Hidden Functional Activity in JOANA

by

Tim Diedrich

Supervisor: Prof. Dr. Florian Büttner

Time frame: July 5th, 2021 – September 20th, 2021

Goethe University

Frankfurt am Main

Faculty: Computer Science

German Cancer Research Center

Table of Contents

Abstract

In gene set analysis, the goal is to accurately identify gene sets that are significantly altered in their annotated gene expressions, for example, in diseased cells compared to healthy cells. While there are many different methods available, such as Gene Set Enrichment Analysis, few include the hierarchical structure of gene sets. Joint continuous ontology enrichment analysis (JOANA) based on a Bayesian network, as proposed in multi-level ontology analysis (MONA), approaches these hierarchy properties in the overlapping gene sets. In this work, we investigate the behavior of JOANA regarding the hierarchical properties and compare JOANA to a gene set analysis method that addresses the importance of the overlap between gene sets rather than hierarchy by diminishing the significance of overlapping genes (PADOG). Two of the analyzed datasets show specifically that JOANA turns on higher-order gene sets and leaves smaller gene sets with a significant overlap to the higher-order term off, reducing the duplicate influence of significantly enriched genes. The average node degree of the JOANA-active terms is lower than the node degree of the PADOG-active terms, indicating that a significant overlap of multiple terms leads to less active terms in highly interconnected areas in JOANA while overlap significance in PADOG is limited. This behavior hinders a straightforward benchmark comparison since JOANA will select fewer higher-order gene sets in closely connected significant gene sets and therefore fail to identify smaller gene sets, resulting in lower scores according to MalaCards benchmarking.

Introduction

When researchers investigate the functional difference in two phenotypes of comparison, e.g., diseased cells and healthy cells, they commonly use differential expression analysis. This expression analysis is then used in gene set enrichment analysis to identify significantly enriched gene sets. This helps to detect diseased cells, such as cancer cells. Identifying the type of disease with the help of gene analysis can also help to focus further medical examinations on the supposed disease, or perhaps hint at diseases that show no significant symptoms but can stay under medical observation with possible early treatment.

Gene ontology enrichment analysis methods can be split into three major categories: over-representation analysis (ORA), functional class scoring (FCS), and pathway topology-based methods [1]. The goal of all gene set analysis methods is to gain insight into how different phenotypes are expressed in gene behavior.

Methods based on ORA utilize all genes as if they were independent of one another and equally effective and relevant in biological processes. Gene independence eases the modeling of the analysis method. As a result, there are many tools available that are based on ORA. However, actual biological interactions are disregarded entirely in methods that are based on gene independence. These biological interactions comprise gene-to-gene interaction, protein interaction, and more. Not only do they interact in paired processes, but they also form whole interactions of multiple processes with several genes, for example [2].

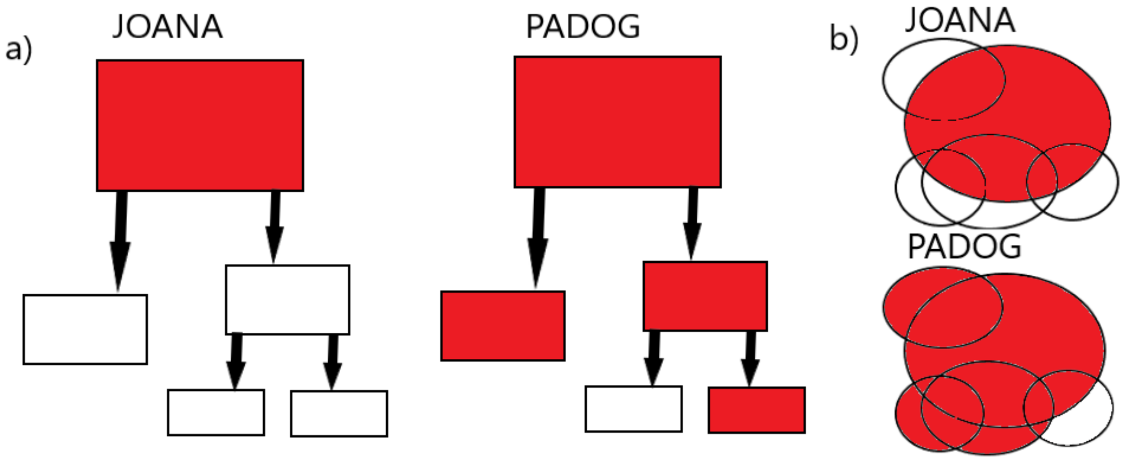


Figure 1. a) JOANA activates (in red) gene sets of higher hierarchy order while other gene set analysis methods such as PADOG mark child terms as active, too. b) Gene sets overlap and shared genes, JOANA avoids redundant use of significant genes by only activating higher-order gene sets compared to PADOG activating multiple gene sets.

Functional class scoring methods aim to use gene expression data without the invalid interaction assumption that all genes are independent. The most popular method is Gene Set Enrichment Analysis (GSEA). GSEA calculates a gene score for all genes and ranks them. A gene set score, also enrichment score (ES), is then calculated with these gene scores associated to the gene set [1]. FCS methods tend to lack specificity, as there are different gene set databases available, such as GeneSigDB and GeneSetDB, and there's no complete consensus on which genes are annotated to which gene set [1]. A widely used database is the Gene Ontology (GO) database [3], which represents biological terms that belong together. GO further describes hierarchical connections between terms.

Knowledge of pathway topology and the relevance of genes can help to produce more significant analysis results, as not all genes are equally important regarding a specific phenotype. A pathway as such is a description of a biological process or phenomenon. Gene sets or terms describe biological pathways, for example, for molecular interaction or human diseases. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database [4] is an example of a collection of pathway topology datasets known to be of a specific phenotype or disease.

One method that makes explicit use of pathway topology is introduced by Tarca et al. as an approach to down-weight genes that have an overlap with multiple gene sets called Pathway Analysis with Down-weighting of Overlapping Genes (PADOG) [5].

While many methods are based on Fisher's exact test or gene set enrichment, they do not include the hierarchy properties of GO [6]. However, these hierarchy properties provide valuable insight into the effect of gene overlap of GO terms. As genes can occur in multiple gene sets/terms, they impact the analysis of all terms that contain them. This redundant influence of genes in multiple gene sets impairs the specificity of the results, leading to many significant gene sets due to a few overlapping significant genes. Interpretability of results with many active gene sets is difficult since the overlap of few significantly enriched genes in a gene set may cause a lot of active terms and analyzing which of the terms of the many are relevant is challenging.

The lack of a gold standard in gene set analysis has been and still is one of the biggest challenges to overcome. So far, there has not been any consensus on best practice, e.g., there is no method that is superior in most cases. The absence of gold standard datasets contributes a major part to this problem [1]. Evaluation with a standard dataset would make gene set enrichment analysis methods comparable, but such a dataset depends on ground truth information of which gene sets are enriched. The ground truth information is not available since it can be obtained only through the output of a gene set enrichment analysis method. This is similar to calibrating weights with calibration weights that are not exactly calibrated themselves.

The joint continuous ontology enrichment analysis (JOANA) [7] is the successor method of a modular framework for gene set analysis integrating multilevel omics data (MONA) introduced by Sass et al. in 2013 [6]. JOANA is a work in progress by Kopf et al. and a gene set analysis method that uses a Bayesian network comparable to MONA.

The Bayesian network in JOANA implements the hierarchical properties of ontology terms in its directed acyclic graph (DAG), thus reducing the redundancy of single gene influence. We investigate the results of JOANA on real-world data to understand the behavior of JOANA given gene set hierarchy properties, i.e., which gene sets will be predicted as active. The main hypothesis investigated states that JOANA will mark higher-order terms, i.e., gene sets that are bigger in size and contain more genes, instead of marking multiple gene sets as active that overlap significantly.

While newly developed methods typically claim to be an improvement to the original methods of ORA and GSEA, the claim is validated through selected datasets that underline the benefits of the new method without featuring ground truth values or simulated datasets, designed to demonstrate those benefits [8].

Objective benchmarking is fundamental when comparing the performance of different methods to one another. MalaCards benchmark scores [9] give every gene set a relevance score depending on the relevance of the gene sets for the phenotype represented in the dataset. One first step towards objective benchmarking or a gold standard is the combination of KEGG datasets with relevance score data of MalaCards. This benchmark collection contains ground truth from MalaCards and the pathway information from KEGG. Results of JOANA or PADOG can be assessed according to this collection.

This assumption introduced for JOANA is validated in this work, as the results hint at this behavior. Joint graphs of JOANA and PADOG for specific datasets underline how JOANA activates higher-order terms while ignoring possible relevant child terms. This is caused by the overlap of the child terms with the higher-order term that is big and significant enough, so their activity is added to the big term’s activity. A t-statistic is used to analyze the node degree of the separate resulting graphs of PADOG and JOANA. For investigated datasets that yield reasonable many active gene sets, the results show a clear tendency. The gene sets marked active by JOANA are significantly less connected, i.e., their average node degree is smaller than the average node degree of PADOG-only active gene sets.

Methods

Pathway Analysis with Down-weighting of Overlapping Genes

In 2012, Tarca et al. introduced a gene set analysis method called Pathway Analysis with Down-weighting of Overlapping Genes (PADOG) [5], which introduces gene weights. The weights augment the relevance of genes present in fewer gene sets, while weights of genes associated with many gene sets are in comparison smaller. The weight of gene g is determined by

$$w(g) = 1 + \sqrt{\frac{max(f)-f(g)}{max(f)-min(f)}}. \qquad (1)$$

$f(g)$ is the number of gene sets that contain gene g and can range from 1 to N_{GS} since a gene is present in at least one gene set and at most in all gene sets. The more $f(g)$ increases, the faster the associated weight drops. The gene set score is the average sum of weighted absolute t-statistic, whereas the t-scores are moderated t-scores $T(g)$ [10]. This is argued to yield more reliable t-statistics because large scores are prevented by reducing gene standard deviations. As a result, the score for gene set GS_i is determined by

$$S_0(GS_i) = \frac{1}{N(GS_i)} \sum_{g \in GS_i} |T(g)| \cdot w(g). \qquad (2)$$

The standardized score S_0^* works as a threshold since the probability to observe an equally large or larger score is assessed by permuting the samples $N_{ite} = 1000$ times. This leads to

$$P_{PADOG}(GS_i) = \frac{\sum_{ite} I(S_{ite}^*(GS_i) \geq S_0^*(Gs_i))}{N_{ite}}. \qquad (3)$$

$S_{ite}^*(GS_i)$ is the standardized score from the ite-th iteration. $I(x)$ will return 1 if expression x is true. If x is evaluated as false, 0 will be returned. Therefore, $P_{PADOG}(GS_i)$ computes the percentage of permutated standardized gene set scores equal to or bigger than the first permutation.

Fisher's Exact Test

Fisher's exact test is a method of analysis for a 2 by 2 contingency table but can be used for bigger tables. It can be applied to an arbitrary number of samples but is often used with a small number of samples. Fisher's exact test checks the null hypothesis for independence between two categories using a hypergeometric probability distribution [11]. These categories can be genes in healthy cells vs. diseased cells, for example. To obtain the p-value all probabilities of a less or equally likely outcome, as the contingency table shows, are added. Only the probabilities with the same marginal frequencies are considered [12].

Since Fisher's exact test assesses the association between two categories or conditions it is a perfect first approach to assess if a gene set is significantly enriched or not. For every gene set, the number of genes that are significantly enriched and annotated in that gene set as well as the number of genes in that gene set that are not significantly enriched are determined. Opposed to that are the number of genes significantly enriched and the number of genes not significantly enriched, which are not annotated in the gene set. Genes with a q-value smaller than 0.1 are considered significantly enriched. The contingency table is then $\begin{pmatrix} A = \#\, significant\ genes\ in\ gene\ set & B = \#\, significant\ genes\ not\ in\ gene\ set \\ C = \#\, insignificant\ genes\ in\ gene\ set & D = \#\, insignificant\ genes\ not\ in\ gene\ set \end{pmatrix}$.

The probability of this outcome is calculated by $\frac{(A+B)!*(C+D)!*(A+C)!*(B+D)!}{(A+B+C+D)!*A!*B!*C!*D!} = p_F$. A gene set will then be marked active if the added probabilities of all possible contingency table combinations, that use the same row and column count as the probability present in the data, with a probability less than p_F, is less than 0.05.

While this approach analyzes all gene sets independently, leaving out any potential information of gene set overlap or pathway topology, it is an exact method for independent data. As such it is a basis for methods that do not take hierarchical properties into account [6].

Beta Distribution

The beta probability distribution over $[0, 1]$ is defined as

$$Beta(x|\alpha,\beta) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \qquad (4)$$

with $B(\alpha,\beta) \triangleq \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ describing the beta function. $\Gamma(\alpha)$ is then defined by

$$\Gamma(\alpha) \triangleq \int_0^\infty x^{\alpha-1}e^{-x}\,dx. \qquad (5)$$

The required lower boundary for its parameters is $\alpha, \beta > 0$, as $B(\alpha, \beta)$ would be 0 otherwise. A uniform distribution can be obtained by setting $\alpha = \beta = 1$ [13].

Since the q-values, i.e., the gene expressions that JOANA uses, can range from 0 to 1, beta distributions are a practical choice to model the distribution of the q-values and a variety of different bioinformatics applications make use of this [14]. The maximum likelihood parameter estimation is not a possible option to fit the function. Values of 0 or 1 can lead to singularities in the log-likelihood function which makes it an unstable approach because the log-likelihood function is not defined for q-values that equal 0 if $\alpha \neq 1$ or for q-values equal to 1 if $\beta \neq 1$ [15]. JOANA uses a mixture of beta distributions to fit the q-values. However, because maximum likelihood parameter estimation does not work properly for a single beta distribution, it is problematic for a mixture of beta distributions, too. Schröder and Rahmann introduce a moment-fitting method that is used in JOANA, with a more stable convergence [7, 15].

Joint Continuous Ontology Enrichment Analysis

The following chapters describe the model, which this thesis analyzes. Joint Continuous Ontology Enrichment Analysis (JOANA) [7] is the successor model of MONA, a modular framework for gene set analysis integrating multilevel omics data introduced by Sass/Buettner et al. in 2013 [6]. With a Bayesian Network as its core and Expectation Propagation to infer hyper-parameters, JOANA and MONA come fitted with the option to analyze more than one species simultaneously with the cooperative model.

Base Model

The base model can be described as a Bayesian network consisting of two layers: the ontology terms layer, followed by a gene response hidden layer (*Figure 2*). The ontology terms layer has one node for each ontology term. The nodes follow a Bernoulli distribution, depicting whether a term is marked as active or inactive. The gene response hidden layer holds one node for each gene. The nodes from the ontology term layer are connected to the gene response layer as described by the ontology under investigation, which could be, for example, GO or KEGG. This means that a term has a set of genes annotated to itself, and the corresponding term layer node is connected to the gene response layer nodes in that set.

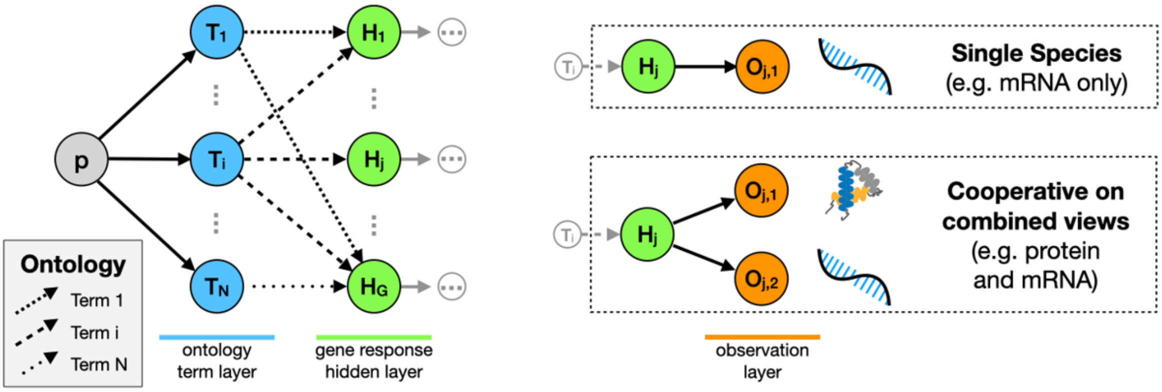
Every term T_i with a prior probability $p \sim Beta(a, b)$ of being active is modeled by a Bernoulli-distributed random variable. The initially unknown value of p is learned from the data through a Beta prior distribution as described in chapter Beta Distribution above.

The hidden nodes H_i are also modeled as boolean variables. Their values are specified through the values of the preceding ontology term nodes that are connected to each node H_j. Stating that $T(H_j)$ defines the terms connected to node H_j, the values for the gene response hidden layer are determined by

$$P(H_j|T) = \begin{cases} 1, & \text{if } \exists T_i \in T(H_i) : T_i = 1 \\ 0, & \text{otherwise} \end{cases}. \tag{6}$$

Therefore, $H_j = 1$ if one or more terms annotated to it are active. If all terms are inactive, the gene response value is 0.

The necessity of the hidden gene response layer is due to possible false positives (FP) and false negatives (FN), stemming from measurement errors. Term-gene assignment may be incorrect or not precise, which is a second reason supporting the need for a hidden layer. Conclusively, this layer helps to model a “coherent integration of biological observations across multiple layers” [6, p.2].



p
T1
Ti
TN
H1
Hj
HG
Ontology
Term 1
Term i
Term N
ontology term layer
gene response hidden layer
Ti
Hj
Oj,1
Single Species (e.g. mRNA only)
Ti
Hj
Oj,1
Oj,2
Cooperative on combined views (e.g. protein and mRNA)
observation layer

Figure 2. Bayesian Network in JOANA. The connections of the ontology term layer ($T_1, \dots, T_N$) to the gene response hidden layer ($H_1, \dots, H_G$) are given by the assignment matrix that tells which gene sets are annotated to which genes. In the single-species model, there is one observation layer node $O_{j,1}$ for every gene response hidden layer node H_j [7].

Single-Species Model

While the Base Model explained above stems from MONA, JOANA differs slightly in the following aspects. For every node H_j, one observation node O_j is added, which is only connected to the corresponding hidden gene response node H_j. It is one node because only a single species will be observed. The probability of a node being on is defined by a mixture of three different Beta distributions defined as

$$P(O_j|H_j) = \begin{cases} Beta(\alpha_1, \beta_1), & \text{if } H_j = 1 \\ \mu_2 Beta(\alpha_2, \beta_2) + \mu_3 Beta(\alpha_3, \beta_3), & \text{if } H_j = 0 \end{cases}. \qquad (7)$$

For an inactive hidden layer $H_j = 0$, the observation layer node is determined by a mixture of two Beta distributions weighted by μ_2 and μ_3. During inference described later, these weights will be optimized. Instead of having just one false positive rate α and one false negative rate β as modeled in MONA [6], JOANA can use the overlap between $Beta(\alpha_1, \beta_1)$ and $Beta(\alpha_2, \beta_2)$ to model hidden-active features, which is not possible in MONA.

The first Beta distribution models all active genes in which $H_j = 1$. The second and third Beta distribution model the inactive genes. The mixture of Beta distributions is optimized using a moment-fitting approach instead of the Maximum Likelihood Estimation since its convergence shows more stability with q-values ranging from 0 to 1 [7].

Cooperative Model

Like the single-species model, the cooperative model observes the hidden gene responses with an extra layer. Depending on the number of species that are introduced to the model, two or more nodes can be inserted for a single hidden gene response to observe a corresponding species (e.g., mRNA or protein). The formula from the single-species model is then extended by the following formulae, one for each s

$$P\left(O_{j,s}|H_j\right) = \begin{cases} Beta\left(\alpha_{1,s}, \beta_{1,s}\right), & \text{if } H_j = 1 \\ \mu_2 Beta\left(\alpha_{2,s}, \beta_{2,s}\right) + \mu_3 Beta\left(\alpha_{3,s}, \beta_{3,s}\right), & \text{if } H_j = 0 \end{cases}, \tag{8}$$

whereas $s = 1 \dots S$ and S is the number of species to observe, which has been tested for two species but is applicable to more species. For all species together, a set of parameters $\left(\alpha_{c,s}, \beta_{c,s}\right)_{c=1\dots3, s=1\dots S}$ must be fitted.

Bayesian Inference

Having a less time-consuming method for inference plays a crucial part in determining the marginal posteriors. The method used with MONA is the expectation propagation (EP) algorithm as introduced by Minka [16]. "These marginal posterior probabilities […] can be interpreted as the outcome of the MONA algorithm in form of the probabilities for each term to be active as best explained by the data" [6, p.4]. JOANA uses the same approach to infer the marginal posteriors.

Including the factorized properties of the posterior, EP minimizes the local Kullback-Leibler divergence iteratively.

The posterior, with parameters $\theta = \{p,\ T,\ H, \alpha^I, \beta^I\}$ referring to the single-species model, factorizes as

$$p(T, H, p, \alpha, \beta | D) = \frac{p(T|p)p(D|H, \alpha, \beta)p(H,T)p(\alpha)p(\beta)p(p)}{p(D)}. \tag{9}$$

Instead of minimizing the Kullback-Leibler divergence with likely intricate factors, Minka proposes an algorithm to minimize local divergence between the factors and their Gaussian approximations [16]. This results in a Gaussian approximation q, as the Gaussian distribution is closed under multiplication.

Results

When comparing the output of JOANA and PADOG, a few significant datasets that have been analyzed give insight into how JOANA works. Examples of joint graphs of the results of JOANA and PADOG show that JOANA does activate higher-order gene sets, i.e., gene sets that are big in size. The smaller gene sets surrounding the big term are mostly marked as active by PADOG and not JOANA. An analysis of the node degree of the resulting active-term graph, separated by methods, i.e., JOANA and PADOG, provides further evidence of the hypothesized behavior (*Figure 1*) that JOANA activates less and rather higher-order gene sets, as overlapping active gene sets are added to the big term and are inactive in turn. Table 1 displays the results of a node degree analysis for numerous datasets along with the t-statistic comparison for PADOG-only and JOANA-only graphs. The statistic shows distinctively that if the percentage of JOANA-active gene sets to all gene sets is below 20%, the average node degree in JOANA is lower than in PADOG. That means the active terms in JOANA are less connected with one another and have little to no overlap. For the two datasets GSE14924_CD4 and GSE24739_G0, a detailed overlap analysis of a big term of interest with all other active terms of the joint graph (*Figure 3*, *Figure 4*) can be found in the supplementary material. The results of both datasets are explained in detail in the following chapters.

Dataset GSE14924_CD4

Dataset GSE14924_CD4 is characterized by the term hsa05168 (*Figure 3*). The size of the term hsa05168 indicates the gene set size; the number of genes the set comprises. The darker the green color of the term is, the more relevant the term is according to the MalaCards relevance score [9]. Gene sets not listed in the relevance score file are not relevant for the given phenotype, but when they are marked active their term color is white. Terms shaped like a circle are marked active by JOANA and PADOG. Diamond-shaped terms are marked active by PADOG only, and square-shaped terms are marked active by JOANA only. Most smaller terms surrounding the term hsa05168 are only marked active by PADOG and not active by JOANA. These terms are also highly interconnected, meaning a lot of the genes involved are annotated on multiple gene sets. The fact that the small gene sets are predominantly not active with JOANA, while the big term hsa05168 is JOANA-active, indicates how JOANA tends to turn on higher-order gene sets with a bigger gene set size and omits smaller gene sets to remove a redundant use of the same significantly enriched genes.

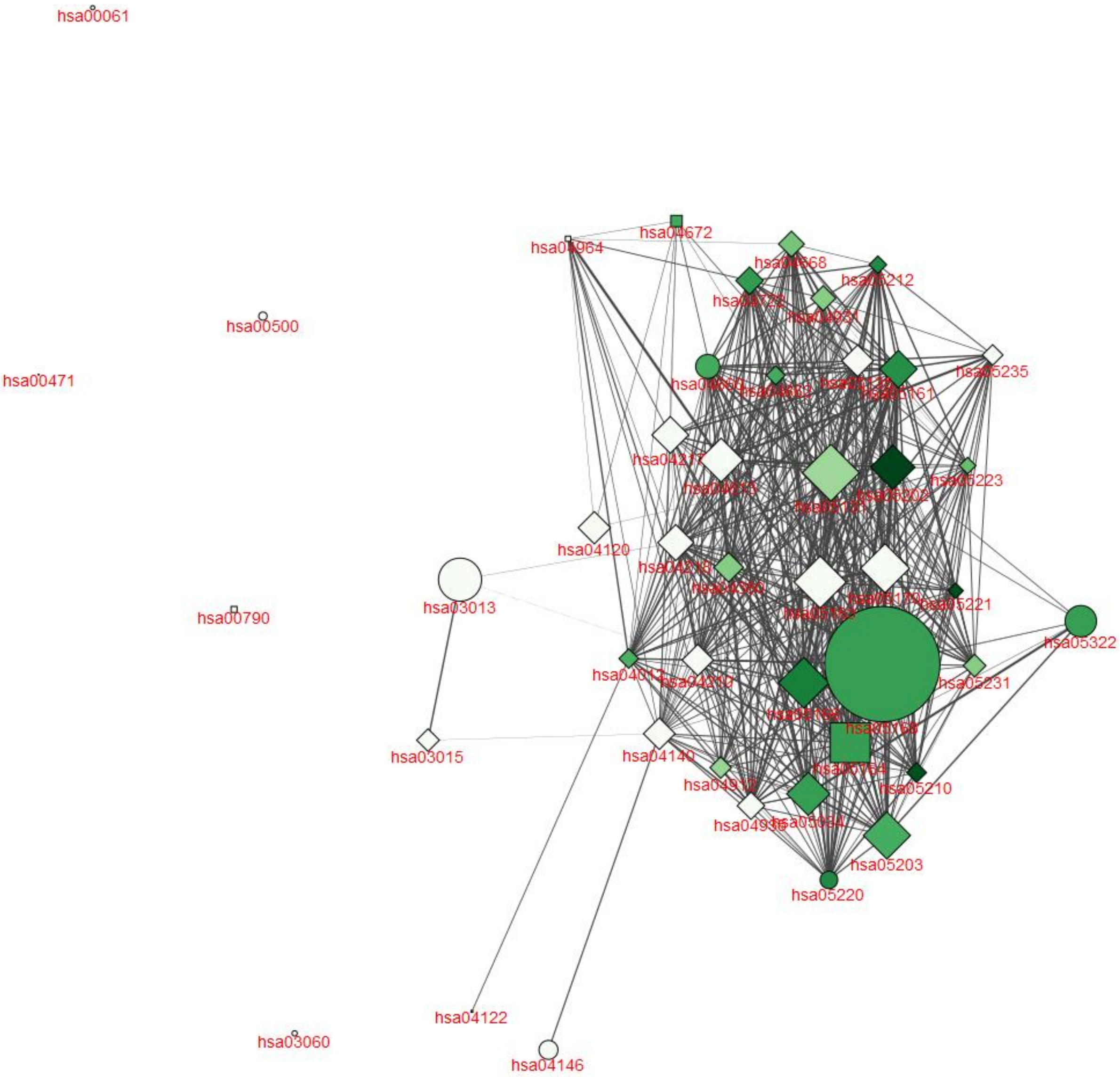


Figure 3. PADOG and JOANA joint active terms for dataset GSE14924_CD4. Node size displays gene set size, circle-shaped nodes are marked active by JOANA and PADOG, square-shaped nodes are marked active by JOANA only, diamond-shaped nodes are marked by PADOG only. The darker the green of a node is, the bigger the relevance score as determined by the Malacard-KEGG relevance for this dataset. The edges display the overlap of genes between gene sets, and the thickness of the edges represents the number of genes overlapping.

The gene overlap comparison of single terms to the big term hsa05168 yields the following results. Gene sets with a big overlap to hsa05168 compared to their set size tend to be affiliated to the big term. This means that JOANA does not mark them as active, because hsa05168 shares enough overlap and is marked active. However, if the remaining genes not included in the overlap have a large enough number of significantly enriched genes, the small term may still be marked as active on its own.

A detailed overlap comparison is provided in the supplementary material (data_archive/GSE14924_CD4/Overlap_Analysis.xlsx) Here, the overlap of every term active with PADOG and/or JOANA is compared to the big term hsa05168. All these comparisons return six numbers: The number of genes annotated in both gene sets that are significantly enriched, and the number of genes annotated in both gene sets that are not significantly enriched. For both gene sets, there is a number of genes significantly enriched and only present in one of the gene sets and a number of not enriched genes unique to the gene set concerning this single comparison.

A closer look at some terms compared with hsa05168 reveals how the underlying gene overlap could lead to the results. The hsa05131 term, marked active only by PADOG, is relatively big and has a significant overlap with hsa05168. 13 out of a total 41 overlapping genes between hsa05131 and hsa05168 have q-values smaller than 0.1 and are considered significantly enriched. The term hsa05202 has a smaller overlap with hsa05168, of which only two genes are significantly enriched, and is marked active by PADOG only. Similar to the term hsa05202, the term hsa05322 has a small overlap with hsa05168, but a much higher percentage of term-specific genes that are significantly enriched. This likely leads to the term hsa05322 being marked active by JOANA.

Dataset GSE24739_G0

The dataset GSE24739_G0 (*Figure 4*) also displays a term big in gene set size (hsa05206). Although there is one other big term (hsa05203), which is also relevant for this dataset, the analysis will be limited to hsa05206 only. Both of these big terms are marked in a dark green, meaning they are MalaCards-relevant terms. While the surrounding terms in this dataset are not marked active by PADOG predominantly, there are more terms only marked by PADOG than those only marked by JOANA in the highly interconnected part of the plot. Furthermore, the PADOG-only terms are on average rather small, compared to the JOANA-only terms and the terms marked active by both methods.

A closer look at the overlap of the big term hsa05206 and the term hsa05220 to the right of hsa05206, and the overlap of hsa05206 and hsa05230 above hsa05206 reveals the expected behavior of JOANA turning on the bigger term (hsa05206) if the overlap is big enough and contains enough significant genes.

The comparison of hsa05206 and hsa05220 results in 7 significant and 32 insignificant genes present in both. The remaining genes in gene set hsa05220 consist of 4 significant and 33 insignificant genes. The term hsa05220 is a PADOG-only term, meaning JOANA did not mark it as significantly enriched, even though it is a very relevant term according to MalaCards relevance score regarding the phenotype of this dataset.

However, gene set hsa05230 is marked as active by JOANA (and PADOG). The overlap comparison shows that a smaller overlap of 3 significant and 23 insignificant genes with hsa05206 and a larger hsa05230-unique proportion of 9 significant and 34 insignificant genes is not enough overlap or are enough self-significant genes for hsa05230 to be marked active by JOANA. Detailed overlap information is available in the supplementary material (data_archive/GSE24739_G0/Overlap_Analysis.xlsx).

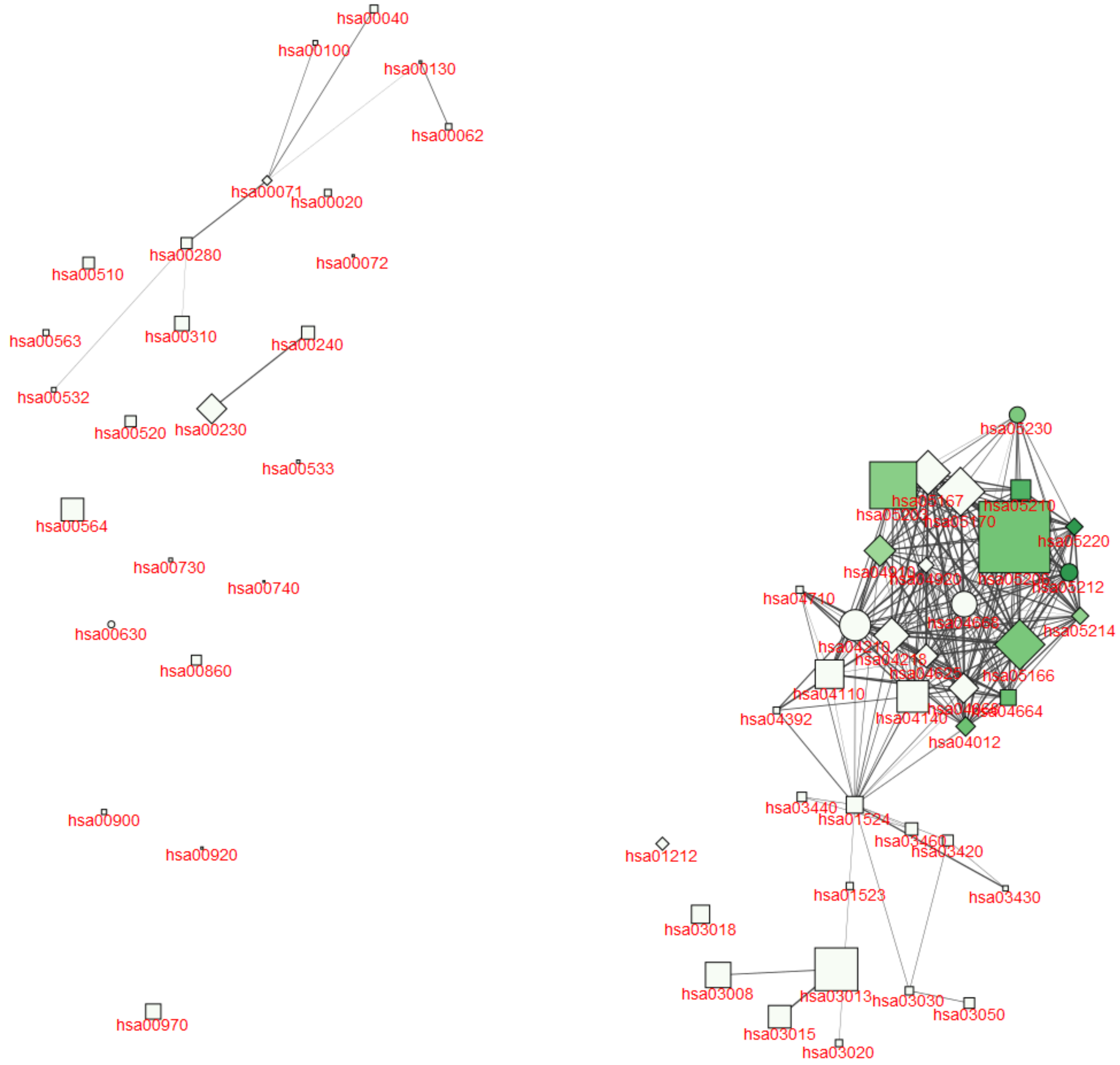


Figure 4. Joint active terms for JOANA and PADOG for dataset GSE24739_G0

Node Degree Comparison

When analyzing the interconnectivity of both methods, PADOG and JOANA, a cut-off point of 20% is introduced to put a threshold on the number of active JOANA terms. All datasets that exceed this threshold are considered invalid results. A total of 338 gene sets are introduced by KEGG [4]. It can be said that 70 active gene sets of all 338 gene sets, is a threshold above which clear results cannot be deduced since one fifth of all gene sets would be active. If most q-values are significantly enriched, JOANA will turn on most gene sets, which may provide an explanation as to why JOANA yielded many active gene sets.

For datasets where less than 70 terms are activated by JOANA, a node degree analysis for the active terms in JOANA and PADOG yields distinct results (*Table 1*).

Table 1. t-statistic for the node degrees between the active terms of JOANA and the active terms of PADOG. JOANA mean is the average node degree for JOANA-only terms. Active JOANA is the number of terms marked active by JOANA. The next two columns contain the same information for PADOG-only terms. The t-test analyzes the significant difference of the node degrees of JOANA and PADOG. The last two columns are the p-value and the corrected p-value for multiple testing.

Dataset	JOANA mean	active JOANA	PADOG mean	active Padog	t-test	pvalue	corrected pvalue
GSE3467	120.136095	338	8.6	11	5.09419425	5.77E-07	7.42E-07
GSE3678	35.5963303	109	28.6046512	43	2.16033707	3.23E-02	3.23E-02
GSE4107	103.965753	292	8.63636364	22	7.42819888	1.06E-12	4.77E-12
GSE4183	78.2323652	240	11.6363636	22	6.64178771	1.79E-10	3.58E-10
GSE8762	120.136095	338	1.71428571	8	4.52511889	8.33E-06	9.37E-06
GSE14924_CD4	2.71428571	15	26.75	40	-8.66598173	1.15E-11	2.59E-11
GSE16515	55.5915493	142	19.5333333	30	6.27714594	2.77E-09	4.53E-09
GSE16759	67.8518519	216	6.23529412	17	6.40415217	8.37E-10	1.51E-09
GSE19188	114.463855	332	6.61538462	14	5.81921567	1.37E-08	1.90E-08
GSE19420	3.83333333	12	10.5925926	27	-7.64657689	3.97E-09	5.96E-09
GSE19728	116.395155	322	16.64	25	7.47645009	6.64E-13	3.98E-12
GSE20153	33.6697248	109	17.2258065	32	4.27653004	3.52E-05	3.73E-05
GSE22780	94.1115538	251	10.8181818	22	7.43467554	1.38E-12	4.97E-12
GSE24739_G0	6.36363636	47	12.1052632	19	-5.46962187	8.91E-07	1.07E-06
GSE24739_G1	11.28	51	27.3333333	33	-10.1456995	4.40E-16	3.96E-15
GSE32676	1.4	16	25.3714286	36	-9.1757954	1.10E-11	2.59E-11
GSE38666_epithelia	119.112462	329	7.47368421	19	7.1159782	6.44E-12	1.93E-11
GSE38666_stroma	8.4	15	31.1666667	36	-17.6157013	7.69E-23	1.38E-21

The number of active terms in JOANA is lower in 4 out 6 cases where JOANA activated less than 20% of terms (marked green), and higher in the other two cases. But in all 6 cases, the average number of edges connected to a term is significantly lower in JOANA than the average number of edges connected to a term in PADOG. In dataset GSE14924_CD4 the average node degree in JOANA is a little more than 2.714 while the PADOG average node degree is 26.75. The JOANA-only graph compared to the PADOG-only graph displays this well, as the JOANA-only graph is sparsely connected (*Figure 5a*), and the PADOG-only graph is densely connected (*Figure 5b*). The GSE24739_G0 dataset shows the same tendency with 6.36 connected edges on average in JOANA (*Figure 5c*) and 12.11 in PADOG (*Figure 5d*), even though JOANA marked 47 terms active and PADOG marked 19 as active. The node degrees of the JOANA-only and PADOG-only graphs are compared with a t-statistic. The t-test results (*Table 1*) show that the difference of the node degrees is significant as all relevant (green) results have absolute values larger than 5 and the respective p-values are all well below 0.05, as are the corrected p-values. Since the t-test results for the green rows are all negative, the node degrees in JOANA are smaller than in PADOG.

As JOANA is expected to activate a higher-order gene set instead of multiple lower-order gene sets, which are highly likely to be connected to one another and therefore increase the node degree, the higher-order gene sets tend to overlap less with one another, decreasing the node degree. Even when the total number of active terms is greater in JOANA than in PADOG, the terms are more dispersed and have less gene overlap, i.e., less connected edges, while PADOG has dense areas, where a lot of active terms are connected to a lot of other active PADOG terms.

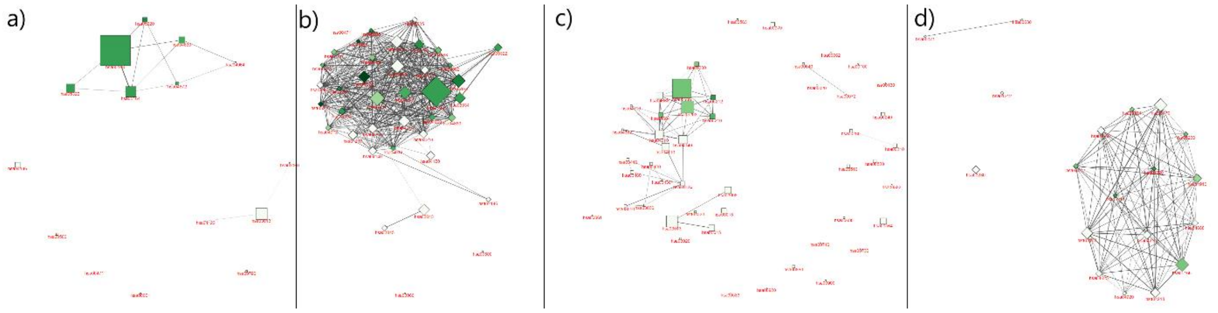


Figure 5. a) JOANA-only active terms for dataset GSE14924_CD4 shows few active terms and low node degrees. b) PADOG-only active terms for dataset GSE14924_CD4, many active terms with highly interconnected area and therefore high average node degree. c) Dataset GSE24739_G0, many JOANA-only active terms but low average node degree as nodes are sparsely connected. d) Dataset GSE24739_G0, less active PADOG-only terms, than JOANA-only terms in c) but high average node degree because of densely connected area.

Discussion

The results described above hint explicitly at the hypothesis mentioned in the introduction that JOANA seems to activate higher-order gene sets. The GSE14924_CD4 dataset serves as an adequate example of this behavior. The term hsa05168 is big in size and is therefore annotated in many genes (*Figure 3*). This means it is a gene set of higher order since it comprises many genes and gene sets of lower order cover a subset of genes accordingly. The surrounding gene sets can be considered such subsets. Since they are activated by PADOG and not JOANA, a possible explanation is that JOANA activates the higher-order term hsa05168 and does not activate the smaller surrounding terms. This diminishes the shared use of significantly enriched genes, which are annotated in several gene sets. The significant genes are used by big gene sets first and are subsequently not available for smaller gene sets. Thus, the JOANA-active gene sets tend to be bigger in size and more significant by drawing significant genes to bigger terms.

The insight into the behavior of JOANA that this thesis provides helps to understand the results of JOANA in comparison to other methods such as PADOG. Having the knowledge that JOANA tends to mark higher-order gene sets gives those gene sets greater significance and reduces the number of gene sets that are relatively important through a few overlapping genes, approaching the problem of redundancy effectively. While JOANA may not be able to reach PADOG performance in terms of MalaCards benchmarking score, besides the single-species analysis it features the multi-omics addendum that allows analysis of more than one species simultaneously. On the contrary, PADOG is a single-species method only. Besides the research results, this thesis yields further questions. One immediate question is whether the JOANA result for dataset GSE14924_CD4, for example, would differ significantly if the gene set term hsa05168 was ignored from the assignment matrix. Furthermore, an assessment will be carried out as to why JOANA yields results with sometimes no terms active and sometimes even every single term or a large number of terms, which makes interpretability difficult and possible results meaningless. The large deviation of the outcome that different methods produce, i.e., the number of gene sets marked as active when analyzing a dataset, is also an important factor. Reliability is diminished if different methods yield results with almost no overlap on which gene sets are marked active.

One problem with JOANA is that some datasets yield unusable results, with the number of active terms either exceeding 70 terms, which is more than one fifth of all gene sets in the datasets examined, or being 0. Such results show that pre-selected input is necessary. If the input is not pre-selected, the validity of the output needs to be assessed. While these seemingly arbitrary results are not a problem specific to JOANA, Maleki et al. mention that this difference in outcome is a problem that occurs between different methods that were compared. Different methods yield different results with little consistency. Such variety of the outcome of methods suggests that most gene set analysis methods range from lacking specificity, and therefore introducing many false positive terms, to lacking sensitivity and missing true positives [1].

While PADOG uses pathway information, it seems to yield a consistent number of active terms in an interpretable range. This stability with various datasets is an advantage that PADOG has over JOANA. However, the ability to analyze multi-omics data with the cooperative model puts JOANA in a significantly different category of methods. If an analysis of single-omics data is needed PADOG is the recommended method of choice [8]. The multi-omics analysis of JOANA is tested for two species, but it is theoretically applicable to more species, just like MONA [6]. This functional flexibility makes JOANA a multipurpose method.

Ideally, in gene set analysis there would be no need for human expertise to assess the validity of the results since the human perspective can be biased and can distort interpretation. However, JOANA depends on human interpretation, as an output of all gene sets marked as active is meaningless and can stem from poor input data. If a significant number of genes are significantly enriched and many of these genes overlap, meaning they are annotated in several gene sets, PADOG will down-weight them to focus on the genes present in few gene sets and find fewer active terms. In JOANA higher-order active gene sets are expected. If JOANA tries to activate higher-order terms, i.e., gene sets, it will activate higher-order gene sets on any hierarchy level. More precisely, if gene set C is a subset of gene set B and gene set B is a subset of gene set A, the overlap of C and B can lead to an activation of B, just as the overlap of B and A can lead to A being marked active. A subset in this context means that, e.g., B and A share a certain number of genes, but gene set A has a bigger gene set size and is considered a higher-order term.

If the overlap of B and A led to gene set A marked as active, the overlap of C and B could still mark B as active. In datasets with many significantly enriched genes, this could explain why JOANA sometimes turns on more than half of all gene sets. In these examples, only the smaller gene sets are inactive. An approach from the other direction might result in very few gene sets marked as active. If C led to B being active and B then led to A being active, only the gene sets at the highest order would be marked active. This seems like a preferable outcome at first, but if the overlap of C and A is negligible, the significance of C will be neglected falsely. The overlap of gene sets C and B and the resulting state of activation or deactivation of B should not impact the influence of the overlap of B and A on its activation.

Disadvantages like the instability of how many gene sets will be turned on make a human interpretation of the results mandatory, though this problem is present in all GO enrichment methods. Even though JOANA engages the problem of redundancy, it does not include gene relevance scores like PADOG, but relies solely on the Bayesian Network and a proper fit of the mixture of Beta distributions, so that the first active Beta distribution covers all differentially expressed genes. Without a proper fit, JOANA may yield meaningless results.

JOANA is a method that provides multi-omics features. The possibility to analyze multiple species simultaneously is a powerful tool. JOANA also functions with different gene set databases such as GO and KEGG and is applicable in many ways because of this variability.

A combination of the strength of JOANA, turning on higher-order gene sets, and reducing redundancy could be combined with the strength of PADOG, which includes pathway topology to down-weight overlapping genes and produces a rather stable method output for the KEGG datasets under investigation.

An analysis of the performance of JOANA with the active outcome gene sets of PADOG and their respective genes as input may yield few but highly significant gene sets big in gene set size. Insignificant gene sets that are not active and filtered with PADOG would not impact the fitting of the Beta distributions. Multiple significant and overlapping gene sets could be analyzed by JOANA, which includes the hierarchy properties, resulting in a few very significant high-order active gene sets. In theory, this would turn off all the small active terms on the left of the joint graph of dataset GSE24739_G0 (*Figure 4*), which are small in gene set size and not relevant according to the MalaCards relevant score, directing the focus on the more densely connected area on the right.

An introduction of the down-weighting of overlapping genes that PADOG uses to the hidden response layer in JOANA might be another approach to join the advantages of both methods, although JOANA might not be suitable for that type of inner modification. Edge weights on the connections between the ontology term layer and the gene response hidden layer, depending on how relevant the gene in combination with the given gene set is to the condition, i.e., phenotype under investigation, may include the relevance information of MalaCards to the JOANA Bayesian Network.

Finally, further research of the effects of directionality of gene enrichment may yield suggestions on improvements of JOANA or other gene set analysis methods. How can positively and negatively enriched genes erase the relevance of one another or maybe even increase the relevance beyond reasonable scale? If the effects of directionality are significant, how can negative influence, i.e., performance repression, be mitigated and positive influence be enhanced to improve JOANA?

Supplementary Material

All supplementary material can be found online at https://github.com/TimDiedrich/HFA-JOANA.

References

[1] Maleki, F., Ovens, K., Hogan, D.J. and Kusalik, A.J., 2020. Gene set analysis: challenges, opportunities, and future research. *Frontiers in genetics*, *11*, p.654.

[2] Tilford, C.A. and Siemers, N.O., 2009. Gene set enrichment analysis. In *Protein networks and pathway analysis* (pp. 99-121). Humana Press.

[3] Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. and Harris, M.A., 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, *25*(1), pp.25-29.

[4] Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. and Tanabe, M., 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, *40*(D1), pp.D109-D114.

[5] Tarca, A.L., Draghici, S., Bhatti, G. and Romero, R., 2012. Down-weighting overlapping genes improves gene set analysis. *BMC bioinformatics*, *13*(1), pp.1-14.

[6] Sass, S., Buettner, F., Mueller, N.S. and Theis, F.J., 2013. A modular framework for gene set analysis integrating multilevel omics data. *Nucleic acids research*, *41*(21), pp.9622-9633.

[7] Kopf, A., Theis, F.J., Buettner, F., Joint continuous ontology enrichment analysis resolves hidden functional heterogeneity from multiple levels or tissues. *Work in progress.*

[8] Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., Law, C., Davis, S., Carey, V., Morgan, M. and Zimmer, R., 2021. Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in bioinformatics*, *22*(1), pp.545-556.

[9] Rappaport, N., Twik, M., Nativ, N., Stelzer, G., Bahir, I., Stein, T.I., Safran, M. and Lancet, D., 2014. MalaCards: A comprehensive automatically-mined database of human diseases. *Current Protocols in Bioinformatics*, *47*(1), pp.1-24.

[10] Smyth, G.K., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, *3*(1).

[11] Kim, H.Y., 2017. Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test. *Restorative dentistry & endodontics*, *42*(2), pp.152-155.

[12] Bower, K.M., 2003, August. When to use Fisher’s exact test. In *American Society for Quality, Six Sigma Forum Magazine* (Vol. 2, No. 4, pp. 35-37).

[13] Bishop, C.M., 2006. Pattern recognition. *Machine learning*, *128*(9).

[14] Ji, Y., Wu, C., Liu, P., Wang, J. and Coombes, K.R., 2005. Applications of beta-mixture models in bioinformatics. *Bioinformatics*, *21*(9), pp.2118-2122.

[15] Schröder, C. and Rahmann, S., 2017. A hybrid parameter estimation algorithm for beta mixtures and applications to methylation state classification. *Algorithms for Molecular Biology*, *12*(1), pp.1-12.

[16] Minka, T.P., 2001. Expectation Propagation for approximate Bayesian inference. *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers Inc, UAI'01, San Francisco, CA, pp. 362-369.

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst habe. Ebenso bestätige ich, dass diese Arbeit nicht, auch nicht auszugsweise, für eine andere Prüfung oder Studienleistung verwendet wurde.

Ort, Datum: Frankfurt am Main, 16.09.2021 Unterschrift: Tim Diedrich

Bitte dieses Formular zusammen mit der Abschlussarbeit abgeben!

Erklärung zur Abschlussarbeit

gemäß § 25, Abs. 11 der Ordnung für den Bachelorstudiengang Informatik vom 06. Dezember 2010:

Hiermit erkläre ich Herr / ~~Frau~~

Tim Diedrich

Die vorliegende Arbeit habe ich selbstständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst.

Frankfurt am Main, den 16.09.2021

T. Diedrich

Unterschrift der Studentin / des Studenten