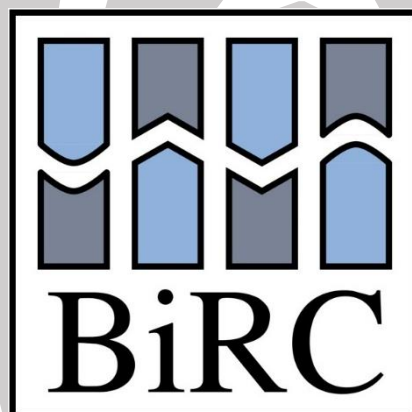


Preprocessing Optimization of LCMS Data from Fingerprints

Solveig Østergaard

201407182



Master's thesis in Bioinformatics

Aarhus University, June 2020

Supervisor: Palle Villesen, Associate Professor

Table of Contents

Abstract	3
Introduction.....	4
The Composition and Aging of Fingerprints	4
Liquid Chromatography/Mass Spectrometry	7
XCMS – The Preprocessing of LCMS Data.....	9
The XCMS Workflow.....	9
The Parameters	11
New Functionality of XCMS.....	12
Optimization of XCMS	12
Other Preprocessing Software.....	13
Data Description	13
Aim and Approach of the Project	16
Results and discussion	18
The Differences Between Final Datasets	18
The Basic Parameter Datasets.....	20
The Optimized Datasets	20
The Feature Distributions of the Datasets	21
Parameter Optimization	22
Collected and Sequential Optimizations	25
QC Optimization	26
Normalization and Choice of Evaluation	27
Conclusion and Future Perspectives	30
Acknowledgements	30
Methods.....	31
The Preprocessing of LCMS Data	31
Preparing the Data – Normalization and Filtering.....	32
Evaluation of the data	32
The combined and sequential optimizations	33
The QC optimization	33

Abbreviations.....	34
References	35
Appendix.....	39
The Distributions of the Datasets	39
The Individual Parameter Optimization.....	41
The Sequential Parameter Optimization	45
The QC Optimization.....	46

Abstract

Fingerprints are used extensively as forensic evidence, and being able to detect and understand the variation in the prints could give us a better chance of identifying individuals in for example criminal cases. One of the variations that could be interesting to study is the degradation of the prints and the possibility of predicting the deposition age for them. This could give us the possibility of establishing if individuals were at a location at a specific time. In this project, fingerprints were deposited at different times, and they were then analyzed with LCMS to detect the differences over time. The focus of the project was the effects of the preprocessing of the LCMS files and how to optimize this step to predict the age of the fingerprints as precisely as possible.

The XCMS package in R was used for the preprocessing, and a manual tuning of several parameters was made. Here, the new version of XCMS was used and three different optimizations were made: one where the best values of some of the predictors were combined, one where the same predictors were optimized sequentially, and one where the predictors were optimized sequentially using the QC samples for the optimization steps. The new version of XCMS turned out to generally perform better than the old version, and all optimizations performed better than a base version of the new XCMS, which means an optimization of the preprocessing step makes sense. The sequential optimization was performing with the best result, while the combined optimization did not perform as well as expected given individual optimizations of the parameters. The QC optimization did not perform as well as the others, but it might be the most interesting result since the optimization still had a considerable effect and the risk of overfitting is as good as nonexistent. Thereby, the QC optimization might be the optimization to use and further study in the future.