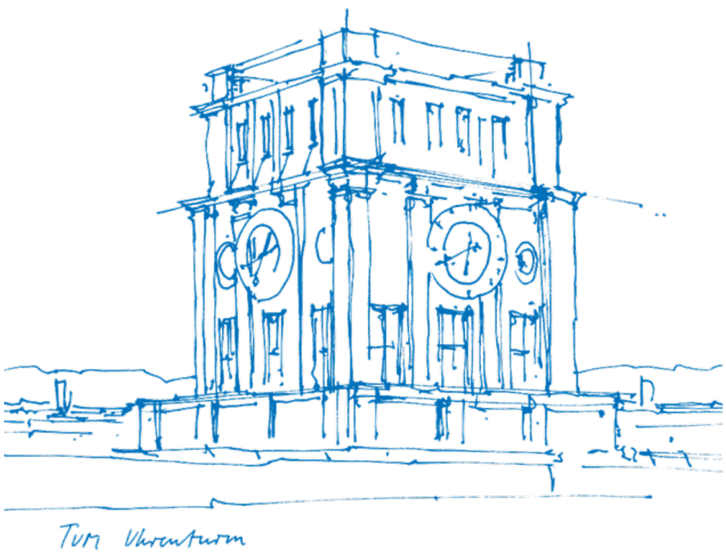
Please adjust group name in tumuser.sty.
Department of Electrical and Computer Engineering
Technical University of Munich

Thesis title

Subtitle of the thesis

Juan David Henao Sanchez



Thesis title

Subtitle of the thesis

Juan David Henao Sanchez

Bitte passen Sie den Lehrstuhlnamen in tumuser.sty an.
Fakultät für Elektrotechnik und Informationstechnik
Technische Universität München

Thesis title

Subtitle of the thesis

Juan David Henao Sanchez

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzende(r):
Prof. Franz X. Gabelsberger

Prüfer der Dissertation:
1. Prof. Dr. Georg Simon Ohm
2. Prof. James Clerk Maxwell

Die Dissertation wurde am 29.04.2016 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 11.07.2016 angenommen.

To Franz X. Gabelsberger, inventor of the street named after him.

Abstract

The abstract of your thesis goes here.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Contents

1 Introduction

This is the introduction of the thesis.

1.1 Pathophysiology of chronical lung diseases

1.1.1 Bronchopulmonary Dysplasia (BPD)

1.1.2 Asthma

1.1.3 Chronic Obstructive Pulmonary Disease (COPD)

1.1.4 Idiopathic Pulmonary Fibrosis (IPF)

1.2 Computational biology and chronic lung diseases

1.2.1 Multi-omics data integration

1.2.2 Clinical prediction

1.2.3 Systems biology

1.3 Aims

2 Methodology

This is the methodology of the thesis.

2.1 Data gathering

2.1.1 Mice data

Transcriptomics

2.1.2 Human data

Transcriptomics

Metabolomics

2.1.3 Public data

Multi-omics bulk data

Neonatal single-cell transcriptomics

2.2 Preprocessing

2.2.1 Normalization

DESeq2

Pareto scaling

Size-effect

2.2.2 Data imputation

Random-forest

knn-Imputation

2.2.3 Batch-effect detection

Principal component analysis (PCA)

Hierarchical clustering

K-BET

2.3 Differential expression analysis

2.3.1 Limma

2.3.2 DESeq2

2.4 Enrichment analysis

2.4.1 Gene list functional enrichment analysis

2.5 Multi-omics factor analysis (MOFA)

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

2.6 Clinical data correlation

2.6.1 Linear regression

2.6.2 Binomial regression

2.6.3 Ordinal regression

2.6.4 Multinomial logistic regression

2.6.5 Dirichlet regression

2.7 Benchmarking of Lasso models dealing with missing values

2.7.1 Knowledge guided multi-level network inference

2.7.2 Two-steps based models

Grouped adaptive Lasso (GALasso)

Stacked adaptive Lasso (SALasso)

2.7.3 Inverse covariance based methods

Convexed conditioned Lasso (CoCoLasso)

Lasso with high missing rate (HMLasso)

2.8 Adult data correlation

2.8.1 Random forest

Imbalanced random forest

Nested cross-validation in random forest

2.8.2 t-test, manova, log-reg

3 Summary of publications

1. Juan Henao, Alida Kindt, Tanja Seegmüller, Kai Foerster, Andreas Flemmer3, Juergen Behr4, Nikolaus Kneidinger, Marion Frankenberger, Fabian Theis, Benjamin Schubert, Markus List, Anne Hilgendorff. Multi-omic signatures relate to the severity of pulmonary outcome in neonates traced into adult disease.

   Summary: This project focused on the detection of endotypes behind Bronchopulmonary Dysplasia (BPD) by proteomics, metabolomics, and clinical data integration using a cohort of 55 neonates with and without BPD. The endotypes were detected using Multi-Omics Latent Factor Analysis (MOFA) REF with sensitivity selection. We caught seven latent factors. However, none showed a sign of endotyping discrimination given the combined distribution of latent scores between no BPD and BPD patients in each latent factor. Nevertheless, the biological interpretation of each latent factor allowed us to discover a persistent inflammatory disease component in BPD.

   We expanded our analysis by looking for individual molecular features with the potential to be biomarkers of severity using ANOVA with a t-test as a post-hoc method comparing no BPD, mild BPD, and moderate/severe BPD. Acknowledging the clinical heterogeneity signal of BPD cases, we reclassified them into no or moderate/severe BPD using a random forest model trained using oxygen supplementation and mechanical ventilation days (clinical variables used to diagnose BPD). We applied a t-test to identify significant molecular features between no BPD and moderate/severe BPD. We complement our analysis by training different random forest models combining significant molecular features and sets of increasing BPD characterization:

   a) **BPD descriptors:** Oxygen supplementation and mechanical ventilation days.
   b) **Main risk variables:** Gestational age and birth weight.
   c) **Deep clinical phenotyping:** *Main risk variables* and a compendium of clinical measurements encompassing comorbidities, medical interventions, and previously defined MRI-based scores.

   The metabolite PC(O-36:5) was detected in both significant analyses and, combined with deep clinical phenotyping, improves the BPD classification along with PC(O-44:5) and gestational age. The protein CCL22 was detected in both significant analyses and improved the BPD classification according to random forest when combined with the main risk variables. Besides, SCGF-alpha, SCGF-beta, and KIR3DL2 were significantly different by ANOVA analysis of no, mild, and moderate/severe BPD comparison.

   We traced our significant proteins in an adult chronic lung disease cohort composed of Chronic Obstructive Pulmonary Disease (COPD), Idiopathic Pulmonary Fibrosis (IPF), and healthy donors by ANOVA analysis comparing the three conditions. CCL22 and KIR3DL2 were detected in COPD, while SCGF-beta was significant in COPD and IPF. Those results support the hypothesis regarding the susceptibility of neonates with a BPD diagnosis to develop chronic lung diseases in adulthood.

   Contribution: I performed the data pre-processing and all the analyses used in this project. Besides, I created all the data visualization and wrote the first draft of the paper, which was reviewed and edited by Anne Hilgendorff, Markus List, and Tanja Segmuller.

2. Erika Gonzalez Rodriguez1, Juan Henao2, Motaharehsadat Heydarian1, Tina Pritzke1, Alida Kindt3, Anna M. Dmitrieva1, Heiko Adler4, 5, Melanie Markmann6, Valeria Viteri-Alvarez1, Prajakta Oak1,

Markus Koschlig1, Xin Zhang1, Kai M. Foerster7, Andreas Flemmer7, Hamid Hossain6,8, Xavier Pastor2, Holger Kirsten9, Peter Ahnert9, Juergen Behr10, Tushar J. Desai11, Benjamin Schubert2, Anne Hilgendorff1,12. Hyperoxia-induced cell cycle arrest drives long-term impairment of lung development and DNA repair in neonates.

3. Juan David Henao Sanchez3,14, Mustafa Abdo1,2, MD, MSc, Benjamin Schubert3,14, PhD, Markus List4, PhD, Henrik Watz2,14, MD, Frauke Pedersen1,2,14, PhD, Alina Bauer3,15, MSc, Dominik Thiele5,14, MSc, Adam M. Chaker6,7, MD, Constanze A. Jakwerth 7,15, PhD, Benjamin Waschki1,8,14, MD, Anne Kirsten2,14, MD, Markus Weckmann9,14, PhD, Oliver Fuchs9,10,14, MD, PhD, Gesine Hansen11,16, MD, Matthias V. Kopp9,14, MD, Erika v. Mutius12,13,15, MD, MSc, Inke R. König4,14, PhD, Klaus F. Rabe1,14, MD, PhD, Thomas Bahmer1,14, MD, Carsten B. Schmidt-Weber7,15, PhD, Ulrich M. Zissler7,15, PhD, and the ALLIANCE Study Group*. Cytokines Derived from Nasal Epithelial Lining Fluid in Patients with Asthma.

4. Henao, J. D., Lauber, M., Azevedo, M., Grekova, A., Theis, F., List, M., ... & Schubert, B. (2023). Multi-omics regulatory network inference in the presence of missing data. Briefings in Bioinformatics, 24(5), bbad309.

   Summary: This project is our contribution to solving one of the most common issues in biological network inferences: the presence of missing data. Here, we extended the previous R package developed by Christoph Ogris, KiMONo (Knowledge guided Multi-Omics Network inference) REF, which uses sparse-group lasso (SGLasso) REF to establish multi-omics edges between molecular features based on a prior network. KiMONo was extended by benchmarking L1-regularized model extensions (sparse models) dealing with missing data and which implementation and code source were performed in R language. We tested SGLasso with imputed data by kNN-imputation (knnSGLasso), two methods based on inverse covariance matrix (CoCoLasso and HMLasso) REF, and four methods using multi-imputed data (S(A)Lasso and G(A)LAsso) REF. The connection between independent and dependent variables (edge in the network) was established if the beta coefficient was non-zero and the r-squared was larger than 0.1.

   We evaluated those sparse models by comparing performance using the same multi-omics information: transcriptomics, CNVs, and methylation. We repeated the comparison using three different datasets extracted from TCGA: Breast invasive carcinoma (TCGA-BRCA), muscle-invasive bladder cancer (TGCA-MIBC), and prostate adenocarcinoma (TGCA-PRAD). We used two prior networks, one extracted from BioGrid for TCGA-BRCA (as the original KiMONo's paper) and one created from FunCoup V5 to evaluate TCGA-MIBC and TCGA-PRAD.

   We simulate the most common missing data scenarios, single-omics (transcriptomics), and multi-omics random missingness from 0 to 50%, increasing by 10%. Besides, we added Gaussian noise by adding 0, 0.5, and 1.5 times the standard deviation to the normal distribution. We also evaluated block-missingness cases, i.e., when samples did not match between omics measurements by random experiment removal from 0 to 50%, increasing by 10%. We repeated the experimental setup five times, changing the random seed. We compared the number of nodes, transitivity, the median of R-squared, and the F1 score calculated in two ways, using the whole data (no missingness) as a reference and a network inferred using stability selection fitting 100 random seeds. In addition, we compared the run time per sparse method, and we reached the ability of each method to detect expression quantitative trait methylation (eQTM) regarding a state-of-the-art method, Matrix eQTL.

   In general, for small datasets (TCGA-BRCA), the sparse models based on the inverse covariance matrix (CoCoLasso and HMLasso) failed to infer networks even at a minimal missing ratio (10%). However, with larger datasets (TCGA-MIBC and TCGA-PRAD), they could outperform the methods based on multiple imputations (S(A)Lasso and G(A)Lasso) and are computationally more efficient. However, SLasso and knnSGLasso tend to perform well, independent of the size of the dataset. The matrix eQTL and sparse models' performance were quite similar. Nevertheless, the sparse models detected marginally more eQTM-linked genes than Matrix eQTL.

Contribution: I created the benchmark framework to automatize the missing data simulations and run the different methods simultaneously. Manuel Acevedo and Michael Laube implemented the inverse covariance matrix-based methods (CoCoLasso and HMLasso), Anastasiia Grekova implemented the knnSGLasso method, and I implemented the multiple imputation-based methods (S(A)Lasso and G(A)Lasso). I joined all the results and conducted the model performance evaluation, run time comparison, and eQTM analysis. I wrote the manuscript draft along with Benjamin Schubert, Christoph Ogris, and Markus List. I created the result visualization with the help of Christoph Ogris.

4 Discussion

This is the discussion of the thesis.

A Appendix

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.