Report on Bachelor Thesis
Faculty of Mathematics and Physics, Charles University

Title: Efficient representation of k-mer sets
Student: Ekaterina Milyutina
Supervisor: Mgr. Pavel Veselý, Ph.D.
Advisors: Karel Břinda, Ph.D.
Study Programme: Computer Science
Reviewer: doc. Mgr. Kolman Petr, Ph.D.

Summary of contributions

The thesis addresses the problem of efficient representation of k-mer sets. Given a string (or a set of strings - e.g. a DNA sequence or several fragments of it) and an integer k, the corresponding k-mer set is the set of all substrings of length k of the input string(s). In the last decade, many techniques used in big data genomics are based on the analysis of k-mer sets.

The thesis is structured as follows. Chapter 1 provides a brief introduction to k-mer sets and known representations of them. Chapter 2 deals with the Shortest Superstring problem (SSP) and approximation algorithms for the problem. Chapter 3 sketches how to exploit SSP approximation algorithms in the representation of k-mer sets. Finally, Chapter 4 describes the results of experiments performed by the student using her implementation of three algorithms for k-mer representation, one of which is the algorithm exploiting SSP approximation.

Evaluation

I tend to think that the student did satisfy the assignment of the thesis, namely to study state-of-arts methods for k-mer set representation and experimentally compare them. However, I have a couple of objections; I mention the most important ones, in the following part.

Main Comments

1. Throughout the thesis, the author is *very* sloppy about definitions and algorithm descriptions, to the extent that some parts are difficult, if not impossible, to understand. A few examples:

- p. 6 The second paragraph starts with the definition of the *k-mer set*. A k-mer set is not an isolated object; it is always related to a fixed string (or strings). However, this is only implicit in the definition. Two sentences later, a definition of *the size of a k-mer set representation* is given, but no k-mer set *representation* has been defined.
- p. 8 Definition of the *de Bruijn graph*. The standard notion of the de Bruijn graph refers to a particular graph (see, e.g., Wikipedia). Here, the proper term is a de Bruijn graph *of a k-mer set K.*
- p. 9 Definition of *unitig*. “where i $\in$ [1, p-1], has both in-degree ...“ should be replaced by “and for i $\in$ [1, p-1], n_i has both in-degree ...”. The meaning is different!
- p. 10 Definition of *simplitig* is missing.
- p. 11, line 6 There is a reference to a *merging process* that has not been provided.
  line 11 What is *a smallest spectrum-preserved set representation*?
  line 22 What is the USP algorithm?
  line 25 What are *optimal* simplitigs?
  line 29 What is a bigraph?
  line 30 Eulerian cycle should be defined with respect to a graph - this is missing.
- p. 12, line 7 What are *matchtings*? Later (p. 13), what are *optimal* matchtings?
- p. 14 The simplitig algorithm - a formal description should appear here.

- p. 15 The definition of the *approximation ratio* is not correct (complete); also, similarly as in many other cases in the thesis, it is a term that is related to something, namely to an algorithm (i.e., the term that should be defined, is *the approximation ratio of an algorithm XY*).
- p. 17, line 6 under the figure: What is *the largest overlap Hamiltonian path* and how to find it?
- p. 17/18 How can the Hamiltonian path be found using depth-first search algorithms?
- p. 18, line 5 Note that finding *any* Hamiltonian path is an NP-complete problem.
- p. 18 Aho-Corasick automaton - a complete description is missing.
- p. 20, 2nd property of the indices in the list: What is *the Hamiltonian path H* that the definition refers to? Similarly in the description of the TGreedy algorithmm step 1.
- p. 22, line 9 *while minimizing the overall length* - it should be properly justified that the overall length is minimized.
- p. 23 and 24 Complete descriptions of the TGreedy and MGreedy algorithms would be appropriate.
- Section 3 - A formal description of the main algorithm for k-mer set representation that is based on the SSP algorithms is missing. Moreover, the names of the section and its subsections are a bit confusing: the point of the section is not that there is a connection between k-mers and SSP but that the SSP algorithms are exploited to get a new k-mer set representation.

2. Most of the time (or maybe even always), the thesis deals with a k-mer set of a string. However, all the computational experiments are done for a k-mer set of a *set of* strings. This should be pointed out and commented a bit.

Other comments

- p. 8 The first paragraph explains, what are the irreplaceable advantages of the bidirectional approach over the unidirectional one; the second paragraph explains that by using the unidirectional model, nothing is lost.
- p. 9 (and other places) *path through the graph* --> path in the graph
- p. 9, figure: CTA does not seem to be a unitig (it is a fork)
- p. 10 What is a FASTA file?? It should be explained in the text.
- p. 16, line 20 (also p. 22, line 8) Upper bound was proved (not introduced)
- p. 24, *theorem proposed and proved* - remove proposed

Typos

- Figures on p. 9 and 10: *megred* → merged
- p. 20, line 16 What is the meaning of *The algorithm of function Hamiltonian traverses the states* ...?
- p. 22 *inner, outer degree are unit* → in and out degrees are one

Overall Assessment

I recommend to accept the thesis, and I recommend the mark 3.

doc. Petr Kolman, Ph.D.

Prague, August 25 2023