NIELS BOHR INSTITUTE

UNIVERSITY OF
COPENHAGEN

PH.D. THESIS

by

Christian Michelsen

Biological Data Science

In ancient genomics, epidemiology,
anesthesiology, and a bit in between

Submitted: November 10, 2022

This thesis has been submitted to the
PhD School of The Faculty of Science,
University of Copenhagen

Supervisor	Troels C. Petersen	Niels Bohr Institute
Cosupervisor	Thorfinn S. Korneliussen	Globe Institute

Christian Michelsen, *Biological Data Science,* In ancient genomics, epidemiology, anesthesiology, and a bit in between, November 10, 2022.

Til kvinderne i mit liv

Contents

Preface i

Acknowledgements iii

Abstract v

Dansk Abstract vii

Publications ix

1 Notes on the design 1

1.1 Document layout 1

Preface

This Ph.D. thesis summarizes my scientific research in collaboration with the Niels Bohr Institute (NBI) and the Globe Institute, University of Copenhagen, and was funded by the Lundbeck Foundation. The research was supervised by Associate Professor Troels C. Petersen (NBI) and Assistant Professor Thorfinn S. Korneliussen (Globe Institute).

Being a multi-disciplinary project, the research presented in this thesis is multi-faceted and covers a wide range of topics with the main scope being the development and integration of a novel statistical methods and machine learning models for the analysis of large-scale biological data. The thesis is organized as follows. First I present a brief introduction to the statistical methods and machine learning models used in the thesis. Then I present the research in the form of four papers, each of which reflects a different aspect of the research.

The first paper presents a novel method I developed for detecting and classifying ancient DNA damage in metagenomic samples taking the full taxanomic information into account. While the first paper focuses on the development of the statistical model in the field of ancient genomics, the second paper focuses on the use of modern machine learning models in medicine and how advanced boosted decision trees can not only improve the accuracy of identifying patients at risk of being readmitted after knee or hip surgery, but doing so in a way that is interpretable as well. In the beginning of 2020 we all experienced how COVID-19 suddenly changed our lives and impacted our societies in dramatic ways. During this time, I ended up working for Statens Serum Institut (SSI), the Danish CDC, on a project to develop an agent based model capable of simulating the spread of COVID-19 in Denmark. This model is presented in the third paper and was used to inform the Danish government on how to best handle the pandemic in the early stages and the effect of lockdowns and other measures. Finally, in the fourth paper we show how advanced Bayesian methods can be utilized to better estimate the diffusion coefficients of molecules in the cell nucleus in XXX experiments.

Acknowledgements

First of all I want to express my sincere gratitude to my long time supervisor, captain, and friend: Troels. You are truly an inspiration to work with. I want to thank you for opening so many doors for me, both academically, professionally, and nautically. I am looking forward to our future adventures together. I also want to thank my co-supervisor, Thorfinn. I want to thank you for introducing me to the field of bioinformatic and helping me to develop my skills in this area. I also want to thank you for your patience and guiding me through the endless amount of (near) identical biological concepts and helping me to understand the minute differences.

I have been fortunate to work with people from a wide range of backgrounds and disciplines during my Ph.D. The author lists on the papers in this thesis include a particle physicist, bioinformaticians, a clinical professor, epidemiologists, a medical doctor, a bio-physicist, a mathematician, a biologist, and the president of the Royal Danish Academy of sciences and letters. Before anything else, I want to thank all of my co-authors for their work and contributions to these papers and for allowing me to be a part of their projects. I have learned a lot from all of you, and I hope that I have been able to contribute something to your work as well.

I am thankful for all the people who have helped me with my work and listened to my complaints when I was stuck, when the code did not compile, or when the small bug was almost impossible to find (which was not a small amount of time). In particular I want to thank the people at Globe who I have spent the most time with; Rasa, Alba, and Rasmus. I also want to thank the Korneliussen Group and the people in my office for helpful advice, suggestions and discussions. This also includes Daniel Nielsen and Rasmus Ørsøe from NBI. Finally, I want to thank Mathias Heltberg for many years of fruitful collaboration and for including me in his projects.

This project would not have been possible had it not been for the Lundbeck Foundation which funded my Ph.D. In addition to the funding itself, I am grateful for the inter-disciplinary aspect of project which has allowed me to meet so many inspiring and talented people and for the freedom to pursue my own interests within the project.

I would also like to express my gratitude to Professor Guido Sanguinetti from the International School for Advanced Studies, SISSA, in Trieste, Italy, for hosting me in his group during the Winter of 2021/2022. My gratitude also goes out to Kosio, Sara, Max, Romina, Noor, Viplove, Anne-Marie, and all the other wonderful people that I met during in Trieste. Thanks for making my stay in Italy so enjoyable and for welcoming me in a way that only non-Danes can do.

I want to thank my friends for always being there for me. A special thanks to my friends from NBI and Borchen who I know that I can always count on, whether or not that includes a trip in the party bus of the Sea, taking Artemis out for a sail, or board games and beer. Thank you for always being there. I also want to thank my family, especially my parents for their unconditional support and encouragement. I am grateful for the opportunities that they have given me and for the sacrifices that they have made for me.

Lastly, I want to thank my future wife and mother of our child, Anna. I would not have been able to do this without you. Thank you for your patience and support. I am looking forward to our future together. I love you to the moon and back.

Abstract

Basically a thesis (book?) class for Tufte lovers like myself. I am aware that `tufte-latex` already exists but I just wanted to create my own thing.

Dansk Abstract

Her et dansk abstract.

Publications

The work presented in this thesis is based on the following publications:

Christian S. Michelsen, Mikkel W. Pedersen, Antonio Fernandez-Guerra, Lei Zhao, Troels C. Petersen, Thorfinn S. Korneliussen (2022). "metaDMG: An Ancient DNA Damage Toolkit".

Christian Michelsen, Christoffer C. Jorgensen, Mathias Heltberg, Mogens H. Jensen, Alessandra Lucchetti, Pelle B. Petersen, Troels C. Petersen, Henrik Kehlet (2022). "Preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty - a machine learning based approach.".

Mathias S. Heltberg, Christian Michelsen, Emil S. Martiny, Lasse E. Christensen, Mogens H. Jensen, Tariq Halasa and Troels C. Petersen (2022). "Spatial Heterogeneity Affects Predictions from Early-Curve Fitting of Pandemic Outbreaks: A Case Study Using Population Data from Denmark". In: Royal Society Open Science 9.9. issn: 2054-5703. doi: 10.1098/rsos.220018.

Susmita Sridar, Mathias S. Heltberg, Christian S. 6 Michelsen Judith M. Hattab, Angela Taddei (2022). "Microscopic single molecule dynamics suggest underlying physical properties of the silencing foci".

1 *Notes on the design*

This class is my personal mix of different book design influences: mainly the works of Edward R. Tufte, (Heltberg et al., 2022; Korneliussen, Albrechtsen, and Nielsen, 2014) known for the big margin and the plentiness of sidenotes and sidecaptions. The margins are however not as prominent as in Tufte's works, the main text takes a bit more space, more like in Robert Bringhurt's typographer's bible (Heltberg et al., 2022).

So it is a bit of a mix of Tufte and Bringhurst, with some of my own choices for other design features, as we will see through this chapter.

1.1 *Document layout*

While `tufte-style-thesis` is a class for typesetting theses, the general layout is pretty much the same as in a regular book. A book is traditionally divided into three major sections: the front matter, the main matter and the back matter.

Bibliography

Heltberg, Mathias Spliid et al. (2022). "Spatial Heterogeneity Affects Predictions from Early-Curve Fitting of Pandemic Outbreaks: A Case Study Using Population Data from Denmark". In: *Royal Society Open Science* 9.9. ISSN: 2054-5703. DOI: 10.1098/rsos.220018.

Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen (2014). "ANGSD: Analysis of Next Generation Sequencing Data". In: *BMC Bioinformatics* 15.1, p. 356. ISSN: 1471-2105. DOI: 10.1186/s12859-014-0356-4. URL: https://doi.org/10.1186/s12859-014-0356-4 (visited on 2019).

This document was typeset using LaTeX and the `tufte-style-thesis` class.
The style is heavily inspired by the works of Edward R. Tufte and Robert Bringhurst.
This is available on here:

`https://github.com/sylvain-kern/tufte-style-thesis/`.

Feel free to contribute!