Report from Dagstuhl Seminar 24511

Coding Theory and Algorithms for Emerging Technologies in Synthetic Biology

R. B.*[1], Olgica Milenkovic*[2], Zohar Yakhini*[3], Yonatan Yehezkeally*[4], Anisha Banerjee†[5], and Frederik Walter†[6]

1 **TU München, DE.** ge37bop@mytum.de
2 **University of Illinois – Urbana Champaign, US.** milenkov@uiuc.edu
3 **Reichman University – Herzliya, IL.** zohar.yakhini@runi.ac.il
4 **TU München, DE.** yonatan.yehezkeally@tum.de
5 **TU München, DE.** anisha.banerjee@tum.de
6 **TU München, DE.** frederik.walter@tum.de

Abstract

The progress in understanding genes and genomes has given a boost to the use of synthetic DNA for biological and technological applications. Synthetic nucleic acids play a central role in synthetic biology and in emerging therapeutic paradigms, e.g., genome editing and nucleic acid vaccines. DNA-based data storage is making a significant progress, and thanks to its extreme data-density, its high durability and its timelessness, it is promising to be the next standard for data archival systems.

Synthetic biology and the use of synthetic DNA for information storage applications bring important algorithmic and data analysis challenges. In synthetic biology, reagent and assay design are often driven by algorithmic approaches. Novel synthesis technologies offer cost-reduction by several orders of magnitude at the cost of increased error-rate, raising new coding-theoretic questions.

This Dagstuhl Seminar brought together researchers working on different aspects of synthetic biology and applications in bio-informatics and informatics; the diverse crowd at the seminar included chemists, biologists, computer scientists, and communication engineers. It successfully honed in on the interplay between these fields, allowing participants a window into the insights of other disciplines, and bringing all of these together to bear on current challenges. Going forward, connections forged during the Dagstuhl Seminar will promote interdisciplinary collaboration between participants and their respective networks.

Seminar December 15–20, 2024 – https://www.dagstuhl.de/24511

2012 ACM Subject Classification Mathematics of computing → Coding theory; Information systems → Information systems applications; Applied computing → Life and medical sciences

Keywords and phrases Bio-informatics, Error-correcting codes, synthetic biology

Digital Object Identifier 10.4230/DagRep.14.12.46

* Editor / Organizer

† Editorial Assistant / Collector

Coding Theory and Algorithms for Emerging Technologies in Synthetic Biology, *Dagstuhl Reports*, Vol. 14, Issue 12, pp. 46–62
Editors: R. B., Olgica Milenkovic, Zohar Yakhini, and Yonatan Yehezkeally
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

R. B. (TU München, DE)
Olgica Milenkovic (University of Illinois – Urbana Champaign, US)
Zohar Yakhini (Reichman University – Herzliya, IL)
Yonatan Yehezkeally (TU München, DE)

Designing DNA-based storage systems intrinsically requires joint efforts between biologists, chemists, engineers, and computer scientists, as prominent differences from classical storage media exist at all stages. For example, cost-effective synthesis introduces insertion and deletion errors on top of the well-understood substitution errors occurring in classical media, and much less is known for correcting these. Error-correction techniques could also be affected by the targeted application due to the intrinsic properties of the stored data and the effects of the different types of errors (e.g., this phenomenon is observed when storing images in DNA-based storage systems). Further, strands in the storage container are not ordered in the memory, thus, during sequencing, it is not possible to distinguish which strand is being read, making error correction even more challenging. Lastly, sequencing techniques allow observing many copies of erroneous sub-sequences of the stored strands, which can be leveraged to reconstruct the stored strands more efficiently.

The first large-scale experiments that demonstrated the potential of *in vitro* DNA storage were reported by Church et al., who recovered 643KB of data [1], and Goldman et al., who accomplished the same task for a 739 KB message [2]. However, both of these groups did not recover the entire message successfully due to the lack of using the appropriate coding solutions to correct errors. Most published studies report that either substitutions or deletions are the most prominent error types in DNA-based storage systems, depending upon the specific technology for synthesis and sequencing. Thus, coding-theoretic aspects of DNA-based storage systems have received significant attention recently. However, these theoretical works have not yet led to viable storage technologies.

The progress in enabling information storage in DNA has been driven by the progress in using synthetic DNA in more general applications. In [3], the authors demonstrated how high throughput synthesis can be used to understand, optimize, and fine-tune the functionality of biological systems. This work involved careful design of the reagents – the composition of a large library of candidates – as well as rigorous statistical data analysis to support the result. This Dagstuhl Seminar, therefore, covered general design and analysis frameworks for high throughput experiments. In particular, one specific (and prominent) type of high throughput experiments that is strongly related to synthetic DNA is CRISPR screening. These experiments involve the silencing or the activation of a large number of elements in genomes to allow for optimizing and tuning certain outcomes, ranging from growth rates in plants and bovine cultures to insinuating immune responses in cancer patients.

Informed by this observation, this seminar ultimately aimed at forging closer connections between information theorists, computer scientists, data scientists, biologists, and chemists to: (i) drive joint progress in coding-theoretic techniques specifically tailored to the emerging synthesis sequencing technologies; (ii) have a better understanding of, and initiate innovation in, the application of computer science techniques for high throughput experimental synthetic biology; and (iii) shape an application-driven design of low-error cost-effective DNA-based storage systems. The seminar schedule was flexibly designed, allowing participants to present their research and expertise while interactively accommodating audience input. The plenaries exposed participants to assorted underlying fields, namely genetic code, CRISPR and gene editing, bio-informatics, informatics and machine learning for medical applications, the utility of coding theory for DNA-based information systems, and market data-storage applications. Meanwhile, working groups enabled participants to leverage interdisciplinary backgrounds and share their knowledge and expertise to envision holistic solutions to contemporary challenges. To advertise the research of junior participants, the schedule included a handful of short talks to showcase their results. During the discussions, a couple of participants noticed the implications of their research on the discussed topics; therefore, "pop-up" talks were scheduled for those participants to share their thoughts. Throughout, much fun was had, and connections were forged in a myriad of ice-breaking activities.

References

1 G. M. Church, Y. Gao, and S. Kosuri, "Next-generation digital information storage in DNA," *Science*, no. 6102, pp. 1628–1628, Sep. 2012.

2 N. Goldman, P. Bertone, S. Chen, C. Dessimoz, E. M. LeProust, B. Sipos, and E. Birney, "Towards practical, high-capacity, low-maintenance information storage in synthesized DNA," *Nature*, no. 7435, pp. 77–80, Jan. 2013.

3 E. Sharon, Y. Kalma, A. Sharp, T. Raveh-Sadka, M. Levo, D. Zeevi, L. Keren, Z. Yakhini, A. Weinberger, and E. Segal, "Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters," *Nature biotechnology*, vol. 30, no. 6, pp. 521–530, 2012.

2 Table of Contents

Executive Summary

Overview of Talks

Working groups

3 Overview of Talks

3.1 Scalable and Robust DNA-based Storage via Coding Theory and Deep Learning

Daniella Bar-Lev (Technion – Haifa, IL)

Joint work of Daniella Bar-Lev, Itai Orr, Omer Sabary, Tuvi Etzion, Eitan Yaakobi
Main reference Daniella Bar-Lev, Itai Orr, Omer Sabary, Tuvi Etzion, Eitan Yaakobi: "Scalable and robust DNA-based storage via coding theory and deep learning." Nat Mach Intell, 2025.
URL https://doi.org/10.1038/s42256-025-01003-z

The global data sphere is expanding exponentially, projected to hit 180 Zettabytes by 2025, whereas current technologies are not anticipated to scale at nearly the same rate. DNA-based storage emerges as a crucial solution to this gap, enabling digital information to be archived in DNA molecules. This method enjoys major advantages over magnetic and optical storage solutions such as exceptional information density, enhanced data durability, and negligible power consumption to maintain data integrity. To access the data, an information retrieval process is employed, where some of the main bottlenecks are the scalability and accuracy, which have a natural tradeoff between the two. In this talk, a modular and holistic approach that combines Deep Neural Networks (DNN) trained on simulated data, Tensor-Product (TP) based Error-Correcting Codes (ECC), and a safety margin mechanism into a single coherent pipeline, was presented. The solution was demonstrated on 3.1MB of information using two different sequencing technologies. Our work improves upon the current leading solutions by up to ×3200 increase in speed, 40% improvement in accuracy, and offers a code rate of 1.6 bits per base in a high noise regime. In a broader sense, the work shows a viable path to commercial DNA storage solutions hindered by current information retrieval processes.

3.2 Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design

Nick Goldman (European Molecular Biology Laboratory – Hinxton, GB)

Main reference Lukas Weilguny, Nicola De Maio, Rory Munro, Charlotte Manser, Ewan Birney, Matthew Loose, Nick Goldman: "Dynamic, adaptive sampling during nanopore sequencing using Bayesian experimental design." Nat Biotechnol no. 41, pp. 1018–1025, 2023.
URL https://doi.org/10.1038/s41587-022-01580-z

Recently I have been working on algorithms that enable more efficient use of Oxford Nanopore Technologies (ONT) nanopore sequencing devices. This paper describes our first work on this topic, and we have additional methods under development that were discussed during the Dagstuhl Seminar.

I was interested to find out more about the needs for improved DNA (or RNA, or modified-DNA, or modified-RNA, etc.) reading from the DNA-storage community.

3.3 Contemporary applications for DNA as an information medium

Robert Grass (ETH Zürich, CH)

Joint work of Francesca Granito

Synthetic DNA is seeing increasingly diverse applications in the context of DNA as an information storing and carrying medium. My talk covered some aspects of these applications. One issue is DNA stability; my lab developed pipelines and approaches to extend the half life time of stored DNA and to develop reconstruction methods, both chemical and computational. In other settings, we also take advantage of DNA degradation, e.g., to verifiably demonstrate that a specific product has been recycled. I also touched on different DNA synthesis technologies as well as on many applications in different domains, presenting work coming from my lab.

3.4 Random access coverage depth problem

Anina Gruica (Technical University of Denmark – Lyngby, DK)

Joint work of Anina Gruica, Daniella Bar-Lev, Alberto Ravagnani, Eitan Yaakobi
Main reference Anina Gruica, Daniella Bar-Lev, Alberto Ravagnani, Eitan Yaakobi: "A Combinatorial Perspective on Random Access Efficiency for DNA Storage", in Proc. of the IEEE International Symposium on Information Theory, ISIT 2024, Athens, Greece, July 7-12, 2024, pp. 675–680, IEEE, 2024.
URL https://doi.org/10.1109/ISIT57864.2024.10619151

In this talk, we will explore the fundamental limits of the random access coverage depth problem in the context of DNA data storage. The focus will be on understanding how many DNA strands need to be read to reliably decode a specific piece of information requested by the user from a large pool of encoded strands. The setup involves information strands encoded into strands using a generator matrix, with the strands being read uniformly at random during the sequencing process. The key question is: How many reads are needed, on average, to recover a particular information strand of interest? We will introduce new techniques to address this problem for arbitrary generator matrices, capturing its combinatorial nature. We also introduce recovery balanced codes and show that for this special family of codes, the expected number of reads is always. This approach provides a way to determine the random access expectation for various families of codes, including MDS codes, Hamming codes, and simplex codes. Finally, we will present results on modified systematic MDS matrices, demonstrating how a specific combination of encoded strands and replicated information strands can reduce the required number of reads for retrieving information strands.

3.5 Synthetic promoters and synthetic gene circuits

Lior Nissim (The Hebrew University of Jerusalem, IL)

Joint work of Lior Nissim, Ming-Ru Wu, Erez Pery, Adina Binder-Nissim, Hiroshi I. Suzuki, Doron Stupp, Claudia Wehrspaun, Yuval Tabach, Phillip A. Sharp, Timothy K. Lu
Main reference Lior Nissim, Ming-Ru Wu, Erez Pery, Adina Binder-Nissim, Hiroshi I. Suzuki, Doron Stupp, Claudia Wehrspaun, Yuval Tabach, Phillip A. Sharp, Timothy K. Lu: "Synthetic RNA-based immunomodulatory gene circuits for cancer immunotherapy.", Cell, no. 171, pp. 1138–1150, 2017.
URL https://doi.org/10.1016/j.cell.2017.09.049

Our lab is dedicated to engineering synthetic gene circuits that precisely regulate and target specific cell states, ensuring maximal efficacy while minimizing off-target effects. Our research spans across fundamental biological studies, cancer immunotherapy, adoptive cell therapies, antiviral vaccines, and more, with a strong emphasis on developing orthogonal biological systems for diverse biomedical applications. This talk will focus on two main platforms we developed:

Synthetic Promoters with Enhanced Cell-State Specificity (SPECS)

Identifying cell-state specific promoters is crucial for both basic and applied research but remains challenging. To address this, we developed a high-throughput synthetic promoter engineering and screening platform capable of generating compact promoters with enhanced specificity for virtually any cell state, without the need for prior gene regulation data. This platform has been successfully applied to discover SPECS for a wide array of normal and disease cell states, such as SPECS for chemotherapy resistant cancer cells primary immune cells.

Scalable Synthetic Gene Circuits

We develop scalable platforms that perform complex logic operations, paving the way for new therapeutic approaches and biotechnological innovations. We recently developed circuit that execute various multi-input logic gates and computational operations.

References

1 L. Nissim & R. H. Bar-Ziv, "A tunable dual-promoter integrator for targeting of cancer cells.", *Molecular Systems Biology* no. 6, p. 444, 2010.
2 L. Nissim, S. D. Perli, A. Fridkin, P. Perez-Pinera, T. K. Lu, "Multiplexed and Programmable Regulation of Gene Networks with an Integrated RNA and CRISPR/Cas Toolkit in Human Cells.", *Molecular Cell*, no. 54, pp. 698–710, 2014.
3 M. Morel, R. Shtrahman, V. Rotter, L. Nissim, R. H. Bar-Ziv, "Cellular heterogeneity mediates inherent sensitivity–specificity tradeoff in cancer targeting by synthetic circuits.", *Proc. Natl. Acad. Sci. U.S.A.* no. 113, pp- 8133–8138, 2016.
4 L. Nissim, et al., "Synthetic RNA-based immunomodulatory gene circuits for cancer immunotherapy.", *Cell* no. 171, pp. 1138–1150, 2017.

3.6 DNA Storage with Expanded Alphabets and Calling Chemically Modified Bases

Olgica Milenkovic (University of Illinois – Urbana Champaign, US)

Joint work of Olgica Milenkovic, Kasra Tabatabei, Chao Pan, Alvaro Hernandez, Min Chen, Charles Schroeder
Main reference S. Kasra Tabatabaei, Bach Pham, Chao Pan, Jingqian Liu, Shubham Chandak, Spencer A. Shorkey, Alvaro G. Hernandez, Aleksei Aksimentiev, Min Chen, Charles M. Schroeder, Olgica Milenkovic: "Expanding the Molecular Alphabet of DNA-Based Data Storage Systems with Neural Network Nanopore Readout Processing", Nano Letters, Vol. 22(5), pp. 1905–1914, 2022.
URL https://doi.org/10.1021/acs.nanolett.1c04203

One of the main issues associated with DNA-based data storage system implementations is the high cost and large latency of DNA synthesis. Current estimates are that recording one bit of information via synthesis takes off the order of seconds, which is prohibitive for practical applications.

To mitigate this issue, in our prior work, we proposed the first known molecular storage system that uses both native and chemically modified bases to expand the molecular alphabet used during recording. While expanded alphabets may be beneficial for recording, they may be a challenge to read (sequence) using existing technologies. To this end, we examine how chemically modified bases can be classified using nano pore and PacBio sequencers, and propose tailor-made machine learning methods that can clean the current and kinetics signals provided by the readers and lead to high readout accuracy.

3.7 Errors and Edits in the Genetic Code

Tzachi Pilpel (Weizmann Institute of Science – Rehovot, IL)

The genetic code serves as a molecular mapping function, translating triplets of nucleotides into the 20 amino acids that comprise proteins. The Central Dogma of Biology outlines the processes of transcription of DNA into RNA and the subsequent translation of RNA into proteins. Decoding of information within genes occurred through codons – triplets of nucleotides in RNA that are interpreted by tRNA molecules with complementary anticodons.

While accurate execution of this code is essential for producing unique protein sequences from each gene, the inherent molecular processes within cells inevitably lead to errors.

We present new experimental and computational methods for mapping errors that occur during transcription and translation. These errors include the incorporation of incorrect amino acids, shifts in the reading frame that alter the resulting amino acid sequence due to a single nucleotide shift, and the violation of translation stop codon signals leading to unintended protein extensions. We further discuss post-synthesis chemical editing of RNA molecules, which modulates the genetic information encoded in DNA and allows a single gene to produce multiple protein products.

One of the primary challenges we face is distinguishing between neutral, deleterious, and beneficial errors and edits. While some events may have no selective advantage, and even disadvantage, evolution appears to favor instances where diversity confers adaptability. We provide compelling examples illustrating how translation errors enable organisms to create and propagate protein diversity faster than DNA mutations can generate genetic variation.

Our findings suggest that errors and edits within the Central Dogma represent a newly recognized mechanism through which evolution fosters functional diversity, upon which natural selection can act.

3.8 From Concept to Automation: Realising a Composite Motif-Based DNA Data Storage System

Nimesh Pinnamaneni (Helixworks Technologies Ltd. – Cork, IE)

In this session, we'll explore the Composite Motif-Based DNA Data Storage system developed by Helixworks as part of the MoSS project. This system takes DNA data storage a step further by significantly increasing the amount of data encoded per synthesis operation, from 2 bits to 48 bits. At the same time, it improves data retrieval accuracy, enabling effective recovery even with low oligo read coverage.

The talk will provide a technical overview of our fully automated, end-to-end system. It integrates encoding, oligo synthesis, sequencing, and data retrieval into one streamlined workflow.

Key outcomes include the successful synthesis of oligonucleotides up to 625 nucleotides long, scaled to $1,000$ oligos per run. The system achieved a 98% motif recovery rate when oligos were sequenced at 32 attomoles each and a 60% recovery rate at just 1 attomole each. This indicates that smaller DNA amounts can be used effectively in combination with error-correction codes for data recovery. Additionally, it features real-time base-to-bit decoding powered by Nanopore sequencing.

3.9 Bivariate Monotonic Functions as Novel Models for Complex Biological Phenotypes

Benno Schwikowski (Institut Pasteur & LIX – Paris, FR & MPI – Berlin, DE)

Joint work of Benno Schwikowski, Iryna Nikolayeva, Océane Fourquet, Anavaj Sakuntabhai, Pierre Bost, Ido Amit, Frederik Gwinner

Main reference Iryna Nikolayeva, Pierre Bost, Isabelle Casademont, Veasna Duong, Fanny Koeth, Matthieu Prot, Urszula Czerwinska, Sowath Ly, Kevin Bleakley, Tineke Cantaert, Philippe Dussart, Philippe Buchy, Etienne Simon-Lorière, Anavaj Sakuntabhai, Benno Schwikowski: "A Blood RNA Signature Detecting Severe Disease in Young Dengue Patients at Hospital Arrival", The Journal of Infectious Diseases, Vol. 217(11), pp. 1690–1698, 2018.

URL https://doi.org/10.1093/infdis/jiy086

Advances in molecular profiling technologies are propelling medicine toward a future that is more predictive, preventative, personalized, and participatory. A core challenge is to leverage the vast, complex datasets generated by these technologies to build computational models that can meaningfully link molecular data to high-level disease phenotypes. Such models could provide valuable insights, improve diagnostics, and guide personalized treatment strategies.

This task demands machine learning models that are flexible enough to capture the biological complexity of these relationships, data-efficient enough to operate on typically small clinical datasets, and interpretable to allow insights to be contextualized within established biological knowledge.

3.10 Coding Over Coupon Collector Channels for Combinatorial Motif-Based DNA Storage

Roman Sokolovskii (Imperial College London, GB)

Joint work of Roman Sokolovskii, Parv Agarwal, Luis Alberto Croquevielle, Zijian Zhou, Thomas Heinis
Main reference Roman Sokolovskii, Parv Agarwal, Luis Alberto Croquevielle, Zijian Zhou, Thomas Heinis: "Coding Over Coupon Collector Channels for Combinatorial Motif-Based DNA Storage", IEEE Transactions on Communications, pp. 1–1, 2024.
URL https://doi.org/10.1109/TCOMM.2024.3506938

Synthesising DNA strands using combinations of short sequences (motifs) from a fixed motif library is an interesting alternative to nucleotide-by-nucleotide synthesis because it could potentially reduce the cost of synthesis – a major obstacle to adopting DNA storage for long-term archival data. We discuss channel errors observed in an empirical dataset provided to us by HelixWorks, propose channel models that mimic these types of errors, and discuss the trade-offs associated with different coding schemes and code parameters for the proposed channel models.

3.11 The chemical complexity of nucleic acids in synthetic and molecular biology, e.g. CRISPR-Cas

Mark Somoza (Leibniz-Institut für Lebensmittel-Systembiologie – Freising, DE)

Main reference Jürgen Behr, Timm Michel, Maya Giridhar, Santra Santhosh, Arya Das, Hamed Sabzalipoor, Tadija Kekić, Etkin Parlar, Igor Ilić, Gisela Ibáñez-Redín, Erika Schaudy, Jory Lietard, Thomas Schletterer, Max Funck, Mark Somoza: "An open-source advanced maskless synthesizer for light-directed chemical synthesis of large nucleic acid libraries and microarrays", ChemRxiv, 2024
URL https://doi.org/10.26434/chemrxiv-2024-j4c90

This talk is intended to provide an introduction to the relevance of non-canonical and chemically-engineered nucleic acids in synthetic as well as molecular biology. First, some of the non-canonical modifications are introduced. These include non-canonical nucleotides relevant to epigenetics, as well as chemically-designed modifications engineered to augment specific properties of DNA or RNA, such as adding nuclease resistance or increasing thermodynamic stability, while simultaneously maintaining essential recognition by nucleic acid processing enzymes. These modifications are known to be essential for nucleic acid therapeutics and therefore are an important research topic. Large-scale synthesis of DNA or RNA libraries with natural or engineered chemical modifications can be accomplished using nucleic acid photolithography.

3.12 Capacity of Noisy Permutation Channels

Jennifer Tang (MIT – Cambridge, US)

Main reference Jennifer Tang, Yury Polyanskiy: "Capacity of Noisy Permutation Channels", IEEE Trans. Inf. Theory, Vol. 69(7), pp. 4145–4162, 2023.
URL https://doi.org/10.1109/TIT.2023.3247812

I discuss how to find the capacity of noisy permutation channels. The noisy permutation channel models DNA storage systems where strands of DNA are abstracted as symbols. The individual symbols go through a process where they are subjected to memoryless noise and a uniform permutation is then applied to the collection of symbols. The capacity the noisy permutation channel represents the amount of information which can be stored in a DNA storage system with such properties. We find the exact capacity for strictly positive noise matrices using Kullback-Leibler (KL) divergence covering and a result bounding the KL divergence between sampling from a distribution with and without replacement.

3.13 Estimation of nanopore DNA signatures

Emanuele Viterbo (Monash University – Clayton, AU)

Joint work of Emanuele Viterbo, Brendon McBain, Adrian Vidal
Main reference Adrian Vidal, V. B. Wijekoon, Emanuele Viterbo: "Concatenated Nanopore DNA Codes", IEEE Transactions on NanoBioscience, Vol. 23(2), pp. 310–318, 2024.
URL https://doi.org/10.1109/TNB.2024.3350001

During the translocation of a single-strand DNA molecule through a nanopore channel, its real-time conductance level is determined by the sequence of δ nucleotides, also known as a δ-mer, at its narrowest constriction. The nanopore sequencer produces noisy piecewise constant signals. Under the δ-mer model, the pore goes through k distinct states when the input DNA sequence has k distinct δ-mer substrings. In this talk, we denoise multiple nanopore signals drawn from the same Gaussian-output, DNA sequence-specific, hidden Markov model (HMM), and recover its state-dependent means. We assume that each HMM state i can only transition to itself or to state $i+1$, and has Gaussian-distributed emissions with mean x_i. Starting from a Gaussian prior on x_i, we use importance sampling to approximate the conditional expectation of x_i given the observations. We test the algorithm's performance using both simulated and experimental nanopore signals generated by Oxford Nanopore Technologies' R10.4.1 nanopore.

References

1 B. McBain and E. Viterbo, "An Information-Theoretic Approach to Nanopore Sequencing for DNA Storage", *IEEE BITS the Information Theory Magazine*, vol. 3, no. 3, pp. 95-108, Sept. 2023, doi: https://doi.org/10.1109/MBITS.2024.3355883.

3.14 Towards a Framework for In-Product Authentication

Frederik Walter (TU München, DE)

Joint work of Frederik Walter, Anisha Banerjee, Antonia Wachter-Zeh, Arya Das, Anne Lüscher, Daniella Bar-Lev, Eitan Yaakobi, Francenca Granito, Hamed Sabzalipoor, Jessica Bariffi, Mark Somoza, Omer Sabary, Robert Grass, Sharon Istvánffy, Yonatan Yehezkeally, Zohar Yakhini

Counterfeit products result in an estimated loss of 45 billion EUR annually, affecting industries such as pharmaceuticals, food, and luxury goods. Ensuring the traceability of products is crucial, particularly for textiles, gemstones, critical components like airplane parts, and biological products. This talk proposes a framework for in-product authentication by embedding unique identifiers directly into products.

The proposed in-product authentication method must meet several criteria: it must be evaluable, reproducible, embeddable, unclonable, unpredictable, unique, one-way, tamper-evident, and property-preserving. The authentication process involves a challenge $\mathbf{u} \in \mathcal{A}_1^k$ resulting in a response $\mathbf{r} \in \mathcal{A}_2^m$, which is then processed to extract features $\mathbf{c} \in \mathcal{A}_3^n$.

Inspired by the EUF-CMA model, we propose a security protocol involving:

1. Authentic substance and evaluation function: $\theta : \mathbf{u} \mapsto \mathbf{c}$
2. Function evaluation by an attacker: $(\mathbf{u}_1, \dots, \mathbf{u}_q) \mapsto (\mathbf{c}_1, \dots, \mathbf{c}_q)$
3. The attacker must: a) Predict the next outcome b) Create a new substance

We identify three attack scenarios: dilution, readout and reproduce, and amplification. To quantify security, we aim to draw parallels to classical cryptography, where security is measured in bits where L bit represents that the most efficient attack has complexity 2^L. For in-product authentication, we consider lab operations, material costs, time costs, and product entropy. We propose various entropy measures: entropy of resulting features, intra-device entropy for different challenges, inter-device distance entropy for the same challenge, and conditional entropy given ℓ previous results.

The outlook includes further elaborating the framework for comparing in-product authentication schemes, creating more schemes and security mechanisms, and establishing a universal security level.

3.15 Secret Sharing for DNA Probability Vectors

Zhiying Wang (University of California – Irvine, US)

Joint work of Zhiying Wang, Wenkai Zhang
Main reference Wenkai Zhang, Zhiying Wang: "Secret Sharing for DNA Probability Vectors", in Proc. of the IEEE International Conference on Communications, ICC 2024, Denver, CO, USA, June 9-13, 2024, pp. 4578–4583, IEEE, 2024.
URL https://doi.org/10.1109/ICC51166.2024.10623016

Emerging DNA storage technologies use composite DNA letters, where information is represented by probability vectors, leading to higher information density and lower synthesis cost. However, the issue of information leakage needs to be addressed due to untrusted storage providers. This paper introduces an asymptotic ramp secret sharing scheme (ARSSS) for distributed secret information storage using composite DNA letters. This innovative scheme, inspired by secret sharing methods over finite fields and enhanced with a modified matrix-vector multiplication operation for probability vectors, achieves asymptotic information-theoretic data security for a large alphabet size. Moreover, this scheme reduces the number of reading operations for DNA samples compared to traditional schemes and therefore lowers the complexity and the cost of DNA-based secret sharing.

3.16 Why "we" love DNA... Coding is not "just" correcting errors

Eitan Yaakobi (Technion – Haifa, IL)

Joint work of Eitan Yaakobi, Omer Sabary,, Daniella Bar-Lev, Ryan Gabrys, Anina Gruica, Alberto Ravagnani
Main reference Daniella Bar-Lev, Omer Sabary, Ryan Gabrys, Eitan Yaakobi: "Cover Your Bases: How to Minimize the Sequencing Coverage in DNA Storage Systems", IEEE Trans. Inf. Theory, Vol. 71(1), pp. 192–218, 2025.
URL https://doi.org/10.1109/TIT.2024.3496587

DNA-based storage has attracted significant attention due to recent demonstrations and given the trends in cost decreases of DNA synthesis and sequencing, it is estimated that within the next 5 years DNA storage may become a highly competitive archiving technology. This technology introduces new challenges in finding coding solutions to address various problems associated with the implementation of DNA-based storage systems. In this talk, I will present several algorithmic and coding problems that are motivated by the DNA-based storage channel. Many of them are motivated by the special behavior of the errors in DNA which are mainly insertions, deletions, and substitutions. Then, I will discuss how to reduce not only the cost but also the latency of DNA storage by initiating the study of the DNA coverage depth problem, which aims to reduce the required number of reads to retrieve information from the storage system. Under this framework, our main goal is to understand the effect of error-correcting codes and retrieval algorithms on the required sequencing coverage depth.

References

1 A. Gruica, D. Bar-Lev, A. Ravagnani, and E. Yaakobi, "A Combinatorial Perspective on Random Access Efficiency for DNA Storage", IEEE Int'l Symp. on Information Theory, Athens, Greece, pp. 675–680, July 2024, doi: https://doi.org/10.1109/ISIT57864.2024.10619151.

3.17 Coding, algorithms and complexity for CRISPR and DNA based data storage

Zohar Yakhini (Reichman University – Herzliya, IL)

The talk covered several topics related to coding, algorithms and design, in the context of synthetic biology. We first addressed information aspects of CRISPR activity. CRISPR-Cas systems are supposed to induce (cleavage or other) activity in a very specific manner. However – in practice we often observe detectable off-target activity, This can be obstructive in many applications and dangerous in therapeutic applications. I described work from my group that investigated off-target activity in CRISPR. I have particularly described CRISPECTOR, a machine learning based software tool that supports the measurement of off target activity in experiments done in cells. We then moved to combinatorial schemes for encoding information in DNA. Specifically – we addressed the effect on the number of required synthesis cycles. Using a coupon collector distribution analysis investigated the trade off with higher required synthesis depth. The talk corresponded with several other talks in the seminar, which addressed either CRISPR (eg Somoza, Rak) or combinatorial encoding and its applications (eg Yaakobi, Wang, Sokolovski).

4 Working groups

4.1 What's Next in Synthetic Biology?

Roee Amit (Technion – Haifa, IL)

4.1.1 Challenge 1: Recycling organisms

Our strategy begins with a search for existing REE-binding enzymes. If suitable enzymes are found, we can bypass the directed evolution process. If not, we will cultivate a selection of microbes under increasing concentrations of selected REEs, guiding them toward REE tolerance through natural selection. From these adapted microbes, REE-specific enzymes will be identified for further refinement.

Enzyme optimization can be approached through in silico methods, such as AI-driven design, or in vitro techniques like low-fidelity PCR. This will result in a diverse library of enzyme candidates, which can be screened using a flow-based affinity assay to pinpoint the strongest REE binders.

Once effective REE-binding enzymes are identified, they can be expressed in mixed microbial colonies or sequential fermentations. These engineered systems will aggregate REEs in a broth, from which the aggregates can be purified. These purified REEs can then be reintegrated into the circular economy towards a more sustainable economy.

4.1.2 Challenge 2: 4 Steps towards a habitable Venus

1. Venus simulator chamber:
   To build the bacteria we want to place on Venus, we need to create the conditions on Earth. Therefore, our Venus simulator must have several properties.
   - Low pH value
   - High CO_2 and H_2S content
   - Bacteria do not touch the surface
   - Suitable pressure and temperature
2. Select suitable bacteria
   For example, Cyanobacteria and others. We will need multiple bacteria working in a chain.
3. Remove pathways to unnecessary metabolisms
   This will likely happen due to evolution in the changing environment.
4. Integrate desired properties
   The most important part is photosynthesis, but others will also be necessary.

4.2 CRISPR Library Design

Roni Rak (Agriculture Research Organization – Rishon LeZion, IL)

CRISPR screening has emerged as a powerful tool for systematically unraveling genetic pathways with unprecedented precision. This lecture explored the fundamentals of CRISPR-based screens, including knockout, activation, and interference approaches, and their role in decoding gene functions and regulatory networks. Emphasis was placed on the application of CRISPR screening in advancing cultured meat production, showcasing how these tools can identify key genetic factors that enhance cell proliferation, differentiation, and metabolic efficiency.

The following break-out discussion also addressed broader applications, such as CRISPR's potential in DNA storage systems, leveraging its precision for error correction and random access data retrieval.

4.3 CRISPR

Mark Somoza (Leibniz-Institut für Lebensmittel-Systembiologie – Freising, DE) and Zohar Yakhini (Reichman University – Herzliya, IL)

Joint work of Arya Das, Mayan Rivlin, Guy Assa, Mark Somoza, Zohar Yakhini
Main reference Jürgen Behr, Timm Michel, Maya Giridhar, Santra Santhosh, Arya Das, Hamed Sabzalipoor, Tadija Kekić, Etkin Parlar, Igor Ilić, Gisela Ibáñez-Redín, Erika Schaudy, Jory Lietard, Thomas Schletterer, Max Funck, Mark Somoza: "An open-source advanced maskless synthesizer for light-directed chemical synthesis of large nucleic acid libraries and microarrays", ChemRxiv, 2024
URL https://doi.org/10.26434/chemrxiv-2024-j4c90

Cas proteins are programable endonucleases that are essential components of archaeal and bacterial immune systems and show great promise in nucleic acid therapeutics and synthetic biology due to their ability to knock-out genes as well as for gene editing. The versatility of Cas endonucleases is limited by off-target effects. These off-target effects differ for each Cas protein and are insufficiently understood. Experimental data from my lab indicates that Cas cleavage of DNA targets is poorly correlated to guide RNA hybridization, which invalidates the most commonly used tools used to predict cleavage. The hypothesis is that the Cas-guide RNA complex modifies the binding between the guide RNA and the target DNA in poorly understood ways. In this breakout session, we will discuss experimental strategies to efficiently map out the relative importance of each base involved it the interaction between the guide RNA and the target DNA.

Participants

- Roee Amit
Technion – Haifa, IL
- Iryna Andriyanova
CY Cergy Paris University, FR
- R. B.
TU München, DE
- Anisha Banerjee
TU München, DE
- Daniella Bar-Lev
Technion – Haifa, IL
- Jessica Bariffi
TU München, DE
- Salim El Rouayheb
Rutgers University – Piscataway, US
- Ohad Elishco
Ben Gurion University – Beer Sheva, IL
- Nick Goldman
European Molecular Biology Laboratory – Hinxton, GB
- Alexandre Graell i Amat
Chalmers University of Technology – Göteborg, SE
- Francesca Granito
ETH Zürich, CH
- Robert Grass
ETH Zürich, CH
- Jasper Groen
TU Delft, NL
- Anina Gruica
Technical University of Denmark – Lyngby, DK
- Serge Kas Hanna
Université Côte d'Azur – Sophia Antipolis, FR
- Cai Kui
Singapore University of Technology and Design, SG
- Olgica Milenkovic
University of Illinois – Urbana Champaign, US
- Lior Nissim
The Hebrew University of Jerusalem, IL
- Tzachi Pilpel
Weizmann Institute of Science – Rehovot, IL
- Nimesh Pinnamaneni
Helixworks Technologies Ltd. – Cork, IE
- Inbal Preuss
Technion – Haifa, IL
- Roni Rak
Agriculture Research Organization – Rishon LeZion, IL
- João Ribeiro
IST – Lisbon, PT
- Eirik Rosnes
Simula Research Laboratory – Oslo, NO
- Omer Sabary
Technion – Haifa, IL
- Benno Schwikowski
Institut Pasteur & LIX – Paris, FR & MPI – Berlin, DE
- Ilan Shomorony
University of Illinois – Urbana Champaign, US
- Roman Sokolovskii
Imperial College London, GB
- Mark Somoza
Leibniz-Institut für Lebensmittel-Systembiologie – Freising, DE
- Kasra Tabatabaei
New England Biolabs – Ipswich, US
- Jennifer Tang
MIT – Cambridge, US
- Emanuele Viterbo
Monash University – Clayton, AU
- Van Khu Vu
National University of Singapore, SG
- Frederik Walter
TU München, DE
- Zhiying Wang
University of California – Irvine, US
- Eitan Yaakobi
Technion – Haifa, IL
- Zohar Yakhini
Reichman University – Herzliya, IL
- Yonatan Yehezkeally
TU München, DE