the primers, creating a complementary strand to each of the original strands.

Effective primer design is crucial for robust random access in large DNA data storage systems, ensuring reliable file amplification without cross-hybridization. Key factors include the melting temperature (Tm), the point at which 50% of double-stranded DNA separates into single strands. Tm depends on parameters like Na+ and Mg2+ concentrations in the PCR mixture.
The challenge in DNA storage is designing large sets of primers to address many files while ensuring specificity and compatibility under the same PCR conditions, especially maintaining consistent Tm.

3 Primer Design

As higlighted in the previous section, it is mandatory to design good primers to ensure an effective and specific selection of oligonucleotides encoding a specific document during the PCR process. The primer design plays a major role in the success of the PCR-based random access. In DNA data storage, unlike genomic studies, numerous primers must coexist with minimal interference. Thus, the process is to firstly design primers with specific characterisctics to optimize the PCR. Secondly, all the primers are checked to ensure their compatibility with each other and with the dataset oligos. The design of the primers is implemented in a set of tools called DSPT (Dna Storage Primer Tools). The following programs, written in C, are currently available :

- **DSPgen** generate a set of primers according to a rigorous list of criteria to meet the PCR requirements. Those criterias are designed by biochemical and biotechnology constraints. This is an adaptation of the IThOS software previously developped for designing primers for genomic purpose [5]. More recent methods for calculating the melting temperature [6] [1] , which increase precision, have been added. Additionally, supplementary filters have been implemented to better meet DNA storage requirements. One of these filters addresses secondary structure, avoiding sequences that can form undesired shapes or loops that hinder DNA amplification and data retrieval. At last, it allows the generation of thousands of primers in less than one second.
- **DSPham** checks the Hamming distance between primers and eliminates the minimum set necessary to maximize the number of primers with a Hamming distance above a user-defined threshold.
- **DSPhyb** detects potential hybridization, from a thermodynamical point of view, between primers and DNA sequences by computing the primers stability called ΔG (Gibbs free energy G) between small similar regions of both primers and DNA sequences.

[1] https://eu.idtdna.com/calc/Analyzer/Home/definitions

The DSPT package is available on the following gitlab: https://gitlab.inria.fr/molecularxiv-pc-2/dna-storage-primer-tools

4 Experiments results

For the experiments, the tools have been tested on the JPEG-AIC-03 [7] dataset encoded with JPEGDNA-SFC4-S-R [8]. It includes 10 images and represents a total of 99156 oligonucleotides of length equal to 300. Different primer sets have been generated (using DSPgen) and checked (using DSPhyb) for potential hybridization with the oligonucleotides provided in the JPEG-AIC-03 dataset. Figure 1 illustrates the number of detected hybridization sites. These results suggest that the encoding process can be improved. Introducing thermodynamic verification for each oligonucleotide during encoding could enhance the quality of the oligonucleotides. This improvement would prevent the oligonucleotides from binding to each other and forming secondary structures. Additionally, this tool could be used to verify headers, indexes, and other metadata to ensure accurate data retrieval.

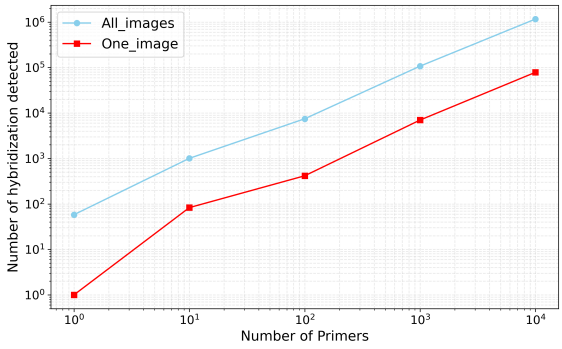

Figure 1: ***DSPhyb - Hybridization detection :*** *Number of hybridization detected while checking if the primers generated can hybridize to the payload of the JPEG-AIC-03 dataset encoded with JPEGDNA-SFC4-S-R.*

5 Conclusion

The DSP software demonstrated excellent performance, particularly in generating efficient primers, which are crucial for PCR-based random access commonly used in molecular biology. With DSP tools, researchers can achieve greater precision and efficiency in their experiments. Additionally, the tools have short execution times, allowing for efficient testing of multiple PCR configurations. Although the software is still under development, the current versions are very promising. The next tool to be developed will check primer compatibility, and a parallel version is planned to further reduce execution times. However, to fully validate these tools further tests are needed, as well as wetlab experiments to ensure the primers meet all specifications.

References

[1] IR Lehman. Discovery of dna polymerase. Journal of Biological Chemistry, 278(37):34733–34738, 2003.

[2] Lilit Garibyan et Nidhi Avashia. Research techniques made simple: polymerase chain reaction (pcr). The Journal of investigative dermatology, 133(3):e6, 2013.

[3] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, et al. Random access in large-scale dna data storage. Nature biotechnology, 36(3):242–248, 2018.

[4] Jian Ye, George Coulouris, Irena Zaretskaya, Ioana Cutcutache, Steve Rozen, et Thomas L Madden. Primer-blast: a tool to design target-specific primers for polymerase chain reaction. BMC bioinformatics, 13:1–11, 2012.

[5] Nouri Ben Zakour, Michel Gautier, Rumen Andonov, Dominique Lavenier, Marie-Françoise Cochet, Philippe Veber, Alexei Sorokin, et Yves Le Loir. GenoFrag: software to design primers optimized for whole genome scanning by long-range PCR amplification. Nucleic Acids Research, 32(1):17–24, 01 2004.

[6] Richard Owczarzy, Bernardo G Moreira, Yong You, Mark A Behlke, et Joseph A Walder. Predicting stability of dna duplexes in solutions containing magnesium and monovalent cations. Biochemistry, 47(19):5336–5353, 2008.

[7] Michela Testolina, Vlad Hosu, Mohsen Jenadeleh, Davi Lazzarotto, Dietmar Saupe, et Touradj Ebrahimi. Jpeg aic-3 dataset: Towards defining the high quality to nearly visually lossless quality range. pages 55–60, 2023.

[8] Xavier Pic, Eva Gil San Antonio, Melpomeni Dimopoulou, et Marc Antonini. Rotating labeling of entropy coders for synthetic dna data storage. Dans 2023 24th International Conference on Digital Signal Processing (DSP), pages 1–5, 2023.