FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

Review report of a final thesis

Reviewer:	Ing. Luboš Krčál
Student:	Bc. Dominika Draesslerová
Thesis title:	Bioinformatics index tool for elastic degenerate string matching
Branch / specialization:	Computer Science
Created on:	6 February 2023

Evaluation criteria

1. Fulfillment of the assignment

[1] assignment fulfilled
▸ **[2] assignment fulfilled with minor objections**
[3] assignment fulfilled with major objections
[4] assignment not fulfilled

> Create a survey of tools for indexing genomes and elastic degenerate string matching. Find and discuss existing self-index data structures.

- The survey is present and provides a sufficient overview of the topic. Both breadth and depth of the survey should be increased. More focus should have been invested especially in the current state-of-the-art related to genome indexing and EDS pattern matching problems.

> Design and implement a self-index based on the BIO-FM index for a collection of genomes allowing efficient search over the elastic degenerate string.

- This is the main contribution of the thesis - an adaptation of a previous algorithm BIO-FMI for indexing and pattern matching in EDS data format.
- The design and implementation were both completed. However, follow-up questions should be addressed in order to determine the extent of the completion of this item.

> Perform an experimental evaluation of the implementation.

- The experimental evaluation is extensive, with many varying attributes. The main caveats pertain to the intended objectives and design of this evaluation.

> Design and implement an application using the BIO-FM index over variable formats.

- The application was implemented according to the assignment.

2. Main written part

50 /100 (E)

- The written part is the weakest part of this thesis.
- The quality of the English language used severely impacts any attempts for an in-depth comprehension of the text. Many sentences could be more comprehensible. I would suggest the author receive extensive assistance from a native or advanced English speaker. Simple proofreading would not help in this case.
- The separation of the survey and the state of the art from the main contributions is not clear. Especially in Chapter 2, with a mix of both.
- The description of the main contribution (index building and pattern matching) on EDS is extremely brief and does not provide sufficient comprehension even to readers with considerable background information.
- The text does not clarify how the EDS version of BIO-FMI operates exactly. Based on sections 2.3.1 to 2.3.3, the main added value is in parsing from EDS format to BIO-FMI format.
- Experimental evaluation is plentiful, however, the results are not discussed accordingly. Often objectives of the experiments are not stated as well.
- The experimental evaluation is not designed well to show the advantages of BIO-FMI over EDS data as opposed to the alternatives. Only a very short section is dedicated to this key comparison.

3. Non-written part, attachments

75 /100 (C)

- The author provides an accompanying source code that covers the functionality described in the text.
- I was unable to verify the correctness of the proposed BIO-FMI extensions from the written description only. The written algorithm description cannot be supplemented by the source code.
- Unfortunately, the accompanying source code does not compile out of the box. A minor intervention was necessary to make the software compile successfully.
- The software behaves as expected on the small sample input files.
- On a comparable machine, the software did not manage to build an index from a single real-world EDS (200 MB, human chromosome). However, experiments on similarly sized synthetic datasets have been done by the author.

4. Evaluation of results, publication outputs and awards

75 /100 (C)

- The implementation of indexer and pattern-matching tools in modern C++ can be extended and used as a basis for further research in related fields.
- The thesis presents an extensive experimental evaluation of artificial datasets, varying many parameters of input data and the underlying BIO-FMI algorithm. Extending the parametrization, clarifying the goals for all experiments, and adding real-world datasets, would be a good starting point for evaluating subsequent EDS indexing and matching algorithms.
- A significant amount of work would be needed to make the results publishable, with an appropriate choice of focus being necessary. The foundation is there.

The overall evaluation 65 /100 (D)

- The contributions of this thesis are not clearly stated. A large part of key chapter #2 describes a BIO-FMI index, which is not part of the contributions of the author. The author should clearly describe her contributions versus contributions from the original BIO-FMI paper, how much of it was reimplemented, and how the EDS pattern matching algorithms differs from the ALN pattern matching and the original matching.
- If the author's contributions on BIO-FMI provably exceed the standalone preprocessing of EDS into the original BIO-FMI format, and if the author can describe how the EDS matching in BIO-FMI is different from the original algorithm, then I propose to increase the overall grade to 75 (C).
- Other parts of the thesis assignment were fulfilled, albeit with several caveats as mentioned elsewhere.
- It appears too much time was sunk into auxiliary tools for data preprocessing and for generating synthetic datasets as opposed to the core contributions.
- For the experimental part, the author could have focused on real-world datasets for EDS, since many are readily available from other projects cited in the work. These were likely out of the scope of this work, yet could have served as the ideal demonstration and could have saved time on the synthetic EDS generator.
- I was unable to gain sufficient understanding and verify the correctness of the proposed BIO-FMI extensions from the written part of the thesis.

Questions for the defense

- What is the main difference between BIO-FMI and your extended algorithm for the EDS format?
- Can you explain why exactly the search time in Figure 4.8 is 6-8 times faster for EDS than ALN?
- When preprocessing EDS before conversion to BIO-FMI, do you also perform some form of simplification of the resulting EDS? Example 2.3.1 shows that all the strings in the EDS segment have the same suffix 'A', which can be extracted from the symbol and reduce the size of the BIO-FMI index
- Why was a comparison with other EDS pattern-matching tools omitted?
- What changes would you make to achieve a well-designed experimental evaluation to focus on the core ideas of this thesis, i.e. complete understanding of BIO-FMI parametrization, and comparison with the extension for EDS data? I.e. can you explain the underlying reason for your observations from Chapter 5?
- In Figure 4.5, why is the search time distribution different for context lengths above 25? Why the search time distribution does not correlate with the number of pattern occurrences, which is presented in Figure 4.1 as strictly decreasing with increasing pattern size?

Instructions

Fulfillment of the assignment

Assess whether the submitted FT defines the objectives sufficiently and in line with the assignment; whether the objectives are formulated correctly and fulfilled sufficiently. In the comment, specify the points of the assignment that have not been met, assess the severity, impact, and, if appropriate, also the cause of the deficiencies. If the assignment differs substantially from the standards for the FT or if the student has developed the FT beyond the assignment, describe the way it got reflected on the quality of the assignment's fulfilment and the way it affected your final evaluation.

Main written part

Evaluate whether the extent of the FT is adequate to its content and scope: are all the parts of the FT contentful and necessary? Next, consider whether the submitted FT is actually correct – are there factual errors or inaccuracies?

Evaluate the logical structure of the FT, the thematic flow between chapters and whether the text is comprehensible to the reader. Assess whether the formal notations in the FT are used correctly. Assess the typographic and language aspects of the FT, follow the Dean's Directive No. 52/2021, Art. 3.

Evaluate whether the relevant sources are properly used, quoted and cited. Verify that all quotes are properly distinguished from the results achieved in the FT, thus, that the citation ethics has not been violated and that the citations are complete and in accordance with citation practices and standards. Finally, evaluate whether the software and other copyrighted works have been used in accordance with their license terms.

Non-written part, attachments

Depending on the nature of the FT, comment on the non-written part of the thesis. For example: SW work – the overall quality of the program. Is the technology used (from the development to deployment) suitable and adequate? HW – functional sample. Evaluate the technology and tools used. Research and experimental work – repeatability of the experiment.

Evaluation of results, publication outputs and awards

Depending on the nature of the thesis, estimate whether the thesis results could be deployed in practice; alternatively, evaluate whether the results of the FT extend the already published/known results or whether they bring in completely new findings.

The overall evaluation

Summarize which of the aspects of the FT affected your grading process the most. The overall grade does not need to be an arithmetic mean (or other value) calculated from the evaluation in the previous criteria. Generally, a well-fulfilled assignment is assessed by grade A.