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. We found out that document queries with Word Mover's distance on word2vec embedding models yield the best results.

refine based o additional dist

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. Moreover, Dr. Zhang 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. Moreover, he 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

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 phenotypic 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 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 B 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 metric. We will also summarize the preprocessing steps for the used evaluation data in this chapter.

In chapter 4, present benchmark results from each metric over the different sets of evaluation date. We will also discuss certain outcome of the results and assess the limitations of different 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 phenotypic differences between cell lines. The basic process involves a series of wet and dry lab operations (fig. 2.1).

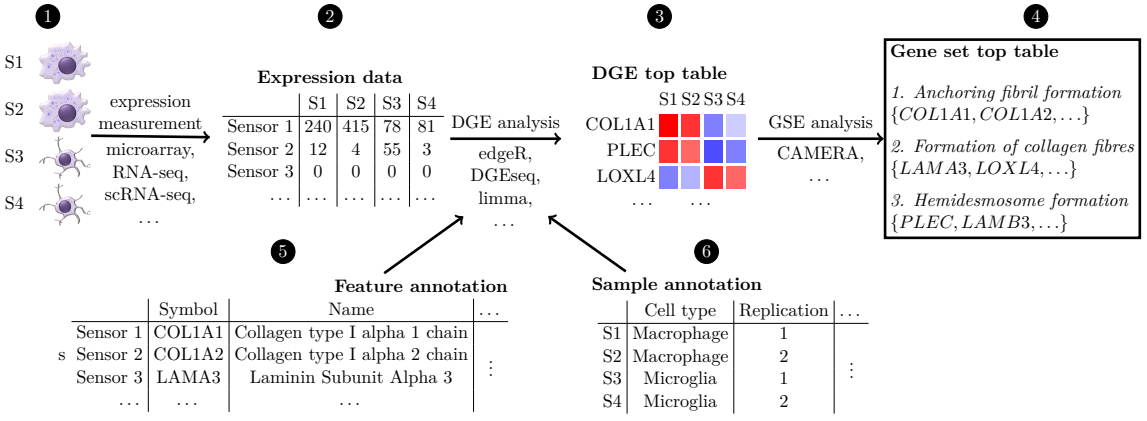


Figure 2.1.: Gene expression analysis work flow

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.

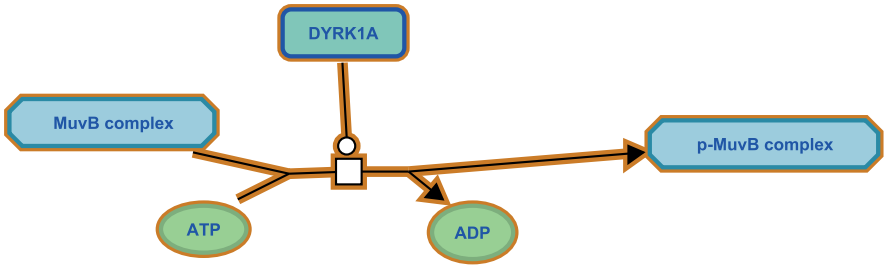


Figure 2.2.: Expert of the mitotic cell cycle. The rectangular boxes and boxes with octagon shape represent proteins. The green nodes represent other organic compounds. The entire pathway involves over different 400 proteins

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 known to exist in living cells [CaOB+17].

2.3. Gene ontology

The Gene Ontology (GO) maintained by the gene ontology consortium [ABB+00, The17a] 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).

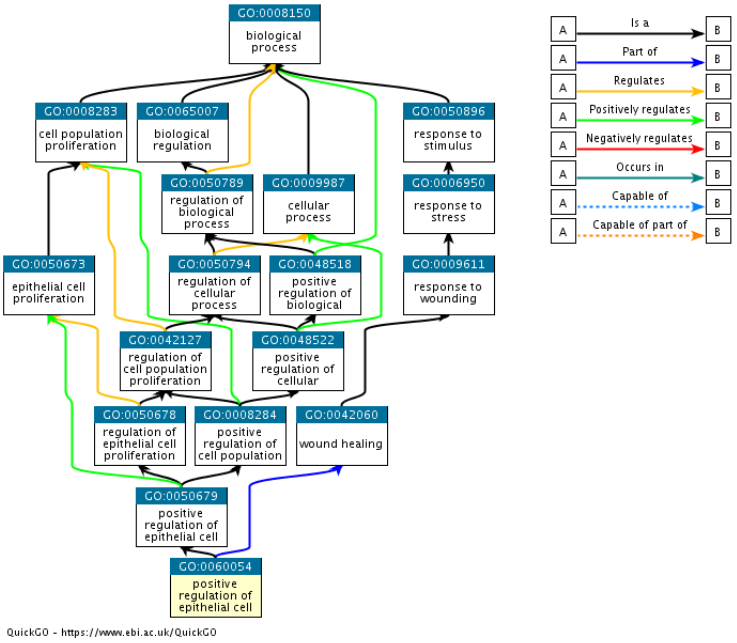


Figure 2.3.: Excerpt of the gene ontology

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. The gene ontology consortium maintains a mapping between fig. 2.3 terms and gene identifiers, where each fig. 2.3 term is associated with multiple gene identifiers. In addition, every term carries information origin (as standard evidence code) and function (as unstructured text).

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. Phenotype traits

Experience shows that the number of observable phenotypes can become unmaintainable depending on the genetic variability of an organism. The European Bioinformatics Institute

(EMBL-EBI) provide a controlled vocabulary of traits to characterize phenotypes. This vocabulary is called Experimental Factor Ontology (EFO) and contains clinical and non-clinical traits. A clinical trait represents either a disease (e.g. diabetes type 1) or a disease marker (e.g. measurements of blood glucose concentration). Non-clinical traits describe morphological properties that are not disease-related (e.g. exe color). Researchers use the standardized phenotype traits to associate diseases or other observable properties with their causing genes. For instance, the Genome-wide association studies (GWAS) database provides a catalog of manually curated mapping between genes and traits based on findings from over 3000 conducted studies [MBC+17]. This allows us to compare genes and gene sets based on their prototypical effects in the organisms.

d excerpt of
VAS

2.5. Natural language processing

According to Wilbur et al. [KFWL17], publication databases like PubMed contain over 25 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.

2.5.1. 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].

2.5.2. 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 a list of 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

Our approach to investigate the similarity between arbitrary gene sets consists of two steps. First, we extract reference trees and gene sets from manually curated data sources. Second, we implement metrics to compare individual gene sets with each other.

3.1. Reference data

All used reference trees and gene sets originate from the Reactome database [JTGV+05] and a study conducted by Sturm et al. [SFP+18]. The reference organism in both data sources is the human species.

Evey gene set from Reactome and Sturm et al. contains at least a name and one or more genes. In our reference data, the gene set name has a semantic meaning and describes either the biological process (Reactome) or the cell type (Sturm et al.) reflected by the containing genes. In general, neither the gene set name nor the type of identifiers used to specify the genes follow a common standard. Reactome, for instance, uses UniProt identifiers (open protein sequence database) [The17b] as primary identifier type for genes. However, we can use external annotation services to translate foreign gene identifier types to a preferred one. An example of such an annotation service is Ensembl BioMart [ZAA+18], where a researcher can use the web side or a web service API. Moreover, a gene set curator can assign a gene set with additional information like alternative gene identifiers (e.g. gene symbols alongside to UniProt identifiers) or a description text. The description text contains more details about meaning or origin of the gene set, but is in general unstructured. In addition, it is possible to enrich gene sets with information form other sources such as GO terms, gene traits, or any type of unstructured gene description.

Every used reference trees is rooted and can have inner notes with two or more children. Nodes in a reference tree represents exactly one gene set. We assume that the reference tree structure resembles the hierarchical relationships between the gene sets. This assumption allows us to interpret the tree path lengths between two gene sets as biological distance. We therefore consider a metric as biologically plausible if the pairwise distances calculated with the particular metric correlate with the pairwise path lengths.

3.1.1. Standard structure

For illustrating the materials and methods used in this work, we define a data point for each gene set as 8-tuple:

$$G = (G_N, G_D, G_{ID}, G_S, G_{GD}, G_{GO}, G_{GOD}, G_T) \tag{3.1}$$

Each data point consists of the gene set name G_N, a gene set description G_{RD}, the actual set of genes as NCBI gene identifiers [MOPT07] G_{ID} and gene symbols G_S, a set of gene descriptions G_{GD}, a multiset of GO terms G_{GO}, a multiset of GO term descriptions G_{GOD}, and a multiset of phenotype traits G_T. Table 3.1 summarizes each element of a data point. We allow duplication in the multiset elements, because frequency in which formal words appear in data points could influence the similarity between gene sets. For instance, data point G_1 and G_2 can share the same set of GO terms but the number in which each GO terms appear in each data point can still be a differentiating factor.

Symbol	Source	Summary
G_N	Reactome, Sturm et al.	Gene set name where $G_N \in \Sigma^*$
G_D	Reactome	Gene set description where $G_D \in \Sigma^*$
G_S	Reactome, Sturm et al.	Set of gene symbols where $G_S \subset \Sigma^*$
G_{GD}	NCBI	Set of gene descriptions $G_{GD} \subset \Sigma^*$
G_{ID}	Reactome, BioMart	Set of NCBI Gene identifiers with $G_{NCBI} \subset L_{NCBI}$
G_{GO}	BioMart	Multiset of GO terms where $G_{GO} \subset L_{GO}$
G_{GOD}	BioMart	Multiset of GO term descriptions where $G_{GOD} \subset \Sigma^*$
G_T	GWAS	Multiset of phenotype traits with $G_T \subset L_{GWAS}$

Table 3.1.: Type of gen set information used in this study.

We assume that G_N is a string from the trivial formal language [Men09] Σ^* where Σ is the entire ASCII alphabet. This assumption also applies to any description G_D provided with the gene set. We chose NCBI as preferred gene identifiers for gene set genes, because Reactome already provide them as alternative to UniProt identifiers. In addition, BioGrid uses NCBI IDs as primary identifiers for their PPI network data. G_{ID} is therefore always a subset of L_{NCBI} where the formal language $L_{NCBI} \subset \{0, \ldots, 9\}^*$ represents all existing NCBI identifiers. Additionally, we collect the gene symbols G_S that correspond to the NCBI identifiers, where each gene symbol is an element of Σ^*. It is more common to mention genes by their gene symbols in literature and therefore are more likely to appear in word embedding models. Moreover, NCBI offers gene descriptions, which we collect as set G_{GD} for each gene set. These gene descriptions can contain information like function, mutations, or pathology of a specific gene as unstructured text (hence $G_{GD} \subset \Sigma^*$). We used the BioMart annotation web services to collect all GO terms from each NCBI gene ID as multiset G_{GO}. All GO terms are part of a standardized collection of identifiers, which we address by $L_{GO} \subset \{GO:\} \cdot \{0, \ldots, 9\}^*$. In addition, we used BioMart web services to download the term description of each GO term as multiset G_{GOD}. The GO consortium reviews the individual term description, but the descriptions follow neither a standard vocabulary nor a format (i.e., $G_{GOD} \subset \Sigma^*$). Finally, we used a downloadable database from the GWAS web site to gather all phenotype traits from each gene as multiset G_T. The phenotype traits are all part of a controlled vocabulary we identify as $L_{GWAS} \subset \Sigma^*$.

3.1.2. Reactome reference trees

Reactome is a publicly available database that provides peer-reviewed pathways [JTGV+05]. The graph database consists of more than 20 rooted trees where each tree represents a high level activity such as “Cell Cycle”, “DNA Repair”, or “Developmental Biology”. Each tree contains a hierarchy of pathways that either resemble a sub-activity within a bigger pathway or share a common effect (e.g. “Disease of Immune System” and “Neurodegenerative diseases” share the effect “disease” and are therefore children of pathway “Disease”). Every pathway in Reactome has an unique ID (e.g. R-HSA-1643685), a name, a set of gene with UniProt as primary gene identifier, alternative identifiers like NCBI gene IDs, and a gene set description as unstructured text. The meaning of the description text varies depending

on how deep the pathway is located in one of the over 20 pathway trees. In pathway "Disease", the descriptions resembles an introducing summary over diseases. In contrast, the descendant pathway "Binding of SHC1 to p-6Y-EGFR mutants" describes all actor (genes and other organic molecules), interactions, and references involved in the pathway.

For the evaluation of metrics, we will interpret every (sub) pathway as autonomous gene set. We utilize the following work flow to extract reference trees and data points for a given Reactome ID: First, we consume the web service API from Reactome to download the reference tree and all aforementioned information for each ancestor pathway. Second, we used the same web API from Reactome to extract gene identifiers. The primary identifiers are UniProt IDs, but Reactom also offers NCBI IDs. We, however, filtered genes from a pathway if their associated UniProt ID has no corresponding National Center for Biotechnology Information (NCBI) IDs. Third, we removed all ancestor pathways that appear more then once within the reference tree. Fourth, we used web services from BioMart, NCBI, and the data source from Genome-wide association studies (GWAS) to annotate every gene set with gene descriptions, GO information, and phenotype traits.

Reactome ID	Tree depth	No. used gene sets	No. removed gene sets
R-HSA-1474290	4	82	0
R-HSA-373755	4	48	0
R-HSA-8982491	5	40	0
R-HSA-422475	3	335	1

Table 3.2.: Reactome pathways used as reference trees.

3.1.3. Immune cell differentiation hierarchy

All cells of the human immune system descend from a common progenitor immune cell. Progenitor cells are specialized stem cells with the ability to differentiate into a dedicated set of cell types. It is possible to identify differentiation state and function of an immune cell by its so called marker genes. Marker genes are genes whose expression levels give evidence to a specific cell function or state. Gregor et al. [SFP+18] utilize this knowledge to evaluate cell-type quantification methods. They investigated the prediction accuracy of algorithms that quantify the fraction of cell types in any given tissue sample. For the evaluation process, Gregor et al. conducted a literature research to curate a hierarchy of 45 cell types. In addition, they collected a set of marker gene symbols for 38 of the 45 cell types if they are backed up by previous studies. Gregor et al. provide the cell type hierarchy and the marker genes as downloadable Tab Separated Values (TSV) documents on a public source repository [SFP+18].

For our analysis, we interpreted the cell types with their individual set of marker genes as gene sets and the cell type hierarchy as reference tree. We performed the following preprocessing steps on the raw data: First, we excluded 5 of 45 cell types that neither have any child cell types with marker genes nor carry marker genes themselves. Second, we performed a post-order traversal of the reference tree to calculate missing sets of marker genes for the inner nodes of the cell type tree. For each inner node without marker genes, we calculate the union over all marker genes set that are direct children of the particular inner node. This is in our opinion reasonable, because pathways in Reactome are always a supersets of their child pathways. Third, we removed 4 of 40 cell types that are cancer cells, because we are unsure if cancer cell provide stable marker genes. This filtering step leaves us with 36 out of 45 immune cell types. Fourth, we used BioMart to translate gene symbols into NCBI gene identifiers. In addition, we used BioMart and the other annotation services mentioned in section 3.1.1 to obtain data points. An illustration of the immune cell reference tree is in fig. 3.1.

We may want run experimen on these cells well

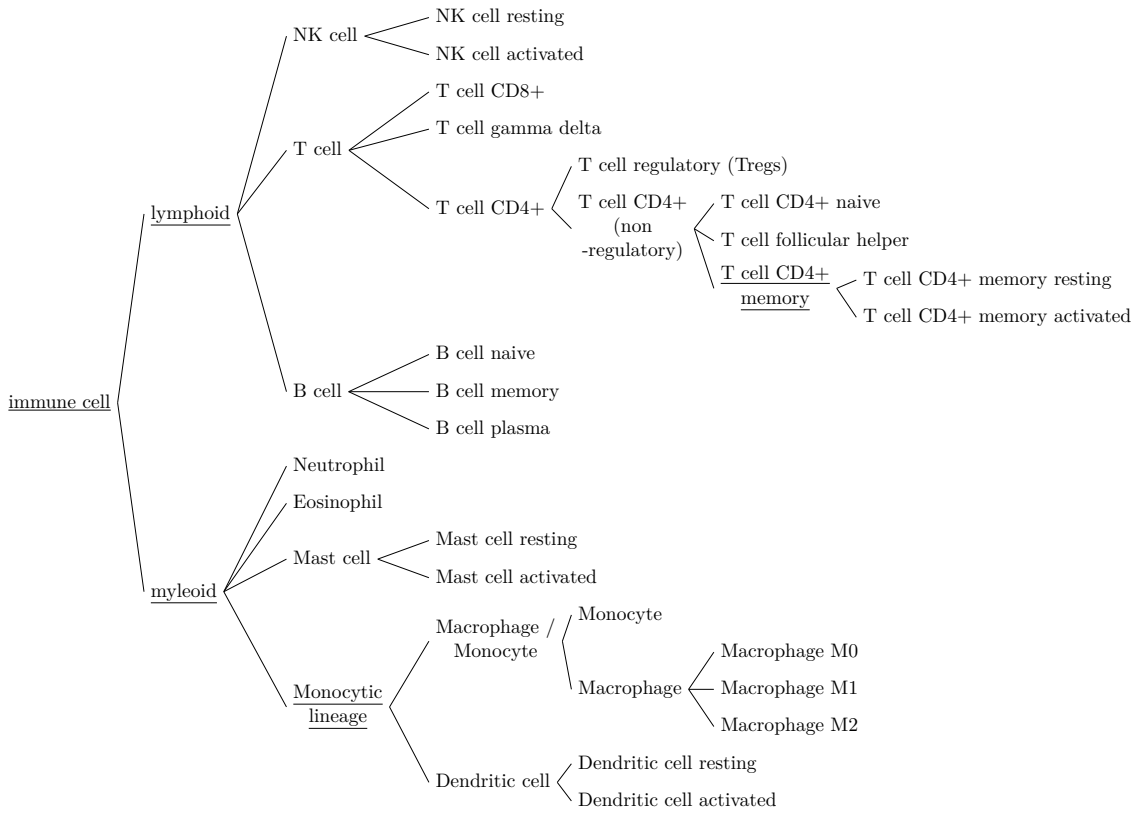


Figure 3.1.: Immune cell type tree used as reference. Data from Georg et al. does not provide marker genes for the underlined cell types

3.2. Gene Set Metrics

Let $\mathcal{G}$ be the universe of all data points G that follow the 8-tuple schema for gene sets in eq. (3.1). We assume that all data points in $\mathcal{G}$ are organized in a rooted tree according to their biological relationship. Furthermore, let $l_{\mathcal{G}}(G_1, G_2)$ be the path length between G_1, $G_2 \in \mathcal{G}$ within this tree. Our goal is to find a mathematical metric [AF90] $d : \mathcal{G} \times \mathcal{G} \to [0, \infty)$ so that $d(X, Y)$ and $l_{\mathcal{G}}(X, Y)$ are statistical depended for any random variable X and Y.

Every gene set metric we present has two degrees of freedom: A projection $p : \mathcal{G} \to X$ and a distance function $D : X \times X \to [0, \infty)$ over a set X. The operator D performs the actual statistical or algorithmic calculation where p projects a data point from $G \in \mathcal{G}$ to a mathematically controlled space X. We use the set $\mathcal{A} = \{N, D, ID, S, GO, GOD, GD, T\}$ to address an specific element of data point $G \in \mathcal{G}$.

Moreover, we will use the following nomenclature to distinguish presented metrics from each other:

$$\begin{aligned} d_{p,D} &: \mathcal{G} \times \mathcal{G} \to [0, \infty) \\ &(G_1, G_2) \mapsto D\left(t(G_1), t(G_2)\right). \end{aligned} \tag{3.2}$$

Table 3.3 shows a list of all metrics used in this work.

3.2.1. Statistical metrics

We define a gene set metric as statistical metric if it does not rely on external topologies such as trees, networks or word embeddings. Moreover, we identify two subgroups of statistical methods: 1) set-based metrics (e.g., over Jaccard coefficient, and overlap coefficient) and 2)

Metric	Description
Statistical metrics	
d_{P_{ID},D_J}	Jaccard distance over NCBI gene identifiers G_{ID}
d_{P_T,D_J}	Jaccard distance over gene traits G_T
d_{P_{ID},D_O}	Overlap distance over NCBI gene identifiers G_{ID}
d_{P_T,D_O}	Overlap distance over gene traits G_T
$d_{\text{bin}_{ID},D_{p=1}}$	Manhattan distance over NCBI gene identifiers G_{ID}
$d_{\text{bin}_T,D_{p=1}}$	Manhattan distance over gene traits G_T
$d_{\text{count}_T,D_{p=1}}$	Manhattan distance over gene trait counts
$d_{\text{bin}_{ID},D_{p=2}}$	Euclidean distance over NCBI gene identifiers G_{ID}
$d_{\text{bin}_T,D_{p=2}}$	Euclidean distance over gene traits G_T
$d_{\text{count}_T,D_{p=2}}$	Euclidean distance over gene trait counts
$d_{\text{bin}_{ID},D_\kappa}$	κ distance over NCBI gene identifiers G_{ID}
$d_{\text{bin}_T,D_\kappa}$	κ distance over gene traits G_T
d_{count_T,D_C}	Cosine distance over gene trait counts
Graph-based metrics	
$d_{GO_{BP},D_{Wang,BMA}}$	Wang method over GO terms (biological process terms)
$d_{GO_{CC},D_{Wang,BMA}}$	Wang method over GO terms (cellular component terms)
$d_{GO_{MF},D_{Wang,BMA}}$	Wang method over GO terms (molecular function terms)
$d_{GO_{BP},D_{Resnik,BMA}}$	Resnik method over GO terms (biological process terms)
$d_{GO_{CC},D_{Resnik,BMA}}$	Resnik method over GO terms (cellular component terms)
$d_{GO_{MF},D_{Resnik,BMA}}$	Resnik method over GO terms (molecular function terms)
$d_{P_{ID},D_{Dijkstra,BMA}}$	Dijkstra method over NCBI gene identifiers in PPI networks
d_{N_{ppi},D_J}	Jaccard distance over direct neighbor NCBI gene identifiers in PPI networks
d_{N_{ppi},D_O}	Overlap distance over direct neighbor NCBI gene identifiers in PPI networks
Word embedding metrics	
$d_{\text{w2vSum}\circ\ P_S,D_C}$	Cosine distance over word vectors of gene symbols G_S
$d_{\text{w2vSum}\circ\ P_D,D_C}$	Cosine distance over word vectors of descriptions G_D
$d_{\text{w2vSum}\circ\ P_{GD},D_C}$	Cosine distance over word vectors of gene descriptions G_{GD}
$d_{\text{w2vSum}\circ\ GOD_{BP},D_C}$	Cosine distance over word vectors of Gene Ontology (GO) terms (biological process)
$d_{\text{w2vSum}\circ\ GOD_{CC},D_C}$	Cosine distance over word vectors of Gene Ontology (GO) terms (cellular components)
$d_{\text{w2vSum}\circ\ GOD_{MF},D_C}$	Cosine distance over word vectors of Gene Ontology (GO) terms (molecular function)
$d_{\text{w2v}\circ\ P_S,D_{WM}}$	Word Mover's distance over word vectors of gene symbols G_S
$d_{\text{w2v}\circ\ P_D,D_{WM}}$	Word Mover's distance over word vectors of descriptions G_D

Table 3.3.: Implemented gene sets metrics.

vector-space-based metrics (e.g., Minkowski distance, cosine distance, and distance based on κ coefficient).

For set-based methods, we define the trivial projection P_A that selects an attribute from the data point G as follows:

$$\begin{aligned} P_A : \mathcal{G} &\to \mathcal{P}(\Sigma^*) \\ G &\mapsto G_A, \end{aligned} \qquad A \in \mathcal{A} = \{N, D, ID, S, GO, GOD, GD, T\} \tag{3.3}$$

This allows us to explicitly state which information of the gene set we want to use for defining metrics. We used the Jaccard index J and the overlap coefficient [MKK16] to define corresponding distance functions:

$$D_J(A,B) = 1 - J(A,B) = \begin{cases} 0 & \text{if } A = B = \emptyset \\ 1 - \dfrac{|A \cap B|}{|A \cup B|} & \text{else.} \end{cases} \tag{3.4}$$

$$D_O(A,B) = 1 - \text{overlap}(A,B) = \begin{cases} 0 & \text{if } A = B = \emptyset \\ 1 - \dfrac{|A \cap B|}{\min(|A|, |B|)} & \text{else.} \end{cases} \tag{3.5}$$

Both distances generate values between 0 and 1, where 0 represents perfect overlap. in contrast to $D_J(A,B)$, $D_O(A,B)$ will always return 0 if one of the two arguments is a subset of the other argument.

For vector-space-based gene set metrics, let $\mathcal{G}_A = \bigcup_{G_i \in \mathcal{G}} P_A(G_i)$ be the set of formal words that appear in at least one data point $G \in \mathcal{G}$ when looking at attribute $A \in \mathcal{A}$. We can then project $G \in \mathcal{G}$ as binary vector in $\mathbb{R}^n$, $n = |\mathcal{G}_A|$ by the following function:

$$\begin{aligned} &\text{bin}_A \colon \mathcal{G} \to \mathbb{R}^n \\ &\quad G^A \mapsto x \text{ such that } x_j = \begin{cases} 1 \iff w_j \in P_A(G_i) \\ 0 \quad else \end{cases} \quad \text{for all } j \in \{1, \ldots, n\}, w_j \in \mathcal{G}^A. \end{aligned} \tag{3.6}$$

For example, the j-th entry in $\text{bin}_{ID}(G)$, is 1 if the j-th NCBI gene identifier is present in data point G.

As mentioned in section 3.1.1, we purposely allow duplicates for certain gene set information (i.e., GO terms, GO descriptions, and phenotype traits). We therefore introduce a transformation count_A that assigns every data point $G \in \mathcal{G}$ to a counts vector $f \in \mathbb{R}^n$ with $n = |\mathcal{G}_A|$:

$$\begin{aligned} &\text{count}_A \colon \mathcal{G} \to \mathbb{R}^n \\ &\quad G_A \mapsto f \text{ such that } f_j = |\{w \in G_A : w = w_j\}| \text{ for all } j \in \{1, \ldots, n\}, w_j \in \mathcal{G}_A. \end{aligned} \tag{3.7}$$

The Minkowski distance [Tre06] describes a class of mathematical distances over $\mathbb{R}^n$, $n \in \mathbb{N}$ and is defined as

$$D_p(x,y) = \begin{cases} \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} & \text{if } p \in [1, \infty) \\ \max\limits_{i \in \{1, \ldots n\}} |x_i - y_i| & \text{if } p = \infty. \end{cases} \qquad \forall x, y \in \mathbb{R}^n \tag{3.8}$$

The k-means clustering algorithm, for instance, uses the Euclidean distance ($p = 2$), for measuring the error between centroids and assigned vectors.

The cosine distance [MKK16] is defined as

$$D_C(x,y) = 1 - \frac{x \cdot y}{\|x\| \, \|y\|} \qquad \forall x, y \in \mathbb{R}^n \tag{3.9}$$

where $x \cdot y$ is Euclidean dot product between x and y. It is the preferred operator to measure the distance between word embedding vectors.

The κ coefficient κ measures the agreement between two classifiers [AP08]. Its general form is defined as:

$$\kappa = \frac{p_0 - p_e}{1 - p_e}, \quad p_e = \frac{1}{N^2} \sum_k n_{k,1} n_{k,2}, \quad \kappa \in [-1, 1]$$

Where p_0 is the agreement between the two classifiers relative to the number of N observations and p_e is the probability that the two classifier agree by chance. The factors n_{ki} represent the number of times where classifier $i \in \{1, 2\}$ predicts category k. If the classifiers have total agreement with each other, the resulting κ coefficient is 1. A κ coefficient of 0 or below implies total disagreement or classification performance worse than random selection. If we interpret gene sets as binary classifiers (e.g. classifiers for gene symbols), we can utilize the κ coefficient as distance operator for x, $y \in \mathbb{R}^n$:

$$D_\kappa(x, y) = 1 - \kappa(x, y) = \frac{1 - p_0(x, y)}{1 - p_e(x, y)} \qquad \begin{aligned} p_0(x, y) &= \frac{c_1(x, y) + c_0(x, y)}{n} \\ p_e(x, y) &= \frac{n_0(x) n_0(y) + n_1(x) n_1(y)}{n^2} \end{aligned} \tag{3.10}$$

Where $n_k(v)$ is the number of entries in vector $v \in \mathbb{R}^n$ that have value k, and $c_k(v_1, v_2)$ the number of times where both vector v_1 and v_2 have value k at position $i \in \{1, \ldots, n\}$.

We implemented the statistical gene set metrics by using Python software packages scikit-learn [PVG+11] for the κ coefficient and SciPy [JOP+] for Jaccard intex, cosine similarity, and Minkowski distances.

3.2.2. Graph-based metrics

For graph-based gene set metrics, we require an additional data structure that models relationships between data point attributes (like NCBI gene IDs G_{ID}) as a mathematical graph. We distinguish between methods that assume a Directed Acyclic Graphs (DAG) (e.g. gene ontologies) and methods that operate on arbitrary graphs (e.g. protein-protein interaction networks). We incorporated the Resnik similarity [Res99] and the Wang similarity [WDP+07] as basis for the DAG-based distances $D_{Wang,BMA}$ and $D_{Resnik,BMA}$. Both Resnik and Wang propose their similarity functions for comparing the similarity between two individual GO terms. Furthermore, we used the Dijkstra algorithm [CLRS09] to measure the path lengths between two individual genes in a PPI network ($D_{Dijkstra,BMA}$). For each of the three methods, we apply the Best Match Average (BMA) strategy [PFB+08] to combine the results of pairwise comparison. Moreover, we define projections for GO information that filters terms based on their category (e.g., GO_{BP} for selected only terms that describe a biological process). Finally, the N_{ppi} projection extends a set of NCBI gene identifiers by its direct neighbors in the PPI network.

Let $DAG_{GO} = (L_{GO}, E_{GO})$ be the DAG over all GO terms L_{GO} where E_{GO} represents the GO term relationships as tuple. Furthermore, let C_t be the set that contains all descendant terms of GO term $t \in L_{GO}$ and t itself. The Resnik similarity is then defined as

$$\begin{aligned} \text{sim}_{Resnik} : L_{GO} \times L_{GO} &\to [0, 1] \\ (t_1, t_2) &\mapsto IC(LCA(t_1, t_2)), \end{aligned} \qquad IC(t) = -\log\left(\frac{|C_t|}{|L_{GO}|}\right),$$

where $LCA(t_1, t_2)$ is the lowest common ancestors of the GO terms t_1, t_2. The function $IC(t)$ describes the information content of term t relative to the entire ontology. Two GO terms are similar in the context of Resnik similarity, if their lowest common ancestor selects a relatively small portion of the gene ontology. We define the corresponding Resnik distance as $D_{Resnik}(t) = 1 - \text{sim}_{Resnik}(t)$ for every Gene Ontology (GO) term t.

For the Wang similarity, let T_t be the set that includes all ancestors of t and t itself. The following function then calculates the Wang similarity

$$\begin{aligned}\mathrm{sim}_{Wang} : L_{GO} \times L_{GO} &\to [0,1] \\ (t_1, t_2) &\mapsto \frac{\sum_{t \in T_{t_1} \cap T_{t_2}} (S_{t_1}(t) + S_{t_2}(t))}{SV(t_1) + SV(t_2)},\end{aligned}.$$

Where $S_t(t')$ is the semantic contribution of t' to the ancestor t and $SA(t)$ the sum of all semantic contributions over all ancestors of t:

$$S_t(t') = \begin{cases} 1 & \text{if } t' = t \\ \max\{w_e \cdot S_t(c) \colon c \text{ is children of } t'\} & \text{else} \end{cases}, \quad SV(t) = \sum_{t' \in T_t} S_t(t'),$$

The factor w_e is called semantic contribution of edge e in E_{GO}. Wang suggest to assign constant values to semantic contribution factors based on the relation type of the edges (e.g., "is-a", "regulates", ...). We define the corresponding Wang distance as $D_{Wang}(t) = 1 - \mathrm{sim}_{Wang}(t)$ for every Gene Ontology (GO) term t.

The comparison of two (multi)sets of GO terms requires an additional combination strategy. Let GO_1, GO_2 be subsets of L_{GO} and D_f a distance function for individual terms (e.g. D_{Resnik}, D_{Wang}). The BMA strategy then calculates the distance between two Gene Ontology (GO) term multisets by the following expression:

$$D_{f,BMA}(GO_1, GO_2) = \frac{\sum_{t_1 \in GO_1} \min_{t_2 \in GO_2} D_f(t_1, t_2) + \sum_{t_2 \in GO_2} \min_{t_1 \in GO_1} D_f(t_1, t_2)}{|GO_1| + |GO_2|}.$$

The function $D_{Wang,BMA}$, for instance, computes the Wang distance over two sets of terms by using the BMA combination strategy. An important property of the BMA strategy is that the identity-of-indiscernibles property for distances functions holds if D_f is a distance function (i.e., $d(x, y) = 0 \iff x = y$).

Moreover, one can generalize the BMA strategy for an arbitrary distance function D_f that operates on elements from an arbitrary set. We use the BMA strategy in combination with the Dijkstra algorithm to define a gene set metric over PPI networks ($D_{Dijkstra,BMA}$). The Dijkstra algorithm uses a greedy approach to find the shortest path between two nodes. In the context of PPI networks, the input nodes resemble NCBI gene identifiers from a data point G.

Next, we define projections for extracting Gene Ontology (GO) terms from data points of a certain category. The tree Gene Ontology (GO) categories are "Biological Process" (BP), "Cellular Component" (CC), and "Molecular Function" (MF), wich also represent the tree root terms for all other GO terms. We use $L_{GO_{BF}}$, $L_{GO_{CC}}$, $L_{GO_{MF}}$ to differentiate between GO term of different categories, i.e. $L_{GO_{BF}} \cup L_{GO_{BF}} \cup L_{GO_{BF}} = L_{GO}$. We do this analogously for GO term descriptions $L_{GOD_{BF}}$, $L_{GOD_{BF}}$, and $L_{GOD_{BF}}$. Furthermore, we define projections for ontology terms GO_X and ontology term descriptions GOD_X as:

$$\begin{aligned} GO_X : \mathcal{G} &\to \mathcal{P}(L_{GO_X}) \\ G &\mapsto P_{GO}(G) \cap L_{GO_X}, \end{aligned} \qquad X \in \{BP, MF, CC\} \tag{3.11}$$

$$\begin{aligned} GOD_X : \mathcal{G} &\to \mathcal{P}(L_{GOD_X}) \\ G &\mapsto P_{GOD}(G) \cap L_{GOD_X}, \end{aligned} \qquad X \in \{BP, MF, CC\} \tag{3.12}$$

This separation is necessary, because we want to see how important the GO category is during the distance evaluation phase. In addition, the Resnik method does not support comparison with mixes GO terms.

Finally, we define the N_{ppi} projection as

$$\begin{aligned} N_{ppi} : \mathcal{G} &\to \mathcal{P}(L_{NCBI}) \\ G &\mapsto P_{ID}(G) \cup \{g' \in L_{NCBI} \mid \exists g \in P_{ID}(G) : \{g, g'\} \in E_{PPI}\}, \end{aligned} \tag{3.13}$$

where E_{PPI} contains the edges of a given PPI network.

The tab-delimited data we used for the PPI network originate from the BioGRID [CaOB+17] homepage. We extracted the edges as tuple of NCBI identifiers from each row and filtered edges that did not originate from homo sapiens. We used an implementation of the Dijkstra algorithm provided by the SciPy package as base for the Dijkstra distance function. The R package GOSemSim [YLQ+10] offered the implementations for the Resnik similarity and the Wang similarity. In addition, we used the R data package org.Hs.eg.db [Car18], which includes the relationships between GO terms for the human species.

3.2.3. Word embedding metrics

The two crucial parts of the presented word embedding metrics are 1) the embedding model and 2) the functions to project gene set information into the word2vec vector space. Additionally, we included the WMD as specialized mathematical distance for the word embedding vector spaces.

For explaining the Word2vec model and the projection function, we need to clarify the difference between formal words and natural language words. A formal word is an element of a formal language $L \subset \Sigma^*$ over an alphabet Σ and can contain every character. On the contrary, a word in the context of natural language consists only of printable characters. We use the term "vocable" to distinguish formal words from natural language words. In a trained Word2vec model $M \in \mathbb{R}^{v \times d}$, we represent every vocable by a d-dimensional row vector in M where v is the size of the model vocabulary. Our reference Word2vec model is a pretrained model provided by Wilbur et al. [KFWL17]. They extracted over 25 million titles and abstracts from publications in PubMed (until March 2016). Wilbur et al. trained the model with the skip-gram variant of the Word2vec learning method.

Link to datase

As mentioned in the previous section, we cannot project formal words directly into the word vector space. It is therefore necessary to extract usable vocables from formal words / arbitrary text. For our evaluation, we conducted the following steps for each formal word: First, we replace every upper case character by a lower case character. This is necessary because our reference model is only trained for lower case vocables. Second, we split each formal word by any kind of word divider (dot, coma, colon, semicolon, whitespace, tabulator, slash, hyphen, brackets) and obtain a list of vocables. Third, we remove numbers and out-of-vocabulary elements from the split list. Finally, we remove stopwords based on the English language. Stopwords represent a set of frequently used vocables in a language that don't add significant value to a list of vocables (e.g. "the" or "of" in the English language).

With this preprocessing steps in mind, we define the transformation of an arbitrary subset of a formal language into a (mult)iset of vocables:

$$\text{w2v} : \mathcal{P}(\Sigma^*) \to \mathcal{P}(\Sigma^*), \quad W \mapsto \bigcup_{w \in W} \text{text2vocs}(w),$$

where text2vocs performs the previously mentioned preprocessing steps for each formal word. Furthermore, we define

$$\text{w2vSum} : \mathcal{P}(\Sigma^*) \to \mathbb{R}^d, \quad W \mapsto \sum_{v \in \text{w2v}(W)} \text{vec}(v),$$

as sum of all vocables extracted for a given set of formal words. The expression $\text{vec}(v)$ returns the vector representation of the vocable v. The w2vSum allows us to build projections from gene set date points into word embedding as composition with other projection function. For instance, we use the composition $\text{w2vSum} \circ P_S$ to denote the transformation of gene symbols into word vectors. Note that Word2vec generally ignores the order of words, which makes it unnecessary to preserve the order of vocables from the input set.

Existing mathematical distance functions over $\mathbb{R}^d$, like the cosine distance D_C, are suitable for building gene set metrics. Alternatively, we can use functions like the Word Mover Distance (WMD) [KSKW15], which are more suited for document-based queries. In the context of WMD, we use the term "document" as synonym for (multi)sets of words. The WMD distance takes two documents and calculates the document distances based on the optimal pairwise pairing of vocables (see fig. 3.2).

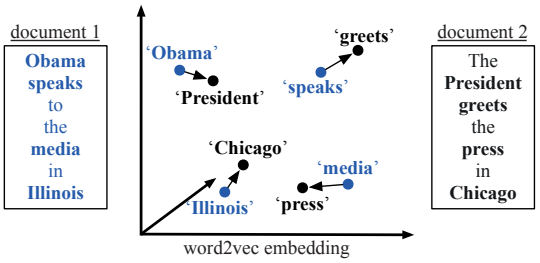


Figure 3.2.: Illustration of WMD distance of two documents [KSKW15]

pand descrip-
n?

We implemented the word-embedding-based gene set metrics by using the Python packages Gensim [ŘS10] and NLTK [LB02]. Gensim is a widely used software package for training Word2vec models, performing common Word embedding operations, and performing document queries with WMD. The NLTK package offers utility functions for downloading English stopwords.

4. Results & Discussion

We evaluate all presented gene set metrics by assessing the correlation of pairwise distance values with pairwise path lengths per gene set data point. We found out that gene set metrics based on the overlap distance yields the best results in 4 out of 5 Reactome and immune cell tree reference datasets. We were able to detect moderate correlation (absolute correlation value > 0.5) in all but one reference datasets. All reference dataset show at least week correlation (absolute correlation value > 0.3) for at least one gene set metric.

4.1. Ground-truth comparison

We assume that the reference tree path lengths between gene sets represent the ground truth. To compare path lengths with metric results, we calculate distance matrices for each reference dataset $\mathcal{G}$. A distance matrix is a symmetric matrix $M \in \mathbb{R}^{n \times n}$ where n is the number of gene set data points in the reference data. Each element M_{ij} contains the pairwise distance $d_{p,D}(g_i, g_j)$ where g_i, g_j are gene set data points from $\mathcal{G}$ and $d_{p,D}$ is one of the presented gene set metrics. The symmetry property and the identity-of-indiscernibles property of distance functions allows us to ignore calculations for the lower triangle and the main diagonal of the distance matrix (as seen in eq. (4.1))

$$M = \begin{pmatrix} \boxed{0} & \boxed{M_{12}} & M_{13} & \dots & M_{1n} \\ & \boxed{0} & \boxed{M_{13}} & \dots & M_{2n} \\ & & \boxed{0} & \ddots & \vdots \\ & & & \ddots & \boxed{M_{n-1,n}} \\ & & & & \boxed{0} \end{pmatrix} \tag{4.1}$$

The evaluation process against the ground-truth consists of 3 major steps. First, we calculate the distance matrix of pairwise path lengths between gene sets. We repeat this for all reference datasets. Second, we calculate the distance matrix for a given gene set matrix $d_{p,D}$ over a given reference data set. We repeat this step for every distance metric and for every reference dataset. Third, we perform correlation coefficient calculations between gene set metrics and the pairwise path lengths.

At the end of the third step, we have two correlation matrices C^P and C^S for each reference dataset. The matrix entries $C^P_{d_i,d_j}$ stands for the Pearson Correlation Coefficient (PCC) [KFVSL$^+$17] between the distance matrix values calculated by gene set metrics

d_i, d_j or the reference tree path length. Analogously, the matrix entries $C^S_{d_i,d_j}$ represent the Spearman's Rank Correlation Coefficient (SRCC) [KFVSL+17] between two gene set metrics d_i, d_j or a gene set metric and the reference path lengths. Researcher use the Pearson Correlation Coefficient (PCC) to measure the linear dependence between two observable variable. The SRCC assesses the monotonic relationship between two variables. Both correlation types have a value range between -1 and 1, where a correlation value near -1 or 1 implies stronger linear or monotonic dependence than a correlation value near 0. To be more precise, we distinguish between 4 correlation levels:

- No correlation if correlation value is in $(-0.3, 0.3)$
- Weak correlation if correlation value is in $(-0.5, -0.3]$ or $[0.3, 0.5)$
- Moderate correlation if correlation value is in $(-0.9, -0.5]$ or $[0.5, 0.9)$
- Strong correlation if correlation value is in $[-1.0, -0.9]$ or $[0.9, 1.0]$

Figure 4.1 illustrates an excerpt of the C^P matrix. It shows only the correlation values between gene set metrics and the pairwise path lengths merged from all reference datasets.

	R-HSA-8982491	R-HSA-1474290	R-HSA-373755	R-HSA-422475	immune_only
Pairwise path length in reference tree	1.0	1.0	1.0	1.0	1.0
Jaccard distance over extended gene set	0.231	0.264	0.413	0.312	0.294
WM distance over summary W2V	0.227	0.446	0.196	0.267	NaN
WM distance over gene symbols W2V	0.279	0.372	0.434	0.358	0.381
Cosine distance over over summary W2V	0.183	0.332	0.114	0.064	NaN
Cosine distance over over NCBI summary W2V	0.14	0.197	0.144	0.22	0.276
Cosine distance over gene symbols W2V	0.34	0.267	0.388	0.231	0.268
Cosine distance GO MF description W2V	0.232	0.105	0.22	0.109	0.274
Cosine distance GO CC description W2V	0.141	0.117	0.15	0.096	0.093
Cosine distance GO BP description W2V	0.036	0.111	0.191	0.142	0.3
GO-distance (go_type=MF, measure=Wang, combine=BMA)	0.455	0.078	0.299	0.151	0.161
GO-distance (go_type=CC, measure=Wang, combine=BMA)	0.078	0.111	0.238	0.158	0.134
GO-distance (go_type=BP, measure=Wang, combine=BMA)	0.177	0.218	0.303	0.19	0.209
Overlap distance over genes	0.485	0.319	0.602	0.544	0.495
Overlap distance over gene traits	0.583	0.348	0.563	0.464	0.458
Minkowski distance (p=2) over genes	-0.1	-0.072	-0.091	-0.205	-0.256
Minkowski distance (p=2) over gene traits	-0.023	-0.041	0.002	-0.132	-0.246
Minkowski distance (p=2) over gene trait frequency	-0.138	-0.091	-0.151	-0.146	-0.281
Minkowski distance (p=1) over genes	-0.168	-0.103	-0.188	-0.227	-0.26
Minkowski distance (p=1) over gene traits	-0.113	-0.074	-0.096	-0.205	-0.256
Minkowski distance (p=1) over gene trait frequency	-0.157	-0.099	-0.194	-0.204	-0.283
Kappa distance over genes	0.273	0.315	0.427	0.443	0.306
Kappa distance over gene traits	0.366	0.317	0.383	0.373	0.182
Jaccard distance over genes	0.307	0.332	0.455	0.426	0.383
Jaccard distance over gene traits	0.348	0.334	0.423	0.385	0.296
Cosine distance over gene trait frequency	0.483	0.449	0.323	0.334	0.398
Random (uniform, (0,1))	0.042	-0.008	0.018	-0.005	0.025

Figure 4.1.: Pearson correlation between gene set metrics and path lengths. The reason for NaN values in the immune cell dataset is the lack of gene set descriptions from the raw data

The majority of the presented gene set metrics, including a random baseline, show at best weak correlation. This especially applies to metrics that uses Minkowski functions as mathematical distance. Other statistical metrics, however, perform better compared to graph-based metrics and metrics that uses word embeddings. The overlap distance in particular has the highest correlation values among all reference data sets except the Reactome subtree "R-HSA-1474290". In dataset "R-HSA-1474290", we see that the WMD distances and cosine distance over gene count vectors show the highest correlations.

The SRCC values in Figure 4.2 support our finding that statistical metrics outperform more sophisticated gene set metrics based on graphs or word embeddings. Especially gene set metrics that uses the overlap distances function show moderate correlation levels. Furthermore, we see that gene identifiers and phenotype traits are enough to achieve moderate correlation where other external gene set annotations don't improve correlation. Moreover, only the Minkowski distance over phenotype traits with $p = 2$ and the cosine distance over GO description word vectors have correlation values with a p-value greater than 5%.

Add p-value heatmap?

	R-HSA-8982491	R-HSA-1474290	R-HSA-373755	R-HSA-422475	immune_only
Pairwise path length in reference tree	1.0	1.0	1.0	1.0	1.0
Jaccard distance over extended gene set	0.244	0.195	0.3	0.178	0.215
WM distance over summary W2V	0.182	0.362	0.162	0.203	NaN
WM distance over gene symbols W2V	0.266	0.349	0.367	0.21	0.335
Cosine distance over over summary W2V	0.174	0.351	0.102	0.06	NaN
Cosine distance over over NCBI summary W2V	0.161	0.251	0.318	0.228	0.267
Cosine distance over gene symbols W2V	0.39	0.277	0.382	0.196	0.278
Cosine distance GO MF description W2V	0.263	0.175	0.255	0.111	0.28
Cosine distance GO CC description W2V	0.135	0.192	0.239	0.114	0.079
Cosine distance GO BP description W2V	0.105	0.174	0.339	0.183	0.291
GO-distance (go_type=MF, measure=Wang, combine=BMA)	0.438	0.122	0.25	0.136	0.14
GO-distance (go_type=CC, measure=Wang, combine=BMA)	0.078	0.124	0.234	0.143	0.118
GO-distance (go_type=BP, measure=Wang, combine=BMA)	0.144	0.217	0.222	0.144	0.172
Overlap distance over genes	0.428	0.253	0.554	0.408	0.614
Overlap distance over gene traits	0.583	0.304	0.501	0.259	0.449
Minkowski distance (p=2) over genes	-0.09	0.012	-0.042	-0.09	-0.201
Minkowski distance (p=2) over gene traits	-0.019	0.023	-0.033	-0.084	-0.192
Minkowski distance (p=2) over gene trait frequency	-0.08	-0.011	-0.106	-0.038	-0.241
Minkowski distance (p=1) over genes	-0.09	0.012	-0.042	-0.09	-0.201
Minkowski distance (p=1) over gene traits	-0.019	0.023	-0.033	-0.084	-0.192
Minkowski distance (p=1) over gene trait frequency	-0.054	-0.001	-0.08	-0.074	-0.228
Kappa distance over genes	0.27	0.337	0.442	0.201	0.458
Kappa distance over gene traits	0.418	0.317	0.371	0.133	0.099
Jaccard distance over genes	0.316	0.319	0.503	0.4	0.604
Jaccard distance over gene traits	0.468	0.296	0.359	0.162	0.213
Cosine distance over gene trait frequency	0.555	0.434	0.264	0.209	0.388
Random (uniform, (0,1))	0.04	-0.008	0.019	-0.004	0.022

Figure 4.2.: Spearman correlation between gene set metrics and path lengths. The reason for NaN values in the immune cell dataset is the lack of gene set descriptions from the raw data

We have two possible explanations for the correlation results. First, the gene sets from Reactome pathways are always subsets of the gene sets from parent pathways. This results

in a data-induced bias towards set-based gene set metrics - especially towards the overlap distance. This is only partially true for the immune cell reference dataset and the overlap coefficient gives best correlation compared to all other gene set metrics.

Second, we do not consider any uncertainty within the gene sets or the reference trees. Even if community-driven curation process insures correctness of the data, it is only a snapshot of the actual biological knowledge base. Wrong genes in gene sets or incorrect assignments within in PPI networks or GO trees could impact the gene set metrics. This could explain why adding more information (gene descriptions, Gene Ontology (GO) terms) leads to weaker correlation that working only with essential gene set information.

4.2. Limitations

We assume that the biological ground-truth is represented as a rooted tree with gene sets as its nodes. Furthermore, we assume that the best way to compare distance metrics with the ground truth is through measuring the correlation between path lengths and metric distances. The last assumption is questionable when we compare gene sets that are siblings in the reference trees. From a path length perspective, siblings will always have the same distance. However, every presented gene set metric will assign varying distance values between different pairings of siblings. This can lead to incorrect PCC and SRCC. We considered to use distance-based tree-inference algorithm to reconstruct and compare trees from the distance metrics instead. However, tree-inference algorithm, as they are well established in phylogenetic applications, can only reconstruct dendrograms where no gene set could appear as inner node. In other words, we would need to sample dendrograms from the reference data sets, which makes it in turn harder to asses the actual usefulness of a gene set metric.

Moreover, we have to mention several technical limitations that comes with the implemented gene set metrics:

- The shortest path algorithms like Dijkstra don't consider false nodes or edges between nodes. Yu et al. formulate this use-case as robust shortest path problem, but also prove that this problem is NP-complete [YY98].
- We are not able to execute gene set metrics on WMD distance if the projected document contains to many vocables (i.e. around > 1000 vocables). We had to terminate execution as even the pairwise distance calculations took more than one hour to complete. We think that the high time complexity of $O(p^3 \log(p))$, where p is the number of unique vocables„ is the reason for this performance issue. The authors of the WMD algorithm also propose an optimization that leads to a $O(p^2)$ [KFWL17], but this optimization is not part of the Gensim Python package we use.

5. Conclusion

We started this project with the goal to measure differences between gene sets with a biologically plausible gold standard. This is the first step in simplifying gene set enrichment analysis, as it often report large lists of gene sets with redundant information. We introduced a mathematical notation to standardize gene set information and gene set metrics. Furthermore, we defined and implemented more than 20 gene set metrics by using basic statistical methods, graph-based models, and word embedding models. We, moreover, used reference data from Reactome and an immune cell type study to evaluate the gene set metrics. We, moreover, downloaded reference data from Reactome and an immune cell type study to evaluate the correlation between metric values and the path lengths in the reference trees.

Our most important finding is that gene set metrics with more sophisticated models like word embeddings or graphs do not improve the correlation with reference trees. On the contrary, we get moderate correlation results with statistical methods based on Jaccard index or overlap coefficient that only operate on gene identifiers or phenotype traits. Uncertainties in gene sets, reference trees, or models could be an explanation for this observation. Adding more information about genes that are wrongly assigned to a gene set could amplify the error in the distance calculation. In contrast, we can achieve moderate correlation if we only use phenotype traits instead of more informative gene identifiers. The findings, however, assume that a reference tree of gene sets is an appropriate representation of biological plausible distances. It is a critical assumption as the manually curated function can contain errors or might does not represent the ground-truth accurately enough.

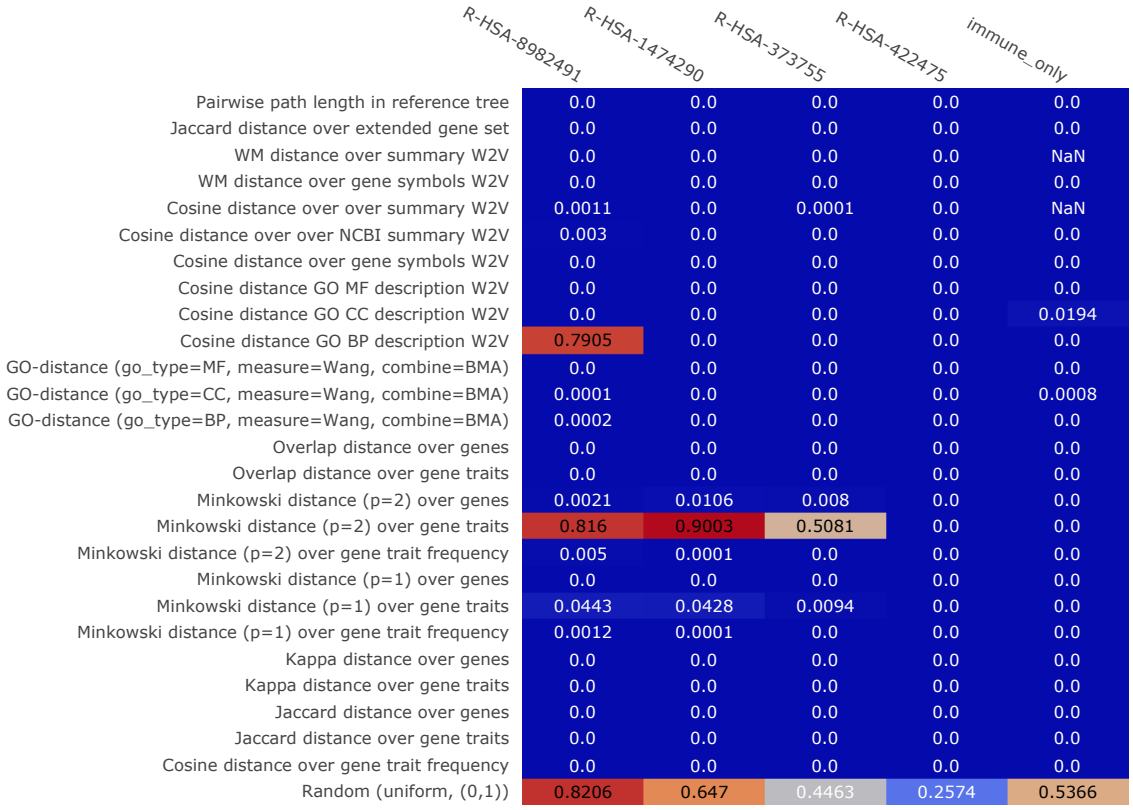
This leads us to suggest the investigation of in-silico models for gene-centric interactions. With in-silico models we can simulate artificial biological systems that mimic simple phenotype traits. We can use such minimalistic systems as more powerful ground-truth where we sample gene sets randomly or by design. However, finding a model that approximates biological systems in a computationally feasible way is a difficult question by its own. The first step is to work on fundamental biochemical network models, which are the main focus of synthetic biology studies [HtrotSF$^+$03]. Besides, every ambition to increase quality and reproducibility of published data is always a step forward, as the GIGO-principle applies also for biology and bioinformatics [BEJP$^+$04].

6. Appendix

A. Additional figures

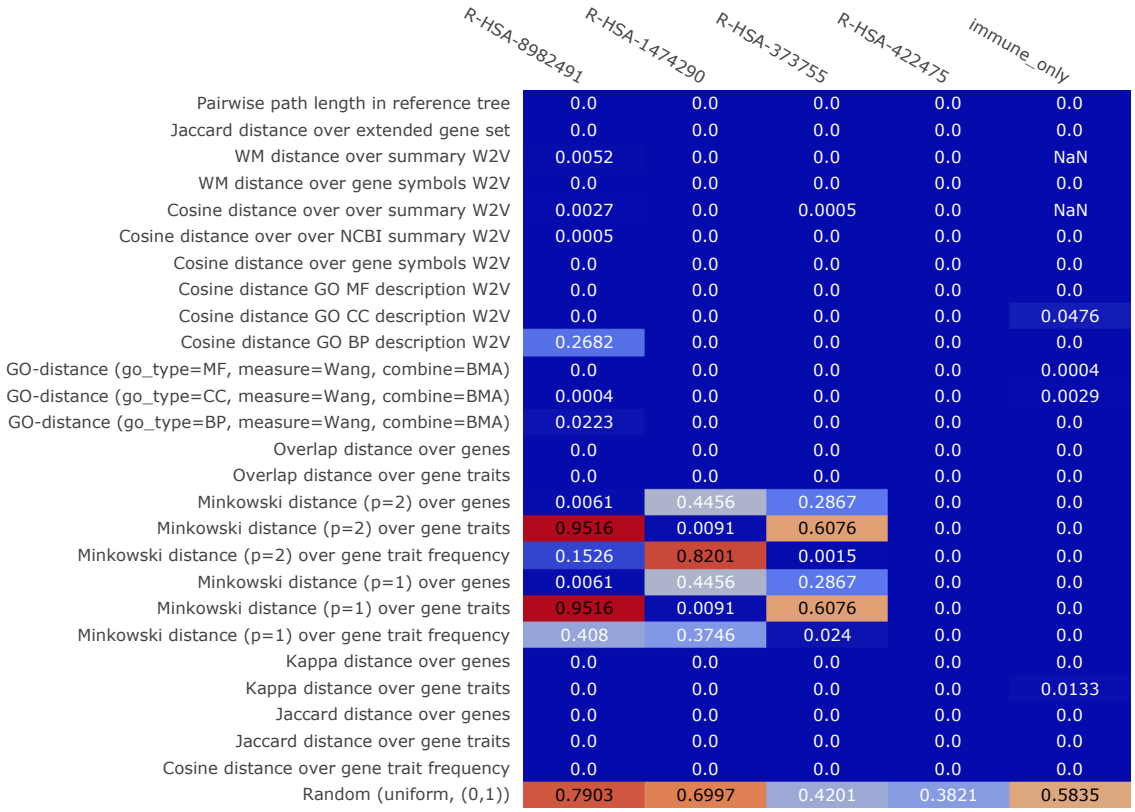
	R-HSA-8982491	R-HSA-1474290	R-HSA-373755	R-HSA-422475	immune_only
Pairwise path length in reference tree	0.012	0.029	0.014	0.161	0.028
Jaccard distance over extended gene set	0.185	0.398	0.238	2.246	0.31
WM distance over summary W2V	24.817	401.254	25.768	8212.347	NaN
WM distance over gene symbols W2V	0.814	16.67	2.325	328.283	9500.508
Cosine distance over over summary W2V	0.886	6.35	1.258	215.654	NaN
Cosine distance over over NCBI summary W2V	3.384	54.313	7.583	737.154	101.634
Cosine distance over gene symbols W2V	0.202	0.761	0.285	17.467	1.483
Cosine distance GO MF description W2V	9.172	27.654	24.075	2348.945	206.04
Cosine distance GO CC description W2V	5.542	52.808	28.758	2191.575	135.862
Cosine distance GO BP description W2V	11.603	154.268	124.624	20266.078	361.471
GO-distance (go_type=MF, measure=Wang, combine=BMA)	11.884	2588.84	301.245	2601.024	1954.441
GO-distance (go_type=CC, measure=Wang, combine=BMA)	5.406	72.14	502.832	3835.697	5870.326
GO-distance (go_type=BP, measure=Wang, combine=BMA)	24.084	826.678	1598.533	77040.866	48340.111
Overlap distance over genes	0.239	3.488	0.7	281.89	11.426
Overlap distance over gene traits	0.22	6.589	1.77	326.871	6.641
Minkowski distance (p=2) over genes	0.001	0.003	0.001	0.221	0.016
Minkowski distance (p=2) over gene traits	0.001	0.007	0.003	0.276	0.013
Minkowski distance (p=2) over gene trait frequency	0.001	0.018	0.006	0.327	0.192
Minkowski distance (p=1) over genes	0.0	0.003	0.001	0.161	0.016
Minkowski distance (p=1) over gene traits	0.0	0.007	0.003	0.204	0.014
Minkowski distance (p=1) over gene trait frequency	0.001	0.016	0.005	0.279	0.192
Kappa distance over genes	0.326	1.407	0.485	38.138	0.97
Kappa distance over gene traits	0.524	1.664	0.536	41.809	0.638
Jaccard distance over genes	0.0	0.001	0.001	0.053	0.008
Jaccard distance over gene traits	0.0	0.004	0.002	0.066	0.009
Cosine distance over gene trait frequency	0.001	0.013	0.004	0.115	0.197
Random (uniform, (0,1))	0.012	0.031	0.016	0.129	0.014

Figure A.1.: Execution time per metric and per entire reference data in seconds



	R-HSA-8982491	R-HSA-1474290	R-HSA-373755	R-HSA-422475	immune_only
Pairwise path length in reference tree	0.0	0.0	0.0	0.0	0.0
Jaccard distance over extended gene set	0.0	0.0	0.0	0.0	0.0
WM distance over summary W2V	0.0	0.0	0.0	0.0	NaN
WM distance over gene symbols W2V	0.0	0.0	0.0	0.0	0.0
Cosine distance over over summary W2V	0.0011	0.0	0.0001	0.0	NaN
Cosine distance over over NCBI summary W2V	0.003	0.0	0.0	0.0	0.0
Cosine distance over gene symbols W2V	0.0	0.0	0.0	0.0	0.0
Cosine distance GO MF description W2V	0.0	0.0	0.0	0.0	0.0
Cosine distance GO CC description W2V	0.0	0.0	0.0	0.0	0.0194
Cosine distance GO BP description W2V	0.7905	0.0	0.0	0.0	0.0
GO-distance (go_type=MF, measure=Wang, combine=BMA)	0.0	0.0	0.0	0.0	0.0
GO-distance (go_type=CC, measure=Wang, combine=BMA)	0.0001	0.0	0.0	0.0	0.0008
GO-distance (go_type=BP, measure=Wang, combine=BMA)	0.0002	0.0	0.0	0.0	0.0
Overlap distance over genes	0.0	0.0	0.0	0.0	0.0
Overlap distance over gene traits	0.0	0.0	0.0	0.0	0.0
Minkowski distance (p=2) over genes	0.0021	0.0106	0.008	0.0	0.0
Minkowski distance (p=2) over gene traits	0.816	0.9003	0.5081	0.0	0.0
Minkowski distance (p=2) over gene trait frequency	0.005	0.0001	0.0	0.0	0.0
Minkowski distance (p=1) over genes	0.0	0.0	0.0	0.0	0.0
Minkowski distance (p=1) over gene traits	0.0443	0.0428	0.0094	0.0	0.0
Minkowski distance (p=1) over gene trait frequency	0.0012	0.0001	0.0	0.0	0.0
Kappa distance over genes	0.0	0.0	0.0	0.0	0.0
Kappa distance over gene traits	0.0	0.0	0.0	0.0	0.0
Jaccard distance over genes	0.0	0.0	0.0	0.0	0.0
Jaccard distance over gene traits	0.0	0.0	0.0	0.0	0.0
Cosine distance over gene trait frequency	0.0	0.0	0.0	0.0	0.0
Random (uniform, (0,1))	0.8206	0.647	0.4463	0.2574	0.5366

Figure A.2.: P-values of Pearson correlations coefficients in fig. 4.1



	R-HSA-8982491	R-HSA-1474290	R-HSA-373755	R-HSA-422475	immune_only
Pairwise path length in reference tree	0.0	0.0	0.0	0.0	0.0
Jaccard distance over extended gene set	0.0	0.0	0.0	0.0	0.0
WM distance over summary W2V	0.0052	0.0	0.0	0.0	NaN
WM distance over gene symbols W2V	0.0	0.0	0.0	0.0	0.0
Cosine distance over over summary W2V	0.0027	0.0	0.0005	0.0	NaN
Cosine distance over over NCBI summary W2V	0.0005	0.0	0.0	0.0	0.0
Cosine distance over gene symbols W2V	0.0	0.0	0.0	0.0	0.0
Cosine distance GO MF description W2V	0.0	0.0	0.0	0.0	0.0
Cosine distance GO CC description W2V	0.0	0.0	0.0	0.0	0.0476
Cosine distance GO BP description W2V	0.2682	0.0	0.0	0.0	0.0
GO-distance (go_type=MF, measure=Wang, combine=BMA)	0.0	0.0	0.0	0.0	0.0004
GO-distance (go_type=CC, measure=Wang, combine=BMA)	0.0004	0.0	0.0	0.0	0.0029
GO-distance (go_type=BP, measure=Wang, combine=BMA)	0.0223	0.0	0.0	0.0	0.0
Overlap distance over genes	0.0	0.0	0.0	0.0	0.0
Overlap distance over gene traits	0.0	0.0	0.0	0.0	0.0
Minkowski distance (p=2) over genes	0.0061	0.4456	0.2867	0.0	0.0
Minkowski distance (p=2) over gene traits	0.9516	0.0091	0.6076	0.0	0.0
Minkowski distance (p=2) over gene trait frequency	0.1526	0.8201	0.0015	0.0	0.0
Minkowski distance (p=1) over genes	0.0061	0.4456	0.2867	0.0	0.0
Minkowski distance (p=1) over gene traits	0.9516	0.0091	0.6076	0.0	0.0
Minkowski distance (p=1) over gene trait frequency	0.408	0.3746	0.024	0.0	0.0
Kappa distance over genes	0.0	0.0	0.0	0.0	0.0
Kappa distance over gene traits	0.0	0.0	0.0	0.0	0.0133
Jaccard distance over genes	0.0	0.0	0.0	0.0	0.0
Jaccard distance over gene traits	0.0	0.0	0.0	0.0	0.0
Cosine distance over gene trait frequency	0.0	0.0	0.0	0.0	0.0
Random (uniform, (0,1))	0.7903	0.6997	0.4201	0.3821	0.5835

Figure A.3.: P-values of Spearman's Rank correlations coefficients in fig. 4.2

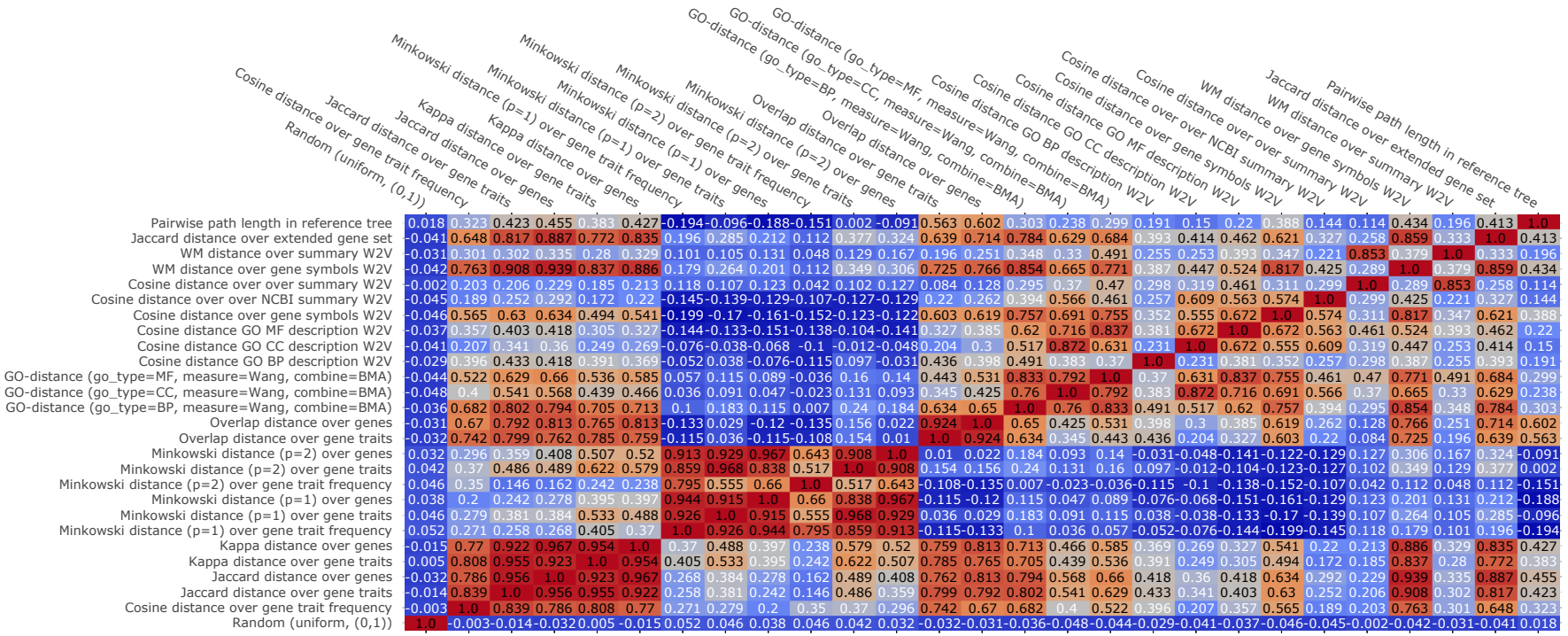


Figure A.4.: Pearson correlation summary over Reactome dataset R-HSA-373755

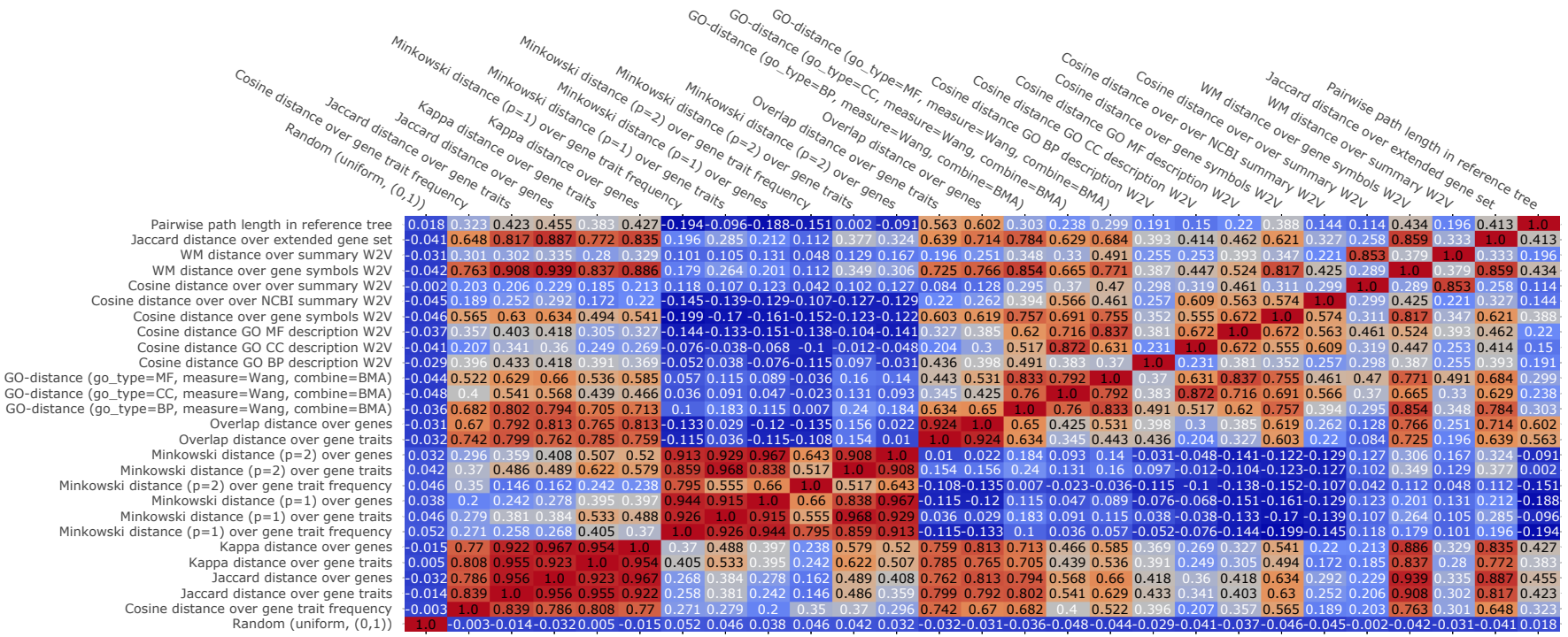


Figure A.5.: Spearman correlation summary over Reactome dataset R-HSA-373755

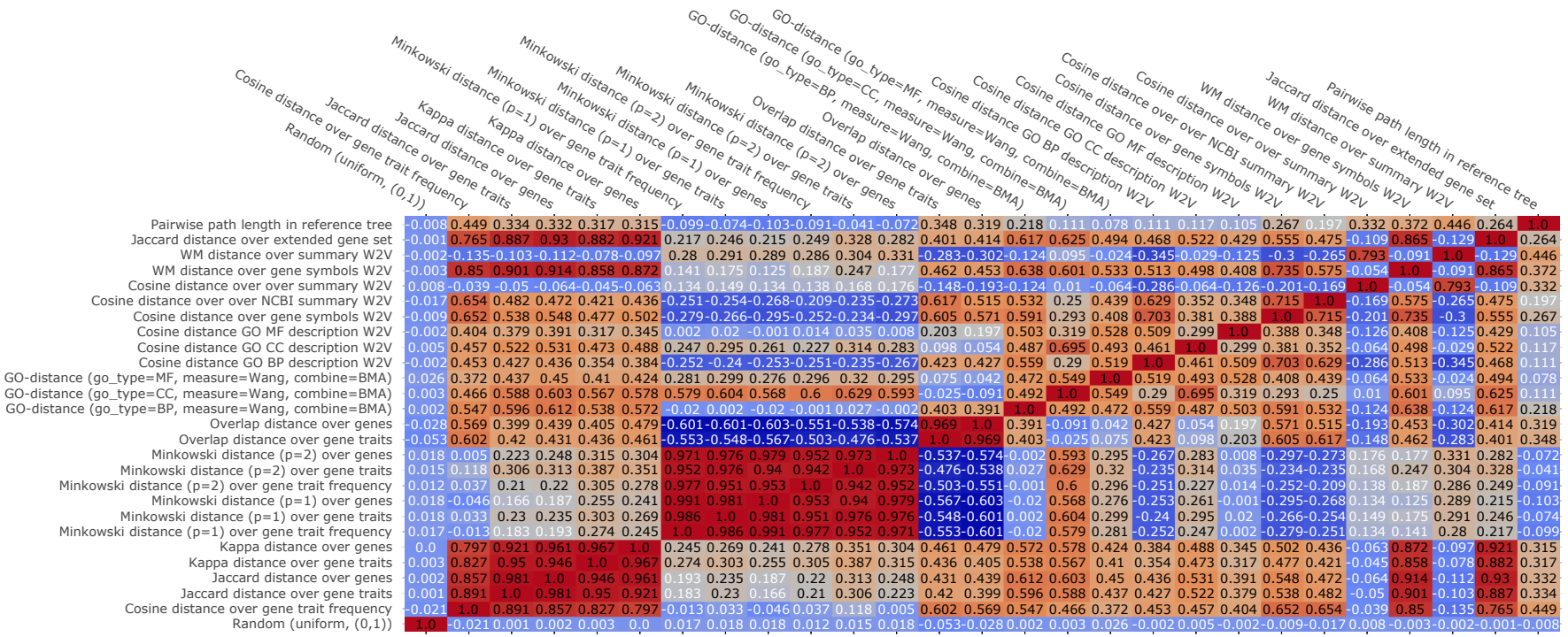


Figure A.6.: Pearson correlation summary over Reactome dataset R-HSA-1474290

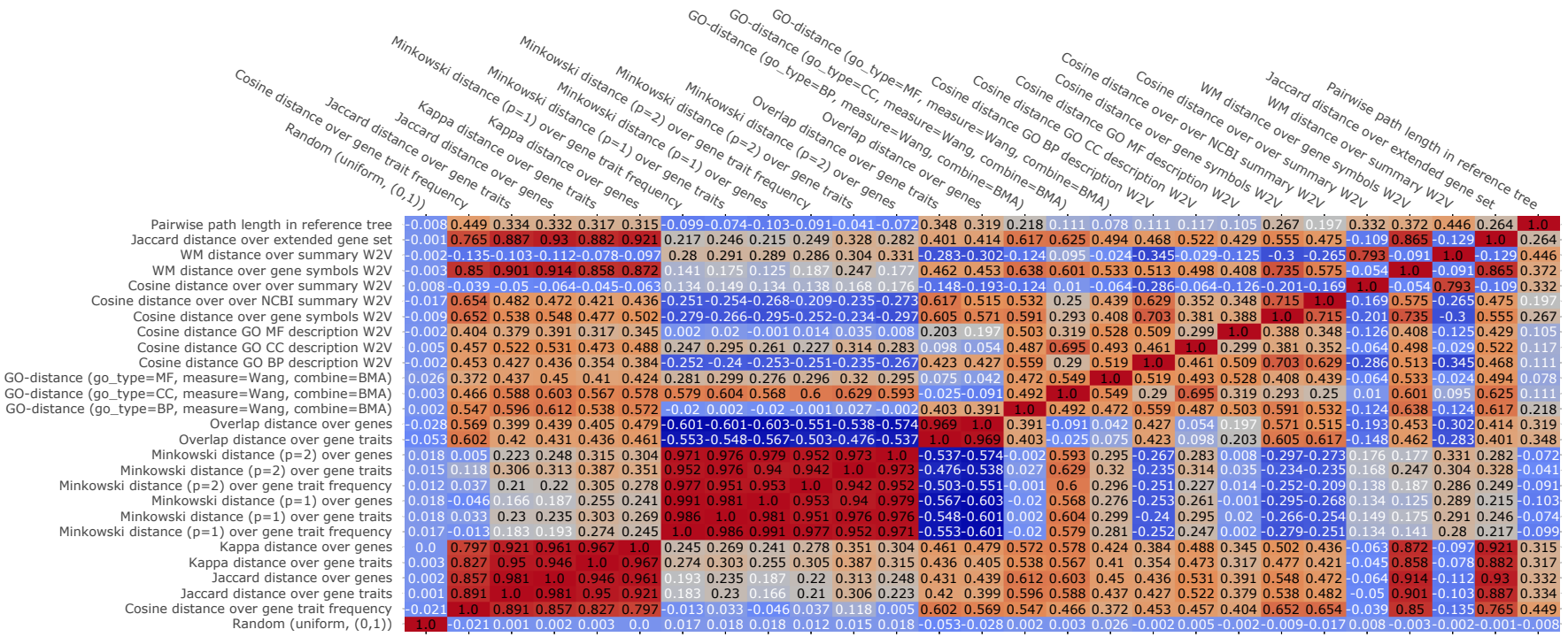


Figure A.7.: Spearman correlation summary over Reactome dataset R-HSA-1474290

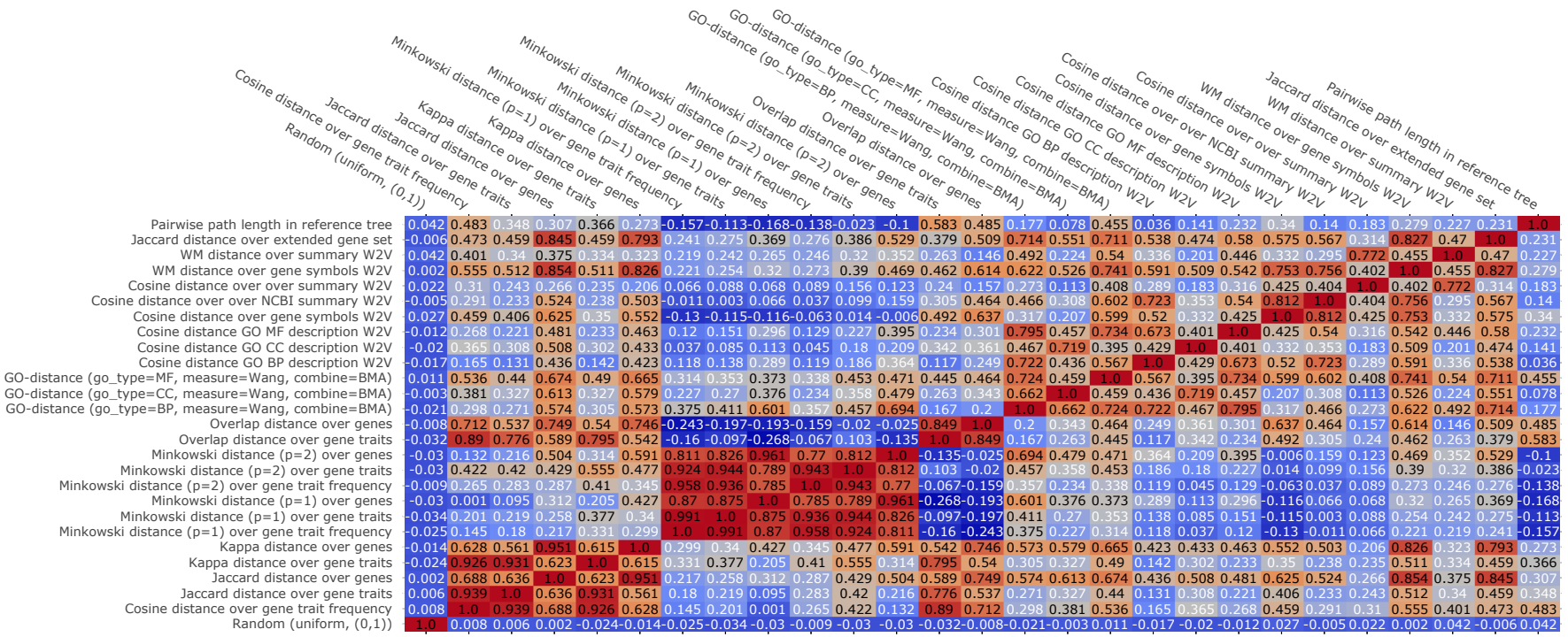


Figure A.8.: Pearson correlation summary over Reactome dataset R-HSA-8982491

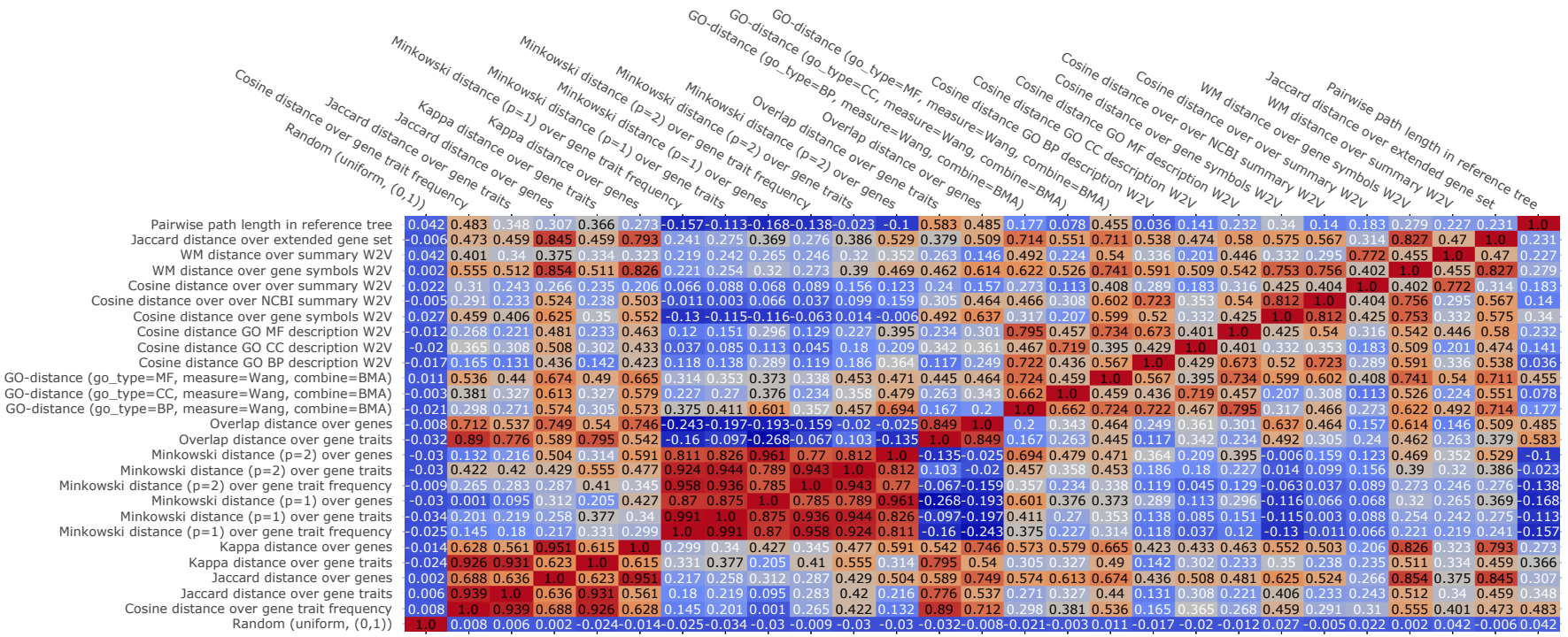


Figure A.9.: Spearman correlation summary over Reactome dataset R-HSA-8982491

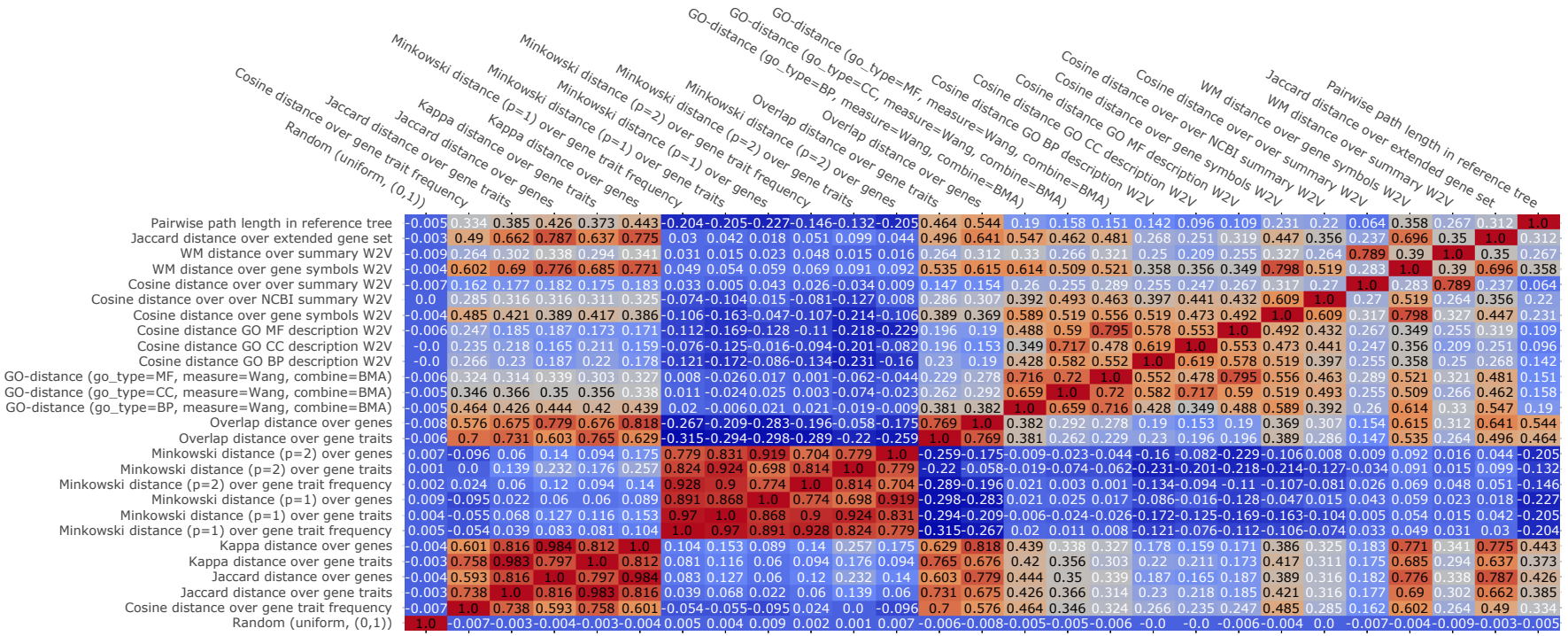


Figure A.10.: Pearson correlation summary over Reactome dataset R-HSA-422475

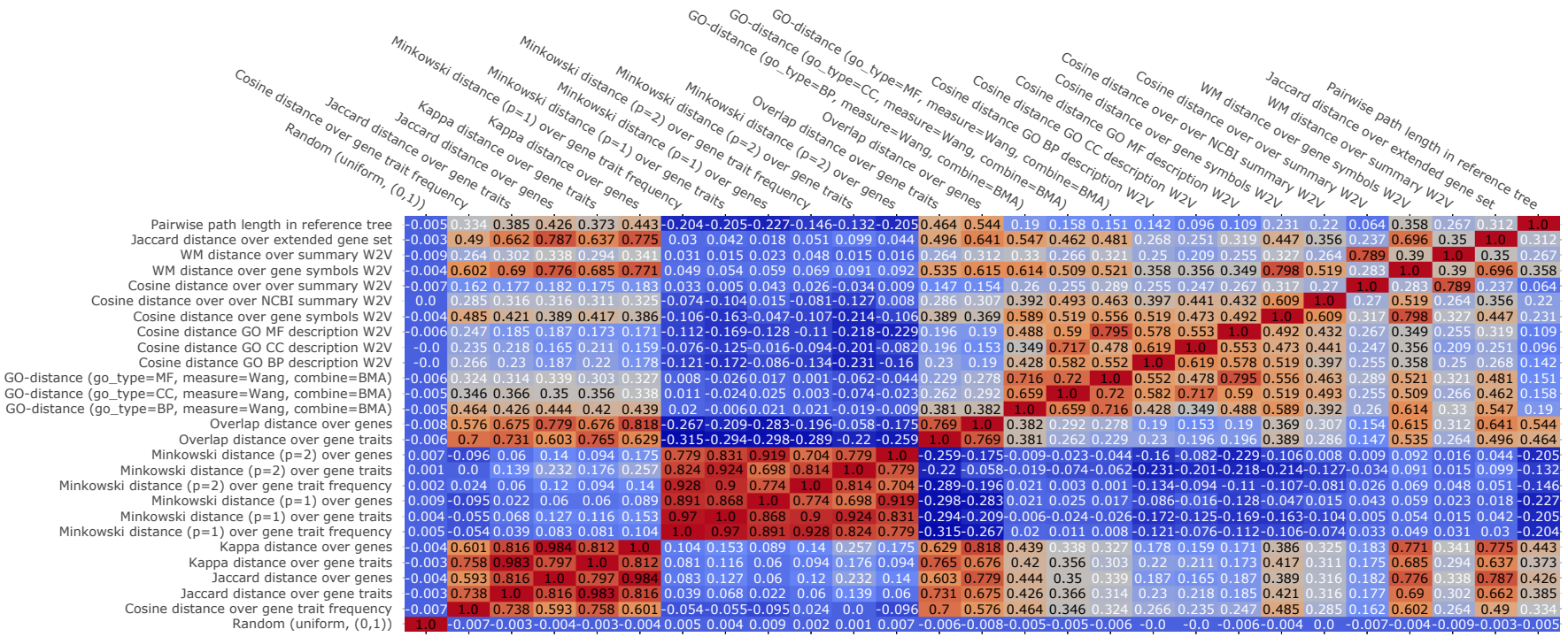


Figure A.11.: Spearman correlation summary over Reactome dataset R-HSA-422475

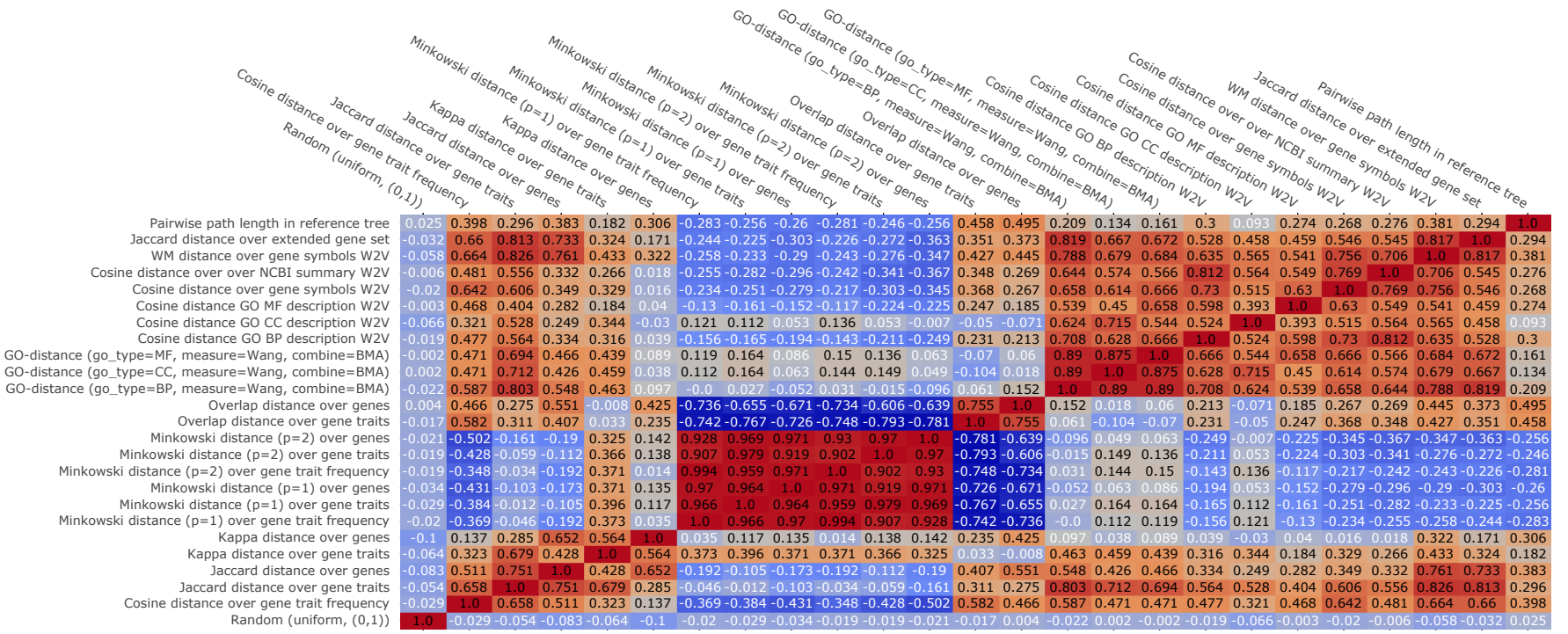


Figure A.12.: Pearson correlation summary over the immune cell reference tree

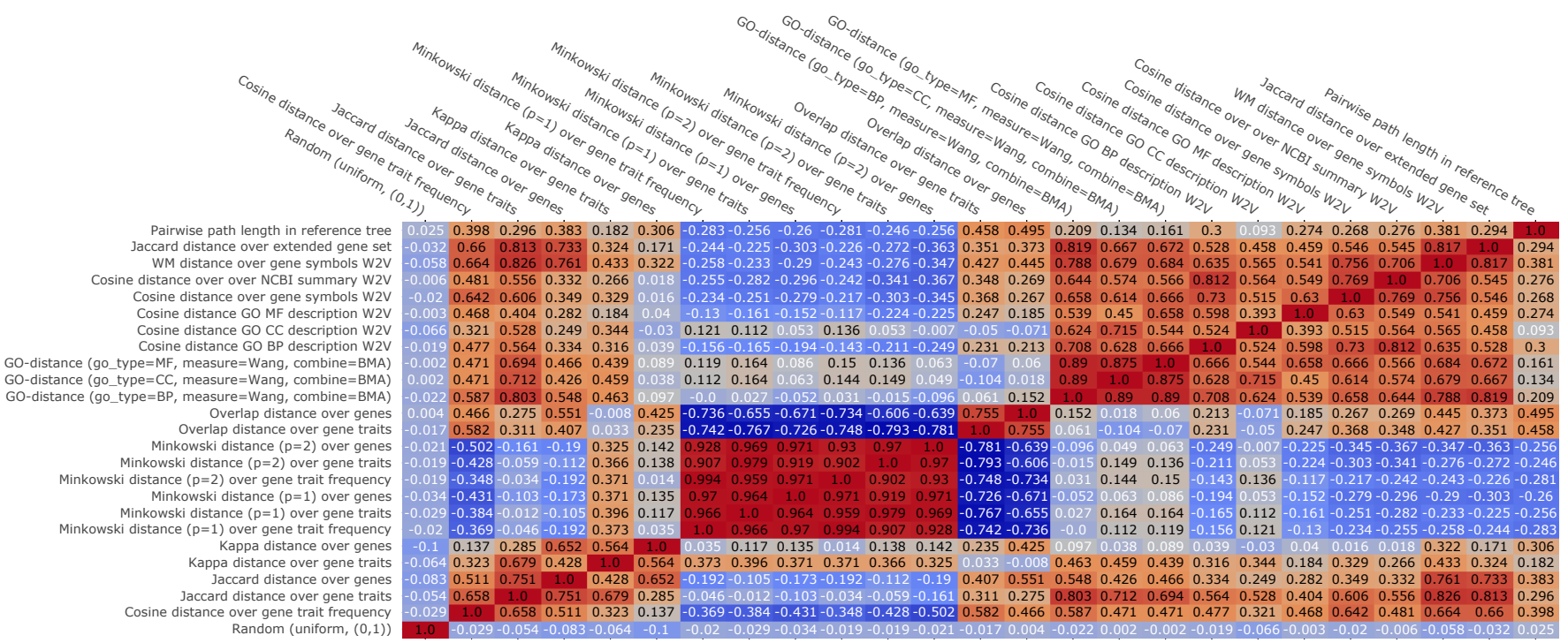


Figure A.13.: Spearman correlation summary over the immune cell reference tree

B. ROGER - Roche Omnibus of Gene Expression Regulation

B.1. State of the art

B.1.1. Transcriptomic data management

B.1.2. Differential Gene Expression Analysis

B.1.3. Gene Set Enrichment Analysis

B.2. Reimplementation

B.2.1. Data structures & architecture

B.2.2. Visualizations & data access

C. List of acronyms

API Application Program Interface. 7, 9

BMA Best Match Average. 13, 14

BOW Bag Of Words. 6

DAG Directed Acyclic Graphs. 13

DGE Differential gene expression. 4

DNA Deoxyribonucleic Acid. 1, 4

EFO Experimental Factor Ontology. 6

EMBL-EBI European Bioinformatics Institute. 5, 6

GO Gene Ontology. 5–9, 11–15

GSE Gene set enrichment. 4, 5

GWAS Genome-wide association studies. 6, 8, 9

NCBI National Center for Biotechnology Information. 8, 9, 11–15

NLP Natural Language Processing. 6

PCC Pearson Correlation Coefficient. 17, 18

PPI Protein Protein Interactions. 1, 2, 8, 11, 13–15

RNA Ribonucleic Acid. 1, 3, 4

RNA-Seq RNA sequencing. 3, 4

scRNA-Seq signle cell RNA sequencing. 3

SRCC Spearman's Rank Correlation Coefficient. 18

TSV Tab Separated Values. 9

WMD Word Mover Distance. 6, 15, 16

D. List of figures

E. List of tables

F. References

[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.

[AF90] A. V. Arkhangel'skiĭ and V. V. Fedorchuk. *The Basic Concepts and Constructions of General Topology*, chapter 1, pages 1–90. Springer Berlin Heidelberg, Berlin, Heidelberg, 1990.

[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.

[AP08] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December 2008.

[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.

[BEJP+04] Olaf R. P. Bininda-Emonds, Kate E. Jones, Samantha A. Price, Marcel Cardillo, Richard Grenyer, and Andy Purvis. *Garbage in, Garbage out*, chapter 12, pages 267–280. Springer Netherlands, Dordrecht, 2004.

[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.

[Car18] Marc Carlson. *org.Hs.eg.db: Genome wide annotation for Human*, 2018. R package version 3.7.0.

[CLRS09] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms, Third Edition*. The MIT Press, 3rd edition, 2009.

[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.

[HtrotSF+03] M. Hucka, , the rest of the SBML Forum:, A. Finney, , the rest of the SBML Forum:, H. M. Sauro, , the rest of the SBML Forum:, H. Bolouri, , the rest of the SBML Forum:, J. C. Doyle, , the rest of the SBML Forum:, H. Kitano, , the rest of the SBML Forum:, A. P. Arkin, , the rest of the

SBML Forum:, B. J. Bornstein, , the rest of the SBML Forum:, D. Bray, , the rest of the SBML Forum:, A. Cornish-Bowden, , the rest of the SBML Forum:, A. A. Cuellar, , the rest of the SBML Forum:, S. Dronov, , the rest of the SBML Forum:, E. D. Gilles, , the rest of the SBML Forum:, M. Ginkel, , the rest of the SBML Forum:, V. Gor, , the rest of the SBML Forum:, I. I. Goryanin, , the rest of the SBML Forum:, W. J. Hedley, , the rest of the SBML Forum:, T. C. Hodgman, , the rest of the SBML Forum:, J.-H. Hofmeyr, , the rest of the SBML Forum:, P. J. Hunter, , the rest of the SBML Forum:, N. S. Juty, , the rest of the SBML Forum:, J. L. Kasberger, , the rest of the SBML Forum:, A. Kremling, , the rest of the SBML Forum:, U. Kummer, , the rest of the SBML Forum:, N. Le Novère, , the rest of the SBML Forum:, L. M. Loew, , the rest of the SBML Forum:, D. Lucio, , the rest of the SBML Forum:, P. Mendes, , the rest of the SBML Forum:, E. Minch, , the rest of the SBML Forum:, E. D. Mjolsness, , the rest of the SBML Forum:, Y. Nakayama, , the rest of the SBML Forum:, M. R. Nelson, , the rest of the SBML Forum:, P. F. Nielsen, , the rest of the SBML Forum:, T. Sakurada, , the rest of the SBML Forum:, J. C. Schaff, , the rest of the SBML Forum:, B. E. Shapiro, , the rest of the SBML Forum:, T. S. Shimizu, , the rest of the SBML Forum:, H. D. Spence, , the rest of the SBML Forum:, J. Stelling, , the rest of the SBML Forum:, K. Takahashi, , the rest of the SBML Forum:, M. Tomita, , the rest of the SBML Forum:, J. Wagner, , the rest of the SBML Forum:, J. Wang, , and the rest of the SBML Forum:. The systems biology markup language (sbml): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531, 2003.

[JOP+] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed <today>].

[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.

[KFVSL+17] Weaver Kathleen F., Morales Vanessa, Dunn Sarah L., Godde Kanya, and Weaver Pablo F. *Pearson's and Spearman's Correlation*, chapter 10, pages 435–471. John Wiley & Sons, Ltd, 2017.

[KFWL17] Sun Kim, Nicolas Fiorini, W. John Wilbur, and Zhiyong Lu. Bridging the gap: Incorporating a semantic similarity measure for effectively mapping pubmed queries to documents. *Journal of Biomedical Informatics*, 75:122 – 127, 2017.

[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.

[LB02] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1*, ETMTNLP '02, pages 63–70, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[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.

[MBC+17] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, Zoe May Pendlington, Danielle Welter, Tony Burdett, Lucia Hindorff, Paul Flicek, Fiona Cunningham, and Helen Parkinson. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). *Nucleic Acids Research*, 45(D1):D896–D901, 2017.

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

[Men09] Elliott Mendelson. *Introduction to Mathematical Logic*. Chapman & Hall/CRC, 5th edition, 2009.

[MKK16] Vijaymeena M K and Kavitha K. A survey on similarity measures in text mining. *Machine Learning and Applications: An International Journal*, 3:19–28, 03 2016.

[MOPT07] Donna Maglott, Jim Ostell, Kim D. Pruitt, and Tatiana Tatusova. Entrez gene: gene-centered information at ncbi. *Nucleic Acids Research*, 35(suppl_1):D26–D31, 2007.

[PFB+08] Catia Pesquita, Daniel Faria, Hugo Bastos, António Ferreira, André Falcão, and Francisco Couto. Metrics for go based protein semantic similarity: A systematic evaluation. *BMC bioinformatics*, 9 Suppl 5:S4, 02 2008.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[Res99] Philip Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Int. Res.*, 11(1):95–130, July 1999.

[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.

[ŘS10] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[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.

[The17a] The Gene Ontology Consortium. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, 2017.

[The17b] The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 45(D1):D158–D169, 2017.

[Tre06] F. Treves. *Topological Vector Spaces, Distributions and Kernels*. Dover books on mathematics. Dover Publications, 2006.

[WDP+07] James Z. Wang, Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, March 2007.

[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.

[YY98] Gang Yu and Jian Yang. On the robust shortest path problem. *Computers & Operations Research*, 25(6):457 – 468, 1998.

[ZAA+18] Daniel R Zerbino, Premanand Achuthan, Wasiu Akanni, M Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, Carla Cummins, Astrid Gall, Carlos García Girón, Laurent Gil, Leo Gordon, Leanne Haggerty, Erin Haskell, Thibaut Hourlier, Osagie G Izuogu, Sophie H Janacek, Thomas Juettemann, Jimmy Kiang To, Matthew R Laird, Ilias Lavidas, Zhicheng Liu, Jane E Loveland, Thomas Maurel, William McLaren, Benjamin Moore, Jonathan Mudge, Daniel N Murphy, Victoria Newman, Michael Nuhn, Denye Ogeh, Chuang Kee Ong, Anne Parker, Mateus Patricio, Harpreet Singh Riat, Helen Schuilenburg, Dan Sheppard, Helen Sparrow, Kieron Taylor, Anja Thormann, Alessandro Vullo, Brandon Walts, Amonida Zadissa, Adam Frankish, Sarah E Hunt, Myrto Kostadima, Nicholas Langridge, Fergal J Martin, Matthieu Muffato, Emily Perry, Magali Ruffier, Dan M Staines, Stephen J Trevanion, Bronwen L Aken, Fiona Cunningham, Andrew Yates, and Paul Flicek. Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761, 2018.