WAB

Provadis School of International Management and Technology

Exposé

Comparing Suffix Automata Against Suffix Arrays For Longest Common Substring Queries

A Proof-of-Concept Implementation Based On A Real Use Case

Rubin Chempananickal James
rubin.chempananickal-james@stud-provadis-hochschule.de
Matriculation Number: D876

Department: Information Technology
Module: Algorithmen und Datenstrukturen
Reviewer: Prof. Dr. Volker Scheidemann

February 18, 2026

Contents

Exposé

Problem Statement

The Longest Common Substring (LCS) problem is a fundamental problem in computer science, with applications in multiple domains, bioinformatics being one of the most prominent. This is due to the fact that DNA/RNA sequences are usually encoded as strings for bioinformatics workflows, in a format called FASTA, originally described by Pearson and Lipman in 1985[1].

Objectives

Methodology

Planned Structure

The paper will likely be structured as follows:

- Introduction
- Research Question and Objectives
- Literature Review
- Methodology
- Results and Discussion
- Limitations
- Conclusion and Future Work
- References
- AI Declaration
- Declaration of Authorship

[1] Lipman and Pearson 1985.

Bibliography

Blumer, A. et al. (1985). „The smallest automation recognizing the subwords of a text“. In: *Theoretical Computer Science* 40. Eleventh International Colloquium on Automata, Languages and Programming, pp. 31–55. ISSN: 0304-3975. DOI: https://doi.org/10.1016/0304-3975(85)90157-4. URL: https://www.sciencedirect.com/science/article/pii/0304397585901574.

Lipman, David J. and William R. Pearson (1985). „Rapid and Sensitive Protein Similarity Searches“. In: *Science* 227.4693, pp. 1435–1441. DOI: 10.1126/science.2983426. eprint: https://www.science.org/doi/pdf/10.1126/science.2983426. URL: https://www.science.org/doi/abs/10.1126/science.2983426.

Manber, Udi and Gene Myers (1990). „Suffix arrays: a new method for on-line string searches“. In: *Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '90. San Francisco, California, USA: Society for Industrial and Applied Mathematics, 319–327. ISBN: 0898712513.