Karlsruhe Institute of Technology

Integration of internal and external gene expression and drug-perturbation data to empower novel immune therapies against Parkinson's Disease

Master Thesis of

Rudolf Biczok

At the Department of Informatics
Institute of Theoretical Computer Science

and in cooperation with

Roche Pharma Research and Early Development
Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd

Reviewer:	Prof. Dr. Alexandros Stamatakis
Second Reviewer:	Prof. Dr. Ralf Reussner
External Advisor:	Dr. Jitao David Zhang

Time Period: 1st August 2018 – 31st January 2019

Statement of Authorship

I hereby declare that this document has been composed by myself and describes my own work, unless otherwise acknowledged in the text.

Heidelberg, 31st January 2019

Abstract

The primary objective of gene set enrichment analysis is to annotate a genes of interest with a-priory knowledge in the form of curated gene sets. The problem in this method lies in the large number of reported gene sets and their varying information content. Previous publications suggest to use unsupervised learning methods like hierarchical clustering or self-organized maps to increase interpretability, but there is no metric to assess the quality of these clustering methods or the difference between gene set itself. We therefore evaluated statistical methods (minkowski, jaccard, kappa-statistic), network-based methods (shortest path in protein-protein networks), and method based on natural language processing for their capability to measure a biologically plausible distance between gene sets. We used pathway trees from Reactome and a curated tree of immune cell types with corresponding gene sets to compare these distance methods.

TODO what i the conclusion

Deutsche Zusammenfassung

Make german summary at t very end

Acknowledgements

First and foremost I want to thank the two most valuable point of information Prof Dr. Alexandros Stamatakis and Dr. Jitao David Zhang. The open minded nature of Prof. Stamatakis made this research collaboration possible and his leading expertise in computational bioinformatics tremendously helped us to keep this master thesis in a academic format. Dr. Zhang also proved his courage and passion in academia by entrusting a theoretical biology research project to an computer science student. His knowledge as principle bioinformatician / biostatistician in an industrial and academic research environment complemented the methodical / computer science expertise of Stamatakis and me beyond expectations.

I send my greetings and thankfulness to Prof. Dr. Ralf H. Reussner, who did not hesitate to take the responsibility as second reviewer. I was a former participant in a two-term research project under the supervision of Prof. Reussner where he demonstrated a high level of methodical knowledge in the engineering aspect of computer science.

In addition, I want to highlight the support from Gregor Sturm, Sarah Lutteropp, and Lucas Czech. Gregor Sturm is a former master thesis student of Dr. Zhang who shared the curated tree of immune cell types with their respective marker genes. He also eagerly helped me to extract all necessary information from his publicly available code repository to save time on my side. Sarah Lutteropp is a PhD student of Prof. Stamatakis and shared her knowledge about methods and limitations in distance-based (phylogenetic) tree inference algorithms. Lucas Czech is also a PhD student under supervision of Prof. Stamatakis and provided me with an implementation skeleton for creating unsupervised clustering algorithms similar to k-means in C++.

Although I wish to thank every person in my live who inspired me, helped me, or even influenced my belief system, I must restrict myself to the following group of people that deserve a special place in this section: All members of the HITS Exelixis Lab (Prof. Dr. Alexandros Stamatakis, Dr. Alexey Kozlov, Lucas Czech, Sarah Lutteropp, Pierre Barbera, Benoit Morel, and Ben Bettisworth) and ROCHE BEDA group. Every single member treated me like an equal researcher.

I send my finally thanks and greetings to my parents, who are the only person on earth able to restrain my evil mind and my sister, who happened to be the younger sibling and by extension forced me to be a good role model.

ll lucas be a
with the end
this thesis?

Contents

1. Motivation **1**
1.1. Own contribution . . . 2

2. Introduction **3**
2.1. Gene Expression Analysis . . . 3
2.1.1. Gene set enrichment . . . 3
2.2. Protein-protein interactions . . . 3
2.3. Gene ontology . . . 3
2.4. Natural language processing . . . 3

3. Materials and methods **5**
3.1. Reference data . . . 5
3.1.1. Reactome reference tree . . . 5
3.1.2. Immune cell differentiation hierarchy . . . 5
3.2. Distance measurements . . . 5
3.2.1. Mathematical measurements . . . 5
3.2.2. Network-based measurements . . . 5
3.2.3. Word2vec models & measurements . . . 5

4. Results & Discussion **7**
4.1. Ground-truth comparison . . . 7
4.2. Limitations . . . 7

5. Conclusion **9**

6. Appendix **11**
A. ROGER - Roche Omnibus of Gene Expression Regulation . . . 11
A.1. State of the art . . . 11
A.1.1. Transcriptomic data management . . . 11
A.1.2. Differential Gene Expression Analysis . . . 11
A.1.3. Gene Set Enrichment Analysis . . . 11
A.2. Reimplementation . . . 11
A.2.1. Data structures & architecture . . . 11
A.2.2. Visualizations & data access . . . 11
B. List of Acronyms . . . 11

Bibliography **13**

Todo list

TODO what is the conclusion? v
Make german summary at the very end v
Will lucas be a Dr with the end of this thesis? vi
update to newest 2
Ref 2
List of figures? 11
List of tables? 11

1. Motivation

Molecular biology is the aspect of life science thats investigates biological processes on a cellular and molecular level. Biologists in this area seek answers for questions like: "What is the structural and functional difference between neuron cells compared to other cell types in mammal species?", "What influence has chemical compound A when introduced to cell line C?", or "Is the cell line derived by following lab protocol A different from the cell line of protocol B?". The common procedure to research these questions is to conduct wet lab experiments on prepared cell cultures followed by a computer-assistant gene expression analysis. Gene expression is the fundamental biological process of every organism that describes the transcription of Ribonucleic Acid (RNA) from Deoxyribonucleic Acid (DNA) and the translation from RNA to proteins [AJL+02]. Collecting and analyzing the gene expression level of every gene inside an organism allows us to identify differentially expressed genes that cause morphological differences between cell groups or cell types [RMS10]. To the end, bioinformaticians use public databases of gene sets to see which known cell components or biological processes are reflected by the previously inferred list of differential expressed genes. Every gene set represents discovered knowledge in form of name, description and involved genes of a particular biological process. Ideally, the entire procedure results in a list of gene sets that uniquely explain the effects of the original wet lab experiment [WS12] (see section 2.1 for further information about gene expression analysis).

In reality, however, the information gain from reported gene sets is unsatisfying, because 1) gene sets from even the same database source tend to have a high gene overlap, 2) gene sets from publicly available databases can have many genes (>200), and 3) gene set information like title and description can vary in quality depending on the source. Existing literature suggest supervised clustering methods to organize gene sets into a more representative structure. DAVID clustering, for instance, performs agglomerative clustering over pairwise kappa statistic between gene sets [HST+07]. The authors of this algorithm claim that it maximizes the number of pairwise Protein Protein Interactions (PPI) within each gene set cluster. They also, however, state that it is unclear if this optimization criterion is biologically justified. In general, there exist no gold standard to asses the biological similarity between two gene sets. Having such a gold standard for comparing gene sets on the other hand would make it possible to compare or even refine different clustering algorithm.

1.1. Own contribution

We present in this thesis a systematic evaluation of different categories of distance metrics for gene sets. We implemented metrics based on 1) statistic methods, 2) gene ontology trees, 3) protein-protein interaction networks, and 4) natural language processing methods. For comparing the performance of each distance implementation, we extracted gene sets from data sources whose relationships are already know and preserved as rooted trees. These data sources include Xsubsets of the Reactome pathway database [JTGV+05] and a manually curated collection of marker genes for human immune cell types. We bundled all distances, data preprocessing and analysis scripts into a automated work flow that can be readily extended or included into other algorithms.

Besides gene set analysis, we also spend a significant amount of time building a client-server application with a web-based front-end for executing and storing gene expression analysis experiments. .

2. Introduction

2.1. Gene Expression Analysis

2.1.1. Gene set enrichment

Even with the list of significant genes from a gene expression analysis, it is hard to impossible identifying the exact biological event that is the causal reason for the expression of these genes. The reason for this is the share amount of genes in higher species (e.g. homo sapiens has 24,000 coding genes) whose actual functions is either unknown or highly ambiqous.

First, a biologist conducts experiments on prepared cell lines and then extracts RNA from these cells. Second, a bioinformatician utilize gene expression analysis pipelines to detect a list of statistically significant gens within expression profiles. Finally, bioinformaticians use public databases of gene sets to see which cell components or biological processes are reflected by the previously inferred list of significant genes.

2.2. Protein-protein interactions

2.3. Gene ontology

2.4. Natural language processing

3. Materials and methods

3.1. Reference data

3.1.1. Reactome reference tree

3.1.2. Immune cell differentiation hierarchy

3.2. Distance measurements

3.2.1. Mathematical measurements

3.2.2. Network-based measurements

3.2.3. Word2vec models & measurements

4. Results & Discussion

4.1. Ground-truth comparison

4.2. Limitations

5. Conclusion

6. Appendix

A. ROGER - Roche Omnibus of Gene Expression Regulation

A.1. State of the art

A.1.1. Transcriptomic data management

A.1.2. Differential Gene Expression Analysis

A.1.3. Gene Set Enrichment Analysis

A.2. Reimplementation

A.2.1. Data structures & architecture

A.2.2. Visualizations & data access

B. List of Acronyms

DNA Deoxyribonucleic Acid. 1

PPI Protein Protein Interactions. 1

RNA Ribonucleic Acid. 1

List of figures?

List of tables?

Bibliography

[AJL+02] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular Biology of the Cell, Fourth Edition*. Garland Science, 4 edition, 2002.

[HST+07] Da Wei Huang, Brad T. Sherman, Qina Tan, Jack R. Collins, W. Gregory Alvord, Jean Roayaei, Robert Stephens, Michael W. Baseler, H. Clifford Lane, and Richard A. Lempicki. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biology*, 8(9):R183, Sep 2007.

[JTGV+05] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G.R. Gopinath, G.R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(suppl_1):D428–D432, 2005.

[RMS10] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[WS12] Di Wu and Gordon K. Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133, 2012.