Report on Bachelor Thesis

Faculty of Mathematics and Physics, Charles University

Thesis author	Ekaterina Milyutina
Thesis title	Efficient representation of k-mer sets
Submitted	2023
Program	Computer Science **Specialization** General Computer Science
Supervisor	Mgr. Pavel Veselý, Ph.D.
Advisor	Karel Břinda, Ph.D.

Review author Pavel Veselý (Computer Science Institute of Charles University)

Summary of the thesis. The thesis explores a new approach for representing k-mer sets (i.e., the set of all length-k substrings obtained from genomic data), which are extensively used in bioinformatics. It is based on the observation that all previous representations can be seen as a superstring of the k-mer set, and there may be shorter superstrings. In particular, the thesis experimentally compares the state of the art, represented by greedily computed simplitigs (which are disjoint paths in the de Bruijn graph of the set) with outcomes of well-known approximation algorithms for the shortest superstring problem (SSP), namely, Greedy and TGreedy. While the presented comparison provides some insight, there are three main gaps:

- As DNA is double-stranded, a k-mer and its reverse complement are considered equivalent. This is not taken into account in the resulting program. Still, the obtained results indicate that similar outcomes would be achieved when we consider reverse complements.
- The thesis is missing a comparison with matchtigs, a more recent concept than simplitigs, also based on de Bruijn graphs.
- The implementation only allowed for an evaluation of the selected methods on a rather small bacterial genome and pangenome for k up to 20, not on larger genomic datasets.

Apart from experimental contributions, the thesis contains a new observation that algorithm TGreedy can be implemented in linear time using the well-known Aho-Corasick automaton. This is basically an adjustment of the linear-time implementation of Greedy by Ukkonen (Algorithmica '90), using that the two algorithms are similar.

Text. As the thesis bridges two topics, k-mer set representations and SSP, the author needed to study and describe both, which is done in Chapters 1 and 2, respectively. Then, Chapter 3 outlines the framework for k-mer set representation based on a superstring and a mask to filter out false positives (i.e., k-long substrings of the superstring which are not in the input set). Chapter 4 discusses the experimental setup and results. The main text of the thesis ends with several conclusions drawn from experimental results and an outline of future directions.

The organization into the chapters is reasonable, and the observations based on data from experiments are written quite well. However, the lower level organization is less clear. For instance, in Chapter 2, high-level descriptions of the algorithms (i.e., how they form the resulting superstring) are interleaved with the descriptions of their implementations. Some parts are not written particularly well, especially inside Chapters 1 and 3. The text sometimes goes into a too high level of technical details (e.g., Section 3.1.1. with a description of how to write down a mask either as a binary string or encode it as letter case in the superstring). Meanwhile, in other parts, important aspects are described only very briefly (e.g., the claim that the implementation of TGreedy indeed runs in liner time requires a proof, or the fact that there may be more masks for a given superstring is not emphasized enough).

The text is accompanied with several illustrations, which help to understand the key concepts, although the figures are not of high quality (sometimes the font inside is too small). The typography needs significant improvements (e.g., variables such as k should be in italics; or there is k-1 instead of $k-1$).

Implementation. The implementation of the chosen algorithms in Python was meant to be prototype from the beginning. However, it substantially lacks optimization of its memory requirements that would allow for an execution of the program on larger datasets and for larger values of k. Therefore, the measured time and memory complexities may be substantially higher than for a well-engineered implementation in C or C++. Still, the relative comparison of the three algorithms provides some insight.

The implementation is stable and the validity of the output representation can optionally be verified. However, more in-depth testing (unit tests) is missing. The code is quite well-organized, and the thesis contains an extensive technical documentation for the code.

Overall, **I recommend the thesis for defense.**

August 27, 2023

Signature: