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 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 sets itself. We therefore evaluated statistical methods (minkowski, jaccard, kappa-statistic), tree-based methods (gene ontology), 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 benchmark 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 an 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.

ll lucas be a
with the end
this thesis?

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.

Contents

Todo list

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 [WS12a] (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 learning methods to organize gene sets into a more representative structure. The DAVID algorithm, 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. However, they also 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 the other hand would make it possible to compare gene set as elements of a metric space. It would enable researchers to benchmark and refine clustering algorithms or to discover new insights from the rapidly growing amount of gene set data.

1.1. Own contribution

We present in this thesis a systematic evaluation of different distance metrics for pairwise gene set comparisons. We implemented metrics based on 1) statistic methods, 2) gene ontology trees, 3) protein-protein interaction graph 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 2subsets of the Reactome pathway database [JTGV+05] and a manually curated collection of marker genes for 36human immune cell types [SFP+18]. We bundled all distances implementations, data preprocessing, and analysis scripts into a Python package that can be readily extended or included into other algorithms.

date to west

date to west

Besides gene set analysis, we also spend a significant amount of time building a gen expression analysis framework based on Python. It utilizes a client-server architecture with3 different front-ends for executing and persisting new gene expression experiments (see appendix A for more information).

1.2. Structure of this thesis

The reminder of this document follows the structure of a conventional bioinformatics paper. In chapter 2, we will explain the anatomy of a gene expression analysis pipeline in greater details. In addition, the introduction chapter will cover basic concepts about natural language processing with Word2Vec models and the composition about biological data sources (e.g. gene ontologies, PPI networks).

Chapter 3 gives more details about the algorithms behind each implemented distance metric. We will also summarize the preprocessing steps for the used evaluation data in this chapter.

In chapter 4, present benchmark results from each distance metric over the different sets of evaluation date. We will also discuss certain outcome of the results and assess the limitations of different distance metrics.

ention the key ints here

And finally, we will draw a final conclusion about our conducted experiments and we will give a future outlook about potential follow-up research in chapter 4.

2. Introduction

2.1. Gene expression analysis

Gene set enrichment is the bread and butter of every bioinformatician who tries to discover the genetic reason for morphological differences between cell lines. The basic process involves a series of wet and dry lab operations (fig. 2.1).

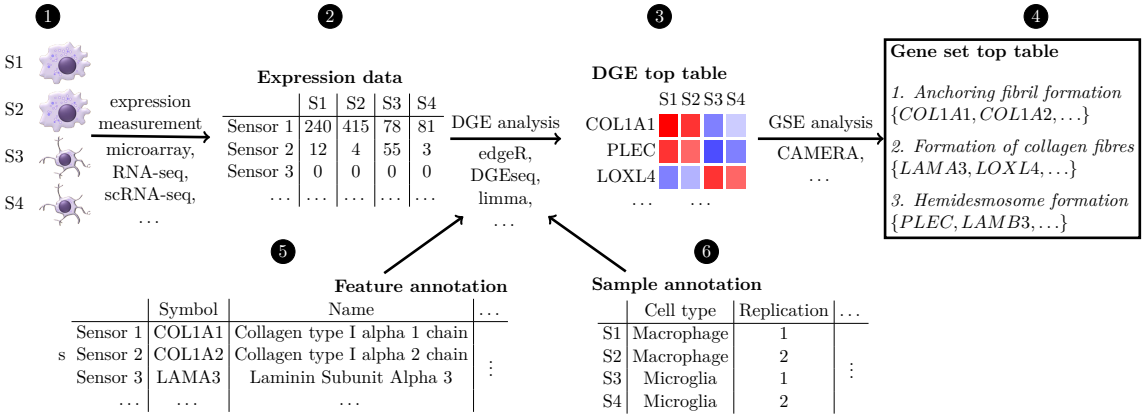

Figure 2.1.: Gene expression analysis work flow

At first (1), a biologist prepares at least two groups of cell lines (generally called samples). The grouping depends on the desired comparison a researcher wants to study, like different immune cell types (e.g. macrophages vs microglia cells [GLH+13]), healthy cells vs. tumor cells, or perturbed vs. non-perturbed cells. By perturbation we mean any type of cell modification (gene knock-out) or manipulation of the cell environment (e.g. adding drug compounds). It is common practice to cultivate more than one cell line under the same experimental condition (aka. technical replicates) to ensure reproducibility.

The next task describes the generation of gene expression profiles (2). This involves fixing the cells, dissolving their membrane, and extracting all RNA fragments. Different methods exist to quantify the RNA concentration per gene and by extension the expression levels. The most frequently used methods are microarray assays, RNA sequencing (RNA-Seq) [AJL+02], and signle cell RNA sequencing (scRNA-Seq) [ESBK13]. Each method requires different laboratory tasks and computational preprocessing algorithm. The end result of this stage is a table that shows the expression levels for every gene in every sample.

The actual meaning of the value can differ depending on the used preprocessing technique. In can, for instance, stand for the total number of RNA fragments counted per gene when using RNA-Seq.

After obtaining the raw data, a bioinformatician use linear models to detect genes that are differentially expressed between sample groups ❸ . For instance, a widely used software package called edgeR models the entire gene expression experiment as negative binomial distribution to control gene-wise dispersion [RMS10]. The package edgeR uses a variant of the Fisher's exact test to detect Differential gene expression (DGE). Other packages like limma [RPW+15] or DGEseq [WFW+10] are flavored depending on the number of technical replicates or used expression measurement technique.

Experience tells us that genes reported from DGE analysis alone give to view information behind the actual biological phenomena. A single gene can be involved in multiple, partially unknown biological processes or is just an artifact from prior DGE analysis. Gene set enrichment (GSE) algorithms try to overcome this issue by using statistical methods and external information about biological processes. One prominent example is implemented in the software package CAMERA [WS12b], which uses competitive gene set tests. The idea behind this competitive tests is to compare every genes inside a gene set relatively to all other genes measured in the experiment. CAMERA in particular uses a modified two-sided t-tests capable of detecting inter-gene correlations. The result of this step is a high score ❹ of gene sets reflected by the prior list of differentially expressed genes. The content of gene sets used for the GSE is arbitrary and the depends on what sources the bioinformatician choses for analysis. Gene sets from MSigDB [LSP+11], for instance, contains genes that characterize a specific cell type or cell condition. The database Reactome [JTGV+05] on the other hand offers gene sets containing these genes that are part of a particular biological process (aka. pathway).

It is important to note the entire process relies on the existence of feature annotation data ❺ and sample annotation data ❻ . The feature annotation gives additional information about the measured genes (e.g. name and description of the actual gene that is associated with a particular sensor slot). The sample annotation hold information about origin, and preparation steps for each analyzed cell sample (e.g. cell type, tissue origin, used chemicals).

2.2. Pathway & protein-protein interaction networks

We can describe every reaction inside or outside a cell as network of protein interactions with organic chemicals, RNA, DNA, or other proteins. Figure 2.2 illustrates an excerpt of such an interaction network during the mitotic cell cycle.

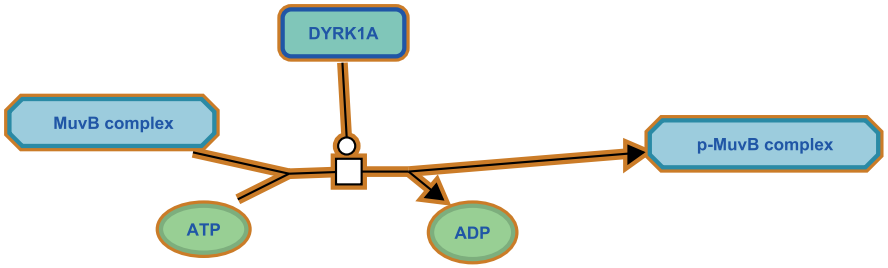

Figure 2.2.: Expert of the mitotic cell cycle. The rectangular boxes and boxes with octagon shape represent proteins. The green nodes represent other organic compounds. The entire pathway involves over different 400 proteins

Public databases like BioGRID [CaOB+17] offer a collection of known protein-protein interaction as annotated edge-list. Each edge represents the interaction between one protein with another and hold information about author, detection method, and interaction type. We can use these data sources to compare gene sets in a graph-based representation. It is important to note that not all protein-protein interactions are necessarily part of an actual biological process. This applies for these protein-protein interactions that researchers discovered outside a cell through regular chemical reaction assay. We use the term pathway to distinguish comprehensive protein-protein interaction networks provided by BioGRID with networks that are know to exist in living cells.

reference for BioGRID

2.3. Gene ontology

The Gene Ontology (GO) maintained by the gene ontology consortium [ABB+00, The17] is a collection fo controlled vocabulary (aka. GO terms). Every term has is unique identifier (e.g. GO:0030234) and is associated in one of three categories: 1) biological process (e.g. wound healing, epithelial cell proliferation), 2) cellular component (cytoplasm, organelle part), and 3) molecular function (e.g. catalytic activity, enzyme regulator activity).

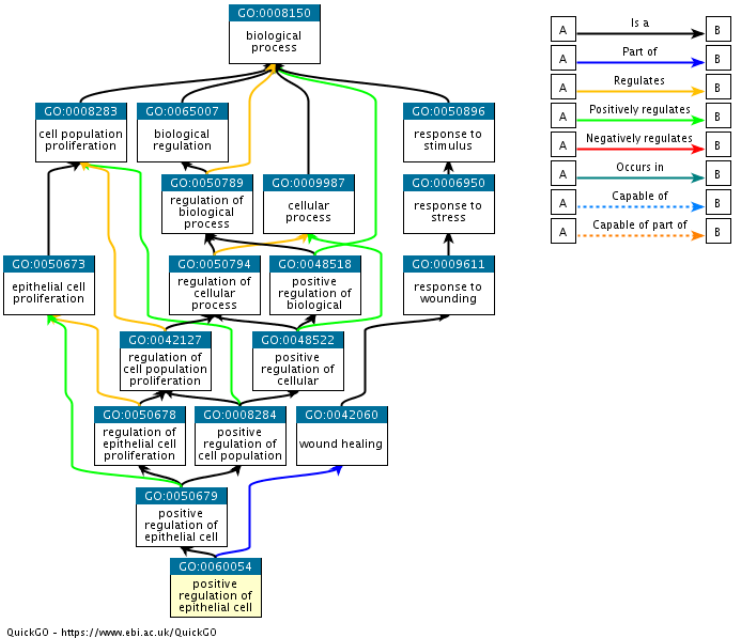

Figure 2.3.: Excerpt of the gene ontology

GO terms can have 8 different type of relationships between each other, as seen in fig. 2.3. If we only consider the “is a” relationships, all GO term of the same category resemble a rooted tree structure. In addition, the gene ontology consortium maintains a mapping between fig. 2.3 terms and gene identifier, where each fig. 2.3 term is associated with multiple gene identifiers.

Researcher can use the gene ontology for GSE analysis, where the reported “gene sets” are GO terms [ARL06]. Additionally, algorithms exist to define an arithmetic similarity measurement for GO terms [YLQ+10].

2.4. Natural language processing

According to Pyysalo et al. [PGM+13], publication databases like the PubMed Central (PMC) contain over 700,000 full-text articles with valuable information about biological

processes. Natural Language Processing (NLP) techniques like word embeddings opens the opportunity to detect linguistic relationships within unstructured text as it is found in literature databases.

2.4.1. Word embedding

Word embeddings are a class of machine learning models that learn feature vectors from a large set of arbitrary text (aka. text corpus). These feature vectors can be used to project words in a vector space while preserving their semantic similarities. For instance, assume two genes which are part of the same pathway. Then it is likely that the name of the genes appear relatively close to each other in multiple publication. Word embeddings reflect the text distances between the gene names to the feature vector space. Mature examples of world embedding are word2vec techniques, which use feedforward neural as supervised learning method networks [MCCD13].

2.4.2. Document queries

Inside a vector space, we can use mathematical functions like the euclidean distance or the cosine distance to measure the similarity between single word. However, gene sets consists of multiple genes and can carry additional information like a description or GO terms. One method to compare two sets of words (aka. documents) with each other is word averaging, where the pairwise sum or average of all word vectors build a representation of the document. Kusner et al. claims, however, that ignoring individual word distances could poor comparability between documents that have few words in commin [KSKW15]. Other solutions to this problem would be the use of alternative vector representations (e.g. Bag Of Words (BOW)) or more sophisticated distances based on the conventional word vectors (e.g. Word Mover Distance (WMD)) [KSKW15].

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

Shortest path problem: Not Robust (show example) * k-shortest path better, but also not that Robust * Robust shortest path problem is NP-complete and therefore prohibited

Variances: Average over pairwise node paths violates d(x,y) = 0 <=> x=y condition Mean over Min of

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

C. List of figures

D. List of tables

Bibliography

[ABB+00] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, Midori A Harris, David P Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C Matese, Joel E Richardson, Martin Ringwald, Gerald M Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25, may 2000.

[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.

[ARL06] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating go graph structure. *Bioinformatics*, 22(13):1600–1607, 2006.

[CaOB+17] Andrew Chatr-aryamontri, Rose Oughtred, Lorrie Boucher, Jennifer Rust, Christie Chang, Nadine K. Kolas, Lara O'Donnell, Sara Oster, Chandra Theesfeld, Adnane Sellam, Chris Stark, Bobby-Joe Breitkreutz, Kara Dolinski, and Mike Tyers. The biogrid interaction database: 2017 update. *Nucleic Acids Research*, 45(D1):D369–D379, 2017.

[ESBK13] James Eberwine, Jai-Yoon Sul, Tamas Bartfai, and Junhyong Kim. The promise of single-cell sequencing. *Nature Methods*, 11:25, dec 2013.

[GLH+13] Florent Ginhoux, Shawn Lim, Guillaume Hoeffel, Donovan Low, and Tara Huber. Origin and differentiation of microglia. *Frontiers in Cellular Neuroscience*, 7:45, 2013.

[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.

[KSKW15] Matt J. Kusner, Yu Sun, Nicholas I. Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 957–966. JMLR.org, 2015.

[LSP+11] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[MCCD13] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[PGM+13] S. Pyysalo, F. Ginter, H. Moen, T. Salakoski, and S. Ananiadou. Distributional semantics resources for biomedical text processing. In *Proceedings of LBM 2013*, pages 39–44, 2013.

[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.

[RPW+15] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, 2015.

[SFP+18] Gregor Sturm, Francesca Finotello, Florent Petitprez, Jitao David Zhang, Jan Baumbach, Wolf H Fridman, Markus List, and Tatsiana Aneichyk. Comprehensive evaluation of cell-type quantification methods for immuno-oncology. *bioRxiv*, 2018.

[The17] The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, 2017.

[WFW+10] Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang. Degseq: an r package for identifying differentially expressed genes from rna-seq data. *Bioinformatics*, 26(1):136–138, 2010.

[WS12a] Di Wu and Gordon K. Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):e133, 2012.

[WS12b] Di Wu and Gordon K. Smyth. Camera: A competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Research*, 40(17):1–12, 2012.

[YLQ+10] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, 26(7):976–978, 2010.