THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Evaluating Directional and Association Methods on Single-cell RNA Sequencing Data.**
Author's name:	**Eliška Dvořáková**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science.
Thesis reviewer:	Doc. Ing. Jiří Kléma, PhD.
Reviewer's department:	Department of Computer Science.

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment — challenging

How demanding was the assigned project?

This thesis compares a couple of gene network inference methods applied to single-cell RNA sequencing data. The topic is current and challenging as single-cell RNA sequencing data often suffer from phenomenons that were not observed in formerly more frequent bulk RNA sequencing data, the most severe of them is dropout. The thesis aims at understanding of the influence of these phenomenons and it also studies the role of data normalization. The assignment is challenging, it requires understanding of RNA sequencing, the nature of statistical inference methods and their evaluation in non-trivial real-world datasets.

Fulfilment of assignment — fulfilled

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

First, the assignment was to get familiar with scRNA-seq and learn the methods for directional and association inference. These goals were obviously fulfilled, the set of inference methods was clearly given by the supervisor. Secondly, the assignment was to evaluate the methods and apply them to real data. These goals were clearly fulfilled too. The only exception could be that the assignments supposes to test at varying discrete levels, which I interpret as a test with contingency tables of various sizes. This probably happened as the function simulate_tables enables to change the table size and the results mix the individual settings, nevertheless, the thesis does not discuss this issue anyhow. On the other hand, the thesis contains a major normalization section that has not been required.

Methodology — correct

Comment on the correctness of the approach and/or the solution methods.

Conceptually, I find the methodology appropriate and correct.

Technical level — B - very good

Is the thesis technically sound? How well did the student employ expertise in the field of his/her field of study? Does the student explain clearly what he/she has done?

Despite the formal details mentioned below, the thesis seems to be technically sound. The student shows her bioinformatics expertise and properly combines the assignments with the input data and selected methods. I would only appreciate slightly deeper statistical analysis of the reached results. Except for the proposal of a new discretization method, the thesis is mainly experimental and for example bar plots could have error bars to see whether the difference between the individual detection methods matters. I would also propose to provide more detailed results for simulated dataset in order to be able to guess the role of the individual settings on the performance of the individual methods.

Formal and language level, scope of thesis — C - good.

THESIS REVIEWER'S REPORT

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?
The thesis is written in nice and fluent English, the global concepts are clear. However, there are numerous details that make the reader stop and think. First of all, the author frequently confuses the words inference and interaction/association/relationship. The phrases "the ability to detect the inference", "variables with no inference" or "biological inference network construction" make no sense. Secondly, the mathematical notation could be more consistent. For example, the notation In Section 4.1.1 does not match with the previous section, including the expression matrix transposition in the provided R code. Third, some introductory figures are taken over without a reasonable explanation. What is the trajectory inference in Fig 1.1? How does it relate to the topic of the work? Fourth, the author could be more detailed in its description of simulated data. Purely from the text, it is unclear how the data originated. What is the noise? How can we have low frequency fields in the contingency tables without any noise? The functional relationship is clearly not one-to-one then. Is it an artefact caused by discretization? Fifth, there are several inconsistencies that make the thesis less understandable. How would you interpret the first two rows in Table 5.1? What is the meaning of H and L in the same table? Two of the links to Figure 5.21 should actually link to Figure 5.20. In Table 5.2, the relationship between p-values and functional indices does not match the 1-x relationship mentioned in Section 4.2.4

Selection of sources, citation correctness	A - excellen
Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?	
The author deals with 41 references. I find the selection of sources adequate. The student's work was distinguished from the earlier work in the field, the contribution of the thesis is thus clear.	

Additional commentary and evaluation (optional)
Comment on the overall quality of the thesis, its novelty and its impact on the field, its strengths and weaknesses, the utility of the solution that is presented, the theoretical/formal level, the student's skillfulness, etc.
Please insert your comments here.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

Summarize your opinion on the thesis and explain your final grading. Pose questions that should be answered during the presentation and defense of the student's work.

This is a very good work on a very current and non-trivial topic. It meets the requirements of the assignment and despite minor shortcomings it fulfills the general requirements for the master's thesis. The reached results accompanied by a discussion provide guidelines for practitioners to process scRNA-seq data. The grade that I award for the thesis is B - very goo

Questions: The real acute leukemia dataset you dealt with showed a high dropout rate around 90%. Your parallel experiments with simulated data worked with dropouts from 0 to 90%, however, the presented results mix all the settings. Would you provide more details specifically for the simulated data with 90% dropout and the simple size comparable with the real data? Were you able to detect any interactions? Did you observe any differences between the inference methods at this dropout level? What were your expectations for the real data based on the previous simulated experiments? Were they actually met?

Date: Signature: