Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Ondřej Sladký
Thesis title	Masked superstrings for efficient k-mer set representation and indexing
Year submitted	2024
Study program	Computer Science
Specialization	General Computer Science
Review author	Mgr. Pavel Veselý, Ph.D. Advisor
Department	Computer Science Institute of Charles University

Ondřej's thesis focuses on representing and indexing k-mer sets, one of the central concepts in computational genomics, where in certain applications, analyzing k-mer sets significantly reduces computational demands for huge volumes of genomic data coming from sequencing. Yet, state-of-the-art methods still have limitations coming from their reliance on certain structural assumptions, namely the so-called "spectrum-like property".

Theoretical part of the thesis. In Chapter 2, Ondřej proposes a new framework for representing sets of k-mers based on the superstring of k-mers equipped with a binary mask. The resulting concept of *masked superstrings* mathematically unifies and generalizes existing representations, and also removes the reliance on the spectrum-like property. Ondřej developed linear-time algorithms for computing masked superstrings based on approximation algorithms for shortest superstrings. However, these needed to be significantly adapted to the problem and an efficient implementation had to be devised. Among his most interesting results in this part is NP-hardness of minimizing the number of runs of ones in the mask.

In Chapter 4, Ondřej developed a new data structure for indexing genomic data, which is simpler and more memory-efficient compared to state-of-the-art methods. His data structure is based on the FM index, which builds on the Burrows-Wheeler transform.

Chapter 5 further generalizes masked superstrings using a so-called *demasking* function. This leads to a mathematically elegant framework that allows to carry out set operations with sets of k-mers in an elegant way. In combination with the indexing, Ondřej proposed a versatile data type for the analysis of k-mer sets. In contrast, most known data structures for k-mers do not allow for set operations, or do so only in a limited way.

Overall, Ondřej proposed a mathematically elegant concept for designing data structures for k-mers and developed efficient algorithms for it, with most of the ideas originating directly from him. He wrote the text clearly and precisely, including proofs of mathematical statements. The text is suitably supplemented with examples, illustrations, and tables. Ondřej concluded the thesis by outlining several open problems and directions for further research.

Implementation and experimental evaluation. Ondřej implemented the proposed algorithms and data structures in C++ into two programs: KmerCamel for computing masked superstrings and FMSI for indexing and set operations. He optimized both programs well, especially with regard to memory efficiency. As a result, KmerCamel is capable of processing large datasets, such as massive bacterial pangenomes or the human genome, even in cases where other programs cannot manage with even 200 GB of memory (KmerCamel required a maximum

of 151 GB). Additionally, KmerCamel computes more efficient representations of k-mer sets — for randomly subsampled pangenomes, the improvement is up to threefold. FMSI achieves a factor of 1.4 to 4.5 improvement in memory efficiency over the state-of-the-art programs when processing queries for the presence of k-mers in bacterial pangenomic data. These results are clearly presented in several plots in Chapter 6.

Summary and outreach. Ondřej's work makes a very valuable progress on the theory of k-mer sets and data structures for them, with promising experimental results. It is conceivable that the outcomes may get into practice of analyzing large genomic data in the future (we have some information that other research groups start to incorporate masked superstrings into their methods, e.g., Barış Ekim, a PhD student at MIT, reimplemented KmerCamel in Rust and made it available at GitHub).

Based on these results, we are preparing two papers, together with Karel Břinda from Inria who co-supervised the thesis. We plan to submit them to top journals in bioinformatics, namely *Genome Biology*, *Genome Research*, or *Bioinformatics* (all in the first decile); these journals published some of the previous work that we improve upon. Moreover, we are already working on a couple of followup papers building on results in this thesis.

Ondřej has already presented his results on several occasions, including international workshops RECOMB-seq 2023 (Istanbul, Turkey) and Sequences 2024 (London, UK), and an invited talk at research institute Inria, France. All of these talks were attended by experts in the area and were met with great acclaim, which reached me. Finally, the thesis recently won the Czech-Slovak undergraduate student competition SVOČ 2024 in category Theoretical Computer Science.

Without any hesitation, I recommend to give the grade of "1", and suggest the thesis for the Dean's award.

Sincerely,
Pavel Veselý