Exposé

Encoding Schemes for Peptide Mass Storage

by

Missaoui Yahya

student ID : 7893687
s9946661@stud.uni-frankfurt.de
November 7, 2025

as part of **Bachelor thesis**
supervised by Prof. Dr. Lena Wiese
at the
Goethe University Frankfurt
Department : Computer Science and Mathematics

Contents

1 Introduction

In a world where digitization is in high demand, the amount of data is growing fast according to some predictions, global data storage could exceed 175 zettabytes by 2025 [1]. Traditional digital storage systems such as magnetic tapes and hard drives are approaching their physical and performance ceilings. For instance, magnetic tapes typically remain reliable for only about 10 to 20 years before their contents must be migrated, a labor-intensive and costly undertaking. Moreover, the areal density that can be packed into magnetic domains is expected to level off soon, restricting any further scalability of these media.[2] Because of these problems, there is more research now into new molecular methods of storing information.

DNA-based storage has seen a considerable improvement in recent years, but peptides are recently also being considered as another option for data storage. They have some benefits compared to DNA, such as a greater variety of building blocks (amino acids), better storage density, and more chemical stability.

Additionally, many synthesis and sequencing techniques are already known from proteomics. For example, commercial peptide synthesis routines enable researchers to assemble peptides residue by residue, providing the single-residue precision required when a peptide strand is intended to embody a string of digital symbols.[2]

On the read-out side, liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS) is the primary instrument for peptide identification; combined with de novo sequencing algorithms it can convert complex fragmentation spectra directly into amino-acid sequences even when no reference database is available.

Together, accurate chemical writing by commercial synthesis and reliable reading through LC-MS/MS assisted de novo sequencing form a well-validated pipeline that this study repurposes as the technical backbone of peptide-based information-storage systems.[2] Recent work has demonstrated that digital information can be directly encoded into peptide chains and then read with mass spectrometry, which allows data to be stored compactly and for a long time.

This exposé presents the plan for a bachelor thesis focused on encoding digital information into peptides. The main goal is to develop and test encoding methods that can translate digital

data into peptide sequences in a reliable and efficient way. The work will look at how to design these codes so they are easy to synthesize, stable, and reduce the chance of errors.

1.1 Scientific Relevance

Recent advancements in molecular data storage have shown that both DNA and peptides are promising alternatives to traditional digital storage media due to their unique physical and chemical properties. DNA storage systems leverage high-density encoding, long-term stability, and energy efficiency, and have already demonstrated the successful recovery of large datasets through techniques such as the fountain code and inner/outer code-based redundancy schemes [3]. These encoding methods not only improve data density but also enable random access and error correction, facilitating the transition of DNA storage from experimental to practical use [3]. However, challenges remain, particularly regarding cost and read-write speed, which ongoing research in enzymatic processing and nanotechnology aims to address [4]. In parallel, peptide-based storage offers new advantages, including a richer set of monomers than DNA and enhanced chemical stability. Notably, the successful encoding and recovery of a MIDI music file using 511 synthetic peptides demonstrates the feasibility of this approach and connects proteomics and peptide synthesis technologies with information storage for the first time [2]. Together, these innovations mark a critical step toward sustainable, high-density molecular data storage systems with cross-disciplinary impact.

1.2 Current State of Research

Molecular data storage is now transitioning from proof-of-concept demonstrations to engineering optimisation on two parallel fronts. DNA systems dominate the field: successive coding innovations—moving from simple bit-to-base maps to the fountain code and concatenated inner/outer Reed–Solomon architectures—have pushed logical densities from 0.215 EB g^{1} to about 17 EB g^{1} while adding random access and robust error correction [3, 4] . Current bottlenecks are economic (≈ US \$0.10–0.30 per nt synthesis) and kinetic (write/read cycles measured in hours), so research is shifting toward enzymatic polymerisation, fast nanopore- or graphene-based read-out, and fully automated micro-fluidic devices to advance DNA data-storage technologies toward eventual commercial viability [4]

Peptide media provide a complementary route: a 2021 study encoded an 848-bit text and a 13.8 kbit MIDI file into 40 and 511 synthetic 18-mer peptides, respectively, and recovered them with tandem mass spectrometry using custom low-density-parity-check and order-tracking codes; each selected amino acid carries 3 bits, giving a theoretical density 3.7 × higher

than DNA and superior chemical durability, though practical density still lags because peptides cannot yet be PCR-amplified [2]

Together these advances define today's research landscape: mature, high-density DNA platforms racing to cut cost and latency, and emerging peptide approaches exploring the richer chemical alphabet for next-generation ultradense, long-lived archival storage.

2 Research Question

Our central inquiry is whether encoding strategies that have proved effective in DNA-based molecular memories, notably variable-length Huffman coding, the droplet-based Fountain architecture, and the constraint-aware Yin Yang codec [3, 5], can be transplanted to peptide sequences, and if so, what adaptations are required. We will test direct portability to the eight-letter peptide alphabet A V L S T F Y E reported in the 2021 peptide storage study [2]. This alphabet was chosen for its balance between chemical stability and analytical clarity: the selected residues exhibit moderate hydrophobicity, making them easy to synthesize and purify, while avoiding amino acids such as cysteine or methionine that are prone to oxidation, or asparagine and glutamine that can undergo side reactions during fragmentation. The result is a chemically robust yet analytically reliable alphabet, well suited for peptide-based data storage experiments. We will further quantify additional design rules that arise from solid-phase peptide synthesis and the liquid chromatography tandem mass spectrometry read-out described in the same work, and evaluate whether the performance metrics used in DNA storage must be complemented by peptide-specific criteria. By answering these questions, we aim to determine how and how far established DNA encoders can inform the development of a robust peptide-based data storage pipeline, without making speculative claims about commercial timelines.

1. Can coding schemes that were developed for DNA, such as Huffman coding, DNA Fountain and Yin Yang coding, be applied to peptide sequences without fundamental changes?

2. If modifications are needed, what specific changes must be introduced to accommodate the peptide alphabet, synthesis constraints and LC MS/MS sequencing?

3. Should additional evaluation metrics be considered for peptide encodings, for example predicted coupling yield during synthesis or fragmentation coverage during sequencing, beyond the logical density and error correction performance used in DNA storage?

3 Methodology

The thesis will proceed through four coordinated phases.

Phase 1. Literature review. Survey recent DNA data-storage encoders and extract benchmarks for logical density, redundancy and error resilience (Huffman coding, DNA Fountain, Yin Yang).

Phase 2. Comparative assessment. Evaluate those schemes against the practical constraints of peptide media, namely the eight-letter alphabet A V L S T F Y E and the write-plus-read workflow (solid-phase synthesis and LC–MS/MS) reported in the 2021 Nature Communications study on peptide data storage.

Phase 3. Code translation. Map the most suitable DNA encoder onto the eight amino acids and add rule sets that mitigate synthesis side reactions and fragment-ion blind zones described in the same study.

Phase 4. In-silico demonstrator. Implement a software prototype that encodes data into virtual peptide strings, applies a channel model with the $\approx 10\,\%$ residue loss and order ambiguity measured experimentally, then decodes the perturbed sequences. Retrieval rate, logical density and redundancy will be compared with the DNA benchmarks from Phase 1 to gauge transfer efficiency.

4 Structure of the Thesis

The thesis is divided into six chapters:

- **Chapter 1** addresses the problem statement, motivation, and research objectives in the context of DNA storage encoding and peptide synthesis as a platform for high-density mass data storage.
- **Chapter 2** provides the theoretical foundations, including biochemical principles of DNA and peptide data encoding, storage density, and stability considerations.
- **Chapter 3** describes the methodology: a comparative empirical analysis of existing encoding and storage strategies based on implementation, evaluation, and feasibility testing.
- **Chapter 4** analyzes classical and current approaches (e.g., DNA-based archival systems, peptide synthesis chains) using defined criteria such as scalability, reliability, and synthesis efficiency.
- **Chapter 5** presents and discusses the results of the experimental analysis and their implications for practical applications in synthetic biology and data storage.
- **Chapter 6** summarizes the key findings and offers a perspective on potential optimizations and directions for future research in peptide-centric mass storage systems.

5 Schedule

- **Week 1-2 – Literature Research**
  - Systematic research in academic databases
  - Categorization of relevant works and setup of a literature management project
- **Week 3 – Theoretical Foundations**
  - Writing the chapter on fundamentals (molecular storage, peptide synthesis, DNA encoding mechanisms)
  - Summary of existing approaches to data storage in biological systems
- **Week 4 – State of Research**
  - Analysis and comparison of current biological data storage technologies
  - Critical evaluation of the methods in terms of scalability, reliability, and biochemical feasibility
- **Week 5 / 6 – Methodology and Analysis**
  - Description of the methodological approach
  - Implementation or simulation for encoding strategies or peptide data pipelines
- **Week 7 – Results and Discussion**
  - Evaluation and interpretation of results
  - Discussion of strengths, limitations, and implications for large-scale peptide synthesis
- **Week 8 – Introduction, Conclusion, and Abstract**
  - Writing the introduction, summary, and outlook
  - Creation of a concise abstract
- **Week 9 – Revision and Refinement**
  - Consistent linguistic and structural optimization

  - Formatting according to university guidelines

- **Final Review and Submission**
  - Final check of citations, layout, and appendices
  - Export, backup, and timely submission

Bibliography

[1] D. Reinsel, J. Gantz, and J. Rydning, "The digitization of the world: From edge to core," White Paper, 2018, international Data Corporation (IDC), Sponsored by Seagate. [Online]. Available: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[2] C. C. A. Ng, W. M. Tam, H. Yin, Q. Wu, P.-K. So, M. Y.-M. Wong, F. C. M. Lau, and Z.-P. Yao, "Data storage using peptide sequences," Journal Article, 2021, nature Communications, 12:4242. [Online]. Available: https://doi.org/10.1038/s41467-021-24496-9

[3] C. Wang, G. Ma, D. Wei, X. Zhang, P. Wang, C. Li, J. Xing, Z. Wei, B. Duan, D. Yang, P. Wang, D. Bu, and F. Chen, "Mainstream encoding–decoding methods of dna data storage," Review Article, 2022, cCF Transactions on High Performance Computing, Volume 4, Pages 23–33. [Online]. Available: https://doi.org/10.1007/s42514-022-00094-z

[4] A. Akash, E. Bencurova, and T. Dandekar, "How to make dna data storage more applicable," Review Article, 2024, trends in Biotechnology, Volume 42, Issue 1. [Online]. Available: https://doi.org/10.1016/j.tibtech.2023.07.006

[5] C. Zhu, Z. Xiao, Z. Li, Y. Zhao, Y. Zhang, Q. Liu, C. Liu, X. Zhang, B. Yang, and H. Yin, "Towards practical and robust dna-based data archiving using the yin–yang codec system," Research Article, 2023, nature Communications, Volume 14, Article number: 470. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10766522/