American Journal of Advanced Research, 2018, 2–1
December. 2018, pages 24-30
doi: 10.5281/zenodo.1438968
http://www.ibii-us.org/Journals/AJAR/
ISBN 2472-9264 (Online), 2472-9256 (Print)

Genomic Big Data Regression Analysis

Alin Tomoiaga*, John Farelly and Shintaro Nakamura

Department of Business Analytics, Manhattan College, 4513 Manhattan College Pkwy, Bronx, NY 10471

*Email: alin.tomoiaga@manhattan.edu

Received on August 01, 2018; revised on September 25, 2018; published on September 29, 2018

Abstract

DNA analysis is now a data intensive discipline. New technology has transformed biomedical research by making a plethora of molecular data available at reduced costs and great speeds. Large consortiums and many individual laboratories have already generated vast datasets: as an example, one such database, the Gene Expression Omnibus (GEO contains more than 1.8 million samples. This data is readily, publicly available but analyzing it requires computational and statistical resources.

A popular concern in biological research is to identify those genomic pathways that are related to the organism's reaction to treatment or disease. There are numerous techniques that try to reduce the false positive errors and rank the pathways according to the degree of the phenotype relationship strength. This goal is accompanied by several challenges: finding parsimonious models with a good balance between simplicity and complexity and designing methods for pathway selection using appropriate significance thresholds. Often, it is difficult to escape the temptation of "ad-hoc" procedures that may work for particular examples but cannot be properly expanded to general cases.

Over the years, many methods have been proposed but over-representation analysis (ORA) remains the most popular. The underlying assumption of ORA is that pathways with an irregular number of differentially expressed genes are responsible for the phenotype to the detriment of lesser differentially expressed pathways. Under the umbrella of logistic regression, we propose a method that aims to improve ORA. We show that traditional hypergeometric ORA methods are fully described by and can be considered a special case of the logistic regression methods. Logistic regression presents the advantage that while it produces simple models, they are richer, and they describe the biological process in a more accurate fashion. While logistic regression has been proposed before as a solution for ORA, we prove the over-encompassing nature of the method and we also propose a flavor of regression that can be aimed at different scenarios. Furthermore, logistic regression has a solid mathematical basis and produces results that have biological justification.

Keywords: big data, regression, genomic

1 Introduction

Genomic pathway analysis places high importance on finding the pathways that are related to the phenotype that is being studied.

For the purposes of this paper, we define a pathway as a collection or subset of genes with a common biological process or molecular function. The subset definition is implicit in over-representation analysis (ORA) and we do not take into consideration any other biological characteristics of pathways. Pathway information for various organisms has been collected and summarized and has been made publicly accessible on public knowledge bases like KEGG [9], MetaCyc [10], Reactome [8], and others.

Several techniques have been developed for pathway analysis: over-representation analysis (ORA), functional class scoring (FCS) and topology-based are some of the most popular types of methods used. These methods are well known and their strengths and weaknesses have been discussed and compared [12]

Although the oldest method, ORA is still very popular as it produces simple models that are easy to interpret. GSEA (Gene Set Enrichment Analysis) [19] is a popular functional class scoring (FCS) method that produces an aggregated pathway level enrichment score and by permuting gene labels makes the need for a significance threshold obsolete. Topology based methods, like the bioconductor SPIA package [20] also take into consideration information about how genes interact with each other (inhibition, activation, etc.).

For the remainder of this paper, we will focus our attention on ORA. The basic idea of ORA methods is summa- rized by name "over-representation analysis": the underlying assumption is that a pathway that contains a larger than expected number of differentially expressed genes is more relevant to the biological process [4, 11]. Therefore, the crux of the problem becomes finding a method that correctly identifies pathways that have an elevated level of gene expression. The many tools that fall under the umbrella of ORA usually produce a Fisher exact test.

Most of the time, ORA is able to correctly identify relevant pathways, rank them according to importance and it has been proven to be effective in various practical situations, but ORA also presents a few challenges that need to be overcome: for certain cases, it produces false positive results when non-relevant pathways are singled-out as important or false negative results, by ignoring important pathways for the phenotype. By ignoring the fact that some pathways share many genes, ORA clouds the pathway significance distinction: for example, some pathways will falsely appear significant just because they overlap other truly significant pathways [6, 14]. Procedures have been proposed to augment ORA to address the issues that arise because of the overlap between pathways [3, 21]. Regression has already been proven to be successful in modelling similar processes to the ones described in [22, 16, 18];

Overlapping pathways do represent an important problem in need of a simple solution, as it is not clear what effect the common genes have on each pathway. Many times, these common genes only account for basic metabolic processes (e.g., energy processing) and their function is not related to the studied phenotype.

We start by showing, analytically, that ORA methods can be considered a subset of logistic regression analysis. Simple ORA many times gets reduces to a Fisher exact test and the supplemental material for [18] does discuss the similarity between logistic regression and Fisher's test, but only focuses on simulations to compare the two methods. We remind the reader that exact logistic regression can be reduced to a Fisher exact test and we show how the two procedures test the same hypotheses at the model level and they are also identical at the data level. We will also show that any combination of pathways (module) can be expressed by a regression model, so even an enhanced ORA analysis will still be a special case of regression.

Then, we will propose a variation of the broader concept of logistic regression and we will show that this model improves on some over-representation analysis traps and it may also may point to pathway modules that have not been given proper attention in prior studies. Therefore, regression produces simple models, not as simple as ORA but more rich and more accurate as it retains more of the characteristics of the underlying process.

2 Methods

We illustrate how we can slightly modify logistic regression to improve on ORA results. The results on the real data presented in this section show that the proposed method provides reasonable results compared to ORA. The collection of pathways was downloaded from the KEGG database, Release 55. Newer releases of course will affect the results.

The first data set on which we tested the method comes from an experiment investigating cellular and metabolic plasticity of white fat tissue (WAT), which stores lipid energy. This experiment, studies the transformation of WAT, under certain physiological and pharmacological conditions, into one resembling brown fat, a thermogenic organ[6, 15].

The data set used for the classical over-representation analysis (ORA) was obtained from a microarray analysis of white fat from mice treated with low dose (0.75 nmol/hr) CL 316,243 (CL) for 0 and 7 days. More on the biological aspects of this phenomenon can be found in the literature [15, 17, 6] The genes were ordered by p-value and the top 5% were selected as differentially expressed (DE).

According to the classical ORA analysis, (Table 3a) of the first comparison, the pathways with a p-values smaller than 0.01 after FDR correction are Parkinson's, Alzheimer's, Huntington's, Leishmaniasis, Phagosome, Cell Cycle, Oocyte Meiosis, Cardiac Muscle Contraction, Toll-like receptor, and PPAR Signaling. Pathways with p-value smaller than 0.05 after FDR correction are Chemokine Signaling Pathway, Lysosome, B Cell Receptor, Systemic Lupus Ery- thematosus, Complement and Coagulation Cascades, Cytokine Cytokine Receptor Interaction, and Chagas Disease. The list of pathways with their associated corrected p-value is summarized in Table 3a.

The immediate problem with the classification in Table 3a is that the theoretical results do not match the bio- logical observations: pathways like Parkinson's disease, Alzheimer's disease, and Huntington's disease are related to degenerative diseases of the central nervous system and obvious or apparent connection fat remodeling; Leishmaniasis describes the protozoan parasitic disease, a disease spread by sand flies. These four pathways do not have any known relationship to the phenomenon under study, yet they are listed at the top of the ORA ranking. Other pathways like Cell cycle and p53 Signaling pathway are more likely to be related to fat remodeling but they are listed lower in the ranking.

2.1 Simple over-representation analysis

Simple over-representation analysis considers the level of expression in one pathway against everything else outside the selected pathway. Without loss of generality, suppose the interest pathway is pathway 1, called $Pathway_1$. Let us consider the table 1, that describes the information about gene expression and pathway $Pathway_1$ gene composition.

Y_i are independent Bernoulli random variables and let us consider $\Pr(Y_i = 1|X_{i1} = 1) = \pi_1$ and $\Pr(Y_i = 1|X_{i1} = 0) = \pi_0$, for all $i = 1, \dots, g$. The information in 1 can be summarized in a contingency table like 1 and usually a hypergeometric statistic is calculated from it.

Gene Label	Observed Expression (binary)	Pathway 1 Indicator (binary)
1	Y_1	X_{11}
2	Y_2	X_{21}
3	Y_3	X_{31}
⋮	⋮	⋮
g	Y_g	X_{g1}

Table 1: Single interest pathway, P$_1$. Gene expression model

	DE	DE^c	Total
P_1	d_1	$g_1 - d_1$	g_1
P_1^c	$d - d_1$	$(g - d) - (g_1 - d_1)$	$g - g_1$
Total	d	$g - d$	g

Table 2: ORA contingency table. Standard over-representation approach contingency table; g_1 and g represent, respectively, the number of genes belonging to pathway P_1 and the total number of genes. d_1 and d represent, respectively, the number of differentially expressed genes belonging to pathway P_1 and the total number of DE genes.

Considering the defined probabilities above, we observe that $d_1 \sim Bin(g_p, \pi_1)$ and $d - d_1 \sim Bin(g - g_1, \pi_0)$.

At this point, the researcher usually needs to make a decision about the degree of extremeness of this observation. The question is: how probable is it, that by random chance alone, more than d_i differentially expressed genes would be observed on a pathway? The concept of calculating a p-value naturally develops.

Usually, for a 2x2 contingency table like 1, the researcher has two options for a statistical test: a chi-square test or a Fisher exact test. The chi-square test does provide good asymptotic approximations and it is computationally fast. The Fisher test on the other hand provides an exact p-value and since computers are getting faster and faster, it is increasingly popular.

By [12], the odds ratio, ρ and under the null hypothesis $H_0\!:\pi_1 = \pi_0$, we have:

$$\Pr(d_1 = x|d = d_t) = \frac{\binom{g_1}{x}\binom{g-g_1}{d-d_1}}{\binom{g}{d_t}}, \text{ where } \rho = \frac{\frac{\pi_1}{(1-\pi_1)}}{\frac{\pi_0}{(1-\pi_0)}}$$

Considering the null hypothesis to be true, for our experiment, the right sided p-value is defined as the probability of obtaining a higher proportion of differentially expressed genes than the observed proportion. We reject H_0 at significance level α when $d_1 \geq C^{(\alpha)}(d)$ where $C^{(\alpha)}(d_t)$ is the smallest integer for which $P_0\{N_{T1} \geq C^{(\alpha)}(d)\big|d = d_t\} \leq \alpha$ if such $C^{(\alpha)}(d_t)$ exists. If it does not exist then let $C^{(\alpha)}(d_t) = \infty$.

Calculating a two-sided p-value is not as straightforward as in the case of symmetric distributions and there are at least two ways of calculating it. The simplest is to consider $2\min(p, 1-p)$ [22]; this method has the disadvantage of being too conservative and in some cases producing a p-value greater than 1. The other option is to calculate p-values for all the 2x2 contingency tables with p-values less than the observed p-value [6]. A discussion of the various methods has been realized in [4].

Logistic regression does describe this simple case. Logistic regression is used for describing models involving categorical response variables. When the response variable is categorical, logistic regression is preferred over linear regression for multiple reasons and the most important two reasons are:

• Linear regression produces predictions that fall outside the categorical response variable's range.

• The linear regression error terms are not normally distributed. This is an assumption of linear regression and the categorical nature of the response variable makes this assumption untenable from the start.

Linear regression makes use of the least squares method to find parameter estimates. The least squares method is not applicable for logistic regression which uses maximum likelihood estimation in order to estimate its parameters.

2.2 The logistic regression model

Gene Label	Observed Expression (binary)	Pathway 1 Indicator (binary)	Pathway 2 Indicator (binary)	...	Pathway k Indicator (binary)
1	Y_1	X_{11}	X_{12}	...	X_{1k}
2	Y_2	X_{21}	X_{22}	...	X_{2k}
3	Y_3	X_{31}	X_{32}	...	X_{3k}
⋮	⋮	⋮	⋮	⋱	⋮
g	Y_g	X_{g1}	X_{g2}	...	X_{gk}

Table 3: Multiple interest pathways. Gene expression model

Consider the dataset in Table 3.

Let $\mathbf{X}_i$ represent a binary representation of $Pathway_i$. $\mathbf{X}_i$, where $i \in \{1, \dots, k\}$ are vectors of length g where $X_{ij} = 1$ if gene j is present on the pathway i and $X_{ij} = 0$, otherwise, for all $j \in \{1, \dots, g\}$.

We consider $\mathbf{Y}$ a g length column vector of random binomial variables Y_i. For our purposes, Y_i can take a value of either 1 or 0 and it signifies whether gene i is differentially expressed or not (1 or 0, respectively). Let y represent a g length observed values vector corresponding to the random variable vector $\mathbf{Y}$. Let $\mathbf{\Pi}$ represent a g length probability vector. Let θ be the parameter vector of length $k+1$. θ contains a parameter for each vector $\mathbf{X}_i$ and θ_0 for the intercept term.

Then the logit model becomes:

$$\text{Logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{k} \theta_j X_{ij}, \quad for\ \ i \in \{1, \dots, g\}$$

2.3 Logistic regression, parameter estimation using MLE

Since $\mathbf{Y}$ is a vector of binomial variables, the likelihood function for $\mathbf{Y}$ becomes:

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{g} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

So,

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{g} \left(\frac{\pi_i}{1-\pi_i}\right)^{y_i}(1-\pi_i)$$

Since

$$\text{Log}\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{k} \theta_j X_{ij}$$

we can rewrite the likelihood function as

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{g} \left(e^{\sum_{j=0}^{k}\theta_j X_{ij}}\right)^{y_i}\left(1 - \frac{e^{\sum_{j=0}^{k}\theta_j X_{ij}}}{1 + e^{\sum_{j=0}^{k}\theta_j X_{ij}}}\right)$$

and

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{g} \left(e^{\sum_{j=0}^{k}\theta_j X_{ij}}\right)^{y_i}\left(1 + e^{\sum_{j=0}^{k}\theta_j X_{ij}}\right)^{-1}$$

We need to find the maximum of the likelihood function. This involves finding the first derivative and setting it equal to 0. Once a solution is found, we need to find the second derivative. If the second derivative evaluated at the solution is positive then we found a maximum.

Since this function is still difficult to differentiate, the log likelihood function is used. Since the log function is monotonic, applying log to the likelihood function will not affect the solution of the initial function.

$$l(\theta|\mathbf{y}) = \sum_{i=1}^{g} y_i * \sum_{j=0}^{k} \theta_j X_{ij} - \log\left(1 + e^{\sum_{j=0}^{k}\theta_j X_{ij}}\right)$$

Taking derivatives with respect to each one of the parameters produces:

$$\frac{\delta}{\delta\theta_j} l(\theta|\mathbf{y}) = \sum_{i=1}^{g} y_i X_{ij} - \pi_i X_{ij}$$

If we set each derivative to 0 we are reduced to solving a system of k+1 equations and k+1 unknowns.

Solving such a system usually requires numerical methods.

2.4 Exact logistic regression

Exact logistic regression adopts a quite different approach on inferring the parameter vector θ.

While the maximum likelihood estimate uses an asymptotic approach, the "exact" method provides a different perspective approach on this problem.

Unlike the asymptotic approach, exact inference has the advantage that it provides an exact solution for all sample sizes and it does not make any assumptions about the sample size. The disadvantage of the exact solution is that it is computationally intensive. Although in the past, the exact solution was mostly applied for small sample sizes, the computational speeds of today's computers make it more appealing and its popularity is increasing.

The idea behind this approach is:

- find a sufficient statistic for the θ parameter vector; showing the statistic is sufficient is done using the Fisher-Neyman factorization theorem.
- then, using the sufficiency principle, we can condition on some of the sufficient statistic components; this way we can control the select parameters that the conditional probability depends on.
- considering a constrained event space, we count the sample points and compute a p-value.

The likelihood conditional function is:

$$L(\theta|\mathbf{y}) = \prod_{i=1}^{g} (e^{\sum_{j=0}^{k} \theta_j X_{ij}})^{y_i} (1 + e^{\sum_{j=0}^{k} \theta_j X_{ij}})^{-1}$$

A sufficient statistic for θ is :

$$t = y'X$$

So the conditional distribution of θ is:

$$\Pr(T_1 = t_1, T_2 = t_2, \dots, T_k = t_k) = \frac{c(\mathbf{t}) e^{\theta' \mathbf{t}}}{\sum_{\mathbf{u}} c(\mathbf{u}) e^{\theta' \mathbf{u}}}$$

where $c(t)$ is the number of distinct sets of y that give the same value for our sufficient statistic t and the denominator is summed over all u such that $c(u) \geq 1$.

If we want to make inferences about one parameter, let us pick θ_1, without loss of generality, then we can consider

$$\Pr(T_1 = t_1 | T_2 = t_2, \dots) = \Pr(T_1|\theta_1) = \frac{c(t_1, t_2, \dots, t_k) e^{\theta_1 t_1}}{\sum_u c(t_2, t_3, \dots, t_k, u) e^{\theta_p u}}$$

We can use this probability to make inferences about θ_1.

For, example if the hypothesis to test is $H_0\colon \theta_1 = 0$ then we can calculate an exact p-value by summing over the critical region C.

$$\sum_{c(t) \in C} \Pr(c(t) | \theta_1 = 0)$$

.

2.6 Fisher's exact test vs. Exact logistic regression

We will show, for a simple case, that Fisher's exact test and exact logistic regression are the same at the model level and they produce the same results at the data level. We consider **Y** to be the response vector variable and $\mathbf{X}_1$ a predictor vector variable, corresponding to pathway $Pathway_1$.

Model comparison

Let us assume, like before that $\Pr(Y_i = 1|Pathway_1) = \pi_1$ and $\Pr(Y_i = 1|Pathway_1^c) = \pi_0$, for all $i = 1, \dots, g$.

We have already discussed how the null hypothesis for the Fisher exact test is:

$$H_0\colon \pi_1 = \pi_0$$

Let us show that the logistic regression hypothesis is identical to the Fisher null hypothesis.

So, the logit model for one pathway is:

$$\text{Logit}(\pi_1|X_{i1} = 1) = \log\left(\frac{\pi_1}{1 - \pi_1}\right) = \theta_0 + \theta_1 X_{i1}, \qquad i \in \{1, \dots, g\}$$

and

$$\text{logit}(\pi_0|X_{i1} = 0) = \log\left(\frac{\pi_0}{1 - \pi_0}\right) = \theta_0 + \theta_1 X_{i1},\ \ i \in \{1, \dots, g\}.$$

Therefore,

$$logit(\pi_1) = \theta_0 + \theta_1$$

and

$$logit(\pi_0) = \theta_0.$$

The logistic regression null hypothesis is $\theta_1 = 0$.

Therefore, if $\theta_1 = 0$ then $\text{logit}(\pi_1) = \text{logit}(\pi_0) = \theta_0$ or $\pi_1 = \pi_0$ is an equivalent form of the hypothesis.

Probability calculations

We have already shown that the probability distribution calculated by the Fisher exact test is:

$$\Pr(d_1 = x | d = d_t) = \frac{\binom{g_1}{x}\binom{g - g_1}{d - d_1}}{\binom{g}{d_t}}$$

On the other hand, the exact logistic regression model calculates the conditional distribution of θ:

$$\Pr(T_1 = t_1 | T_0 = t_0, \dots) = \Pr(T_1|\theta_1) = \frac{c(t_1, t_2, \dots, t_k) e^{\theta_1 t_1}}{\sum_u c(t_2, t_3, \dots, t_k, u) e^{\theta_p u}}$$

where $c(t)$ is the number of distinct sets of y that give the same value for our sufficient statistic t and the denominator is summed over all u such that $c(u) \geq 1$.

For the simple case of one expressed pathway P_1, we will show that the two probability distributions are the same.

First of all, we will establish some correspondences between the quantities defined by the two methods.

t_0 and t_1 are defined as sufficient statistics for θ_0 and θ_1, respectively.

$$t_0 = \sum_{i=1}^{g} y_i.$$

$$t_1 = \sum_{i=1}^{g} y_i * X_{i1}$$

So finding the probability that $T_1 = t_1$ corresponds to $d_1 = x$ in the Fisher procedure.

So conditioning on $T_0 = t_0$ is the same as conditioning on $d = d_t$ in the Fisher procedure. Therefore calculating $P(d_1 = x|d = d_t)$ is the same as calculating $Pr(T_1 = t_1|T_0 = t_0)$.

Now $c(t)$ is the number of distinct sets of y that give the same value for our sufficient statistic t. That is exactly what the product $\binom{g_1}{x}\binom{g-g_1}{d-d_1}$ calculates.

2.7 Simulation and comparison

Applying both methods to the same dataset produces similar results. [1] on page 253 makes a comparison between the two methods.

As an example, let us consider a simple table:

Gene Label	Observed Expression (binary)	Pathway 1 Indicator (binary)
1	1	1
2	1	1
3	1	1
4	0	1
5	0	0

Exact logistic and fisher's exact test produce identical p-values (the right sided p-value is 0.4).

2.8 Logistic regression fully encompasses ORA

So far, we have shown that logistic regression can be configured to express an identical model to ORA and we have proven this both analytically and through a simulation. Furthermore, in the Supplemental Material, we show that a multiple logistic regression with interactions can be configured to model any complex configuration of overlapping pathways, thus it can be used even for situations that exceed ORA's modelling capabilities.

3 Results. Application of our modified logistic model to a real experimental dataset

We illustrate how we can slightly modify logistic regression to improve on ORA results. The results on the real data presented in this section show that the proposed method provides reasonable results compared to ORA.

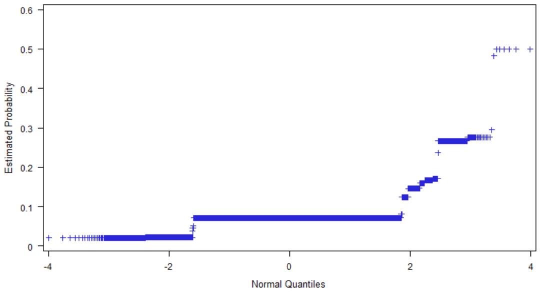


Figure 1 A plot of the predicted gene probability of expression against normal quantiles. All genes have been ranked in decreasing order by their predicted probability of expression. These probability have been plotted against the normal quantiles.

The collection of pathways was downloaded from the KEGG database, Release 55. Newer releases of course will affect the results.

3.1 Removing the false positives using logistic regression predictions

We tested the method on a dataset that comes from an experiment investigating cellular and metabolic plasticity of white fat tissue (WAT), which stores lipid energy. This experiment, studies the transformation of WAT, under certain physiological and pharmacological conditions, into one resembling brown fat[5, 14].

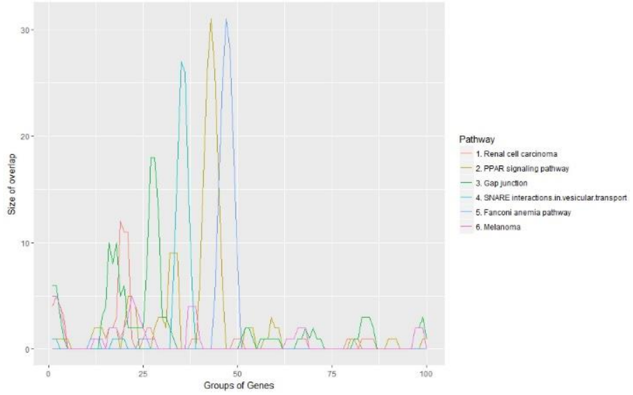


Figure 2 Pathway overlaps.. A logistic regression model is run and pathways are ordered by increasing p-value and genes are ranked decreasingly on their predicted probability of expression. For this figure, we only retain the top 6 pathways that have significant p-values. Then, genes are grouped 100 at a time and we keep track of the number of genes that belong to each of the top six pathways. We observe how pathways like Renal cell carcinoma, Melanoma, Gap junction share many top ranked genes. Because of this overlap, we will discount the overlapping pathways, and we will retain pathways PPAR signaling or Fanconi anemia pathway that contain unique modules of genes.

The data set used for the classical over-representation analysis (ORA) was obtained from a microarray analysis of white fat from mice treated with low dose (0.75 nmol/hr) CL 316,243 (CL) for 0 and 7 days. More on the biology of this phenomenon can be found in the literature [14, 16, 5] The genes were ordered by p-value and the top 5% were selected as differentially expressed (DE).

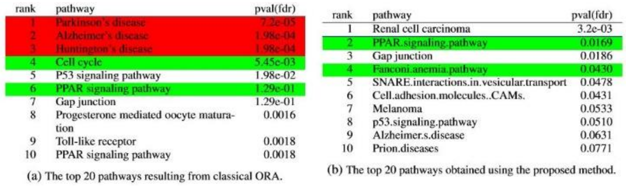
According to the classical ORA analysis, (Figure 3a) of the first comparison, the pathways with a p-values smaller than 0.01 after FDR correction are *Parkinson's, Alzheimer's, Huntington's, Leishmaniasis, Phagosome, Cell Cycle, Oocyte Meiosis, Cardiac Muscle Contraction, Toll-like receptor*, and *PPAR Signaling*. Pathways with p-value smaller than 0.05 after FDR correction are *Chemokine Signaling Pathway, Lysosome, B Cell Receptor, Systemic Lupus Erythematosus, Complement and Coagulation Cascades, Cytokine Cytokine Receptor Interaction*, and *Chagas Disease*. The list of pathways with their associated corrected p-value is summarized in Figure 3.

The immediate problem with the classification in Figure 3a is that the theoretical results do not match the biological observations: pathways like Parkinson's disease, Alzheimer's disease, and Huntington's disease are related to degenerative diseases of the central nervous system and obvious or apparent connection fat remodeling; Leishmaniasis describes the protozoan parasitic disease, a disease spread by sand flies. These four pathways do not have any known relationship to the phenomenon under study, yet they are listed at the top of the ORA ranking. Other pathways like Cell cycle and p53 Signaling pathway are more likely to be related to fat remodeling but they are listed lower in the ranking.

3.2 Logistic regression improves the overlapping pathways problem

Our method aims to apply logistic regression and improve the results of ORA. Our assumption is that genes that belong to overlapping regions conceal the true signal and the analysis should focus on highly expressed, but unique genes that belong to only one pathway. We proceed to identify these unique genes.

At first, we ran a stepwise logistic regression analysis and pathways have been ranked according to their p-values (Figure 3b). The regression analysis also produces probability predictions for each gene and the genes have been ranked according to the expression probabilities that have been predicted by the model. In Figure 1, the predicted gene expression ranking shows how a group of seven genes at the top have a predicted probability of expression noticeably higher than the rest of the genes. This observation again points to the idea that there are distinct modules of genes that share their predicted probability of expression.



rank	pathway	pval(fdr)
1	Parkinson's disease	7.2e-05
2	Alzheimer's disease	1.98e-04
3	Huntington's disease	1.98e-04
4	Cell cycle	5.45e-03
5	P53 signaling pathway	1.98e-02
6	PPAR signaling pathway	1.29e-01
7	Gap junction	1.29e-01
8	Progesterone mediated oocyte maturation	0.0016
9	Toll-like receptor	0.0018
10	PPAR signaling pathway	0.0018

(a) The top 20 pathways resulting from classical ORA.

rank	pathway	pval(fdr)
1	Renal cell carcinoma	3.2e-03
2	PPAR.signaling.pathway	0.0169
3	Gap junction	0.0186
4	Fanconi.anemia.pathway	0.0430
5	SNARE.interactions.in.vesicular.transport	0.0478
6	Cell.adhesion.molecules..CAMs.	0.0431
7	Melanoma	0.0533
8	p53.signaling.pathway	0.0510
9	Alzheimer.s.disease	0.0631
10	Prion.diseases	0.0771

(b) The top 20 pathways obtained using the proposed method.

Figure 3 Analysis of the fat remodeling experiment for the comparison between days 7 and 0, with comparison between ORA (left) with our method (right). Only the top 10 pathways are presented. Pathways in red represent are not related with fat remodeling, while pathways highlighted in green are those for which we know they influence the phenomenon. The white background indicates inconclusive pathways. Classical ORA yields clear false positives, while our method is able to discern true positives

We then proceed to analyze the top pathways in our ranking in Figure 3b) and we eliminate the pathways that share large modules of highly expressed genes. In Figure 2, pathways like Renal cell carcinoma, Melanoma share many genes with other pathways, like PPAR signaling pathway and Fanconi anemia.

By discounting the overlapped genes, our method can filter out pathways like Renal cell carcinoma, Melanoma as false positives and it points to PPAR signaling pathway and Fanconi anemia as involved in the fat remodeling process. These pathways have been proven to be important [6] for fat metabolism.

4 Conclusion

We have shown that logistic regression totally encompasses every case that ORA describes. Furthermore, logistic regression is able to model any complex combination of overlapping pathways. With some simple modifications, we also showed that our method is able to discount false positives from the top of the rankings. We recommend our method for its simplicity and for situations where there are large overlaps between pathways.

Conflict of Interest: none declared.

References

[1] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd edition, 2002.

[2] M. D. Z. X. A. T. P. W. S. Draghici. A method for analysis and correction of cross-talk effects in pathway analysis. *Nucleic Acids Research*, 33(suppl 1):D428–D432, 2005. reactome knowledge base.

[3] S. Drăghici, P. Khatri, R. P. Martins, G. C. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, February 2003.

[4] M. P. Fay. Two-sided Exact Tests and Matching Confidence Intervals for Discrete Data. *The R Journal*, 2(1):53–58, June 2010.

[5] J. G. Granneman, P. Li, Z. Zhu, and Y. Lu. Metabolic and cellular plasticity in white adipose tissue I: effects of beta3-adrenergic receptor activation. *American Journal Of Physiology-Endocrinology And Metabolism*, 289(4):E608–616, 2005.

[6] K. Hirji. *Exact Analysis of Discrete Data*. Taylor & Francis, 2005.

[7] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D'Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. Gopinath, G. 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. reactome knowledge base.

[8] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000. this is the knowledge base for KEGG.

[9] P. D. Karp, M. Riley, S. M. Paley, and A. Pellegrini-Toole. The metacyc database. *Nucleic Acids Research*, 30(1):59–61, 2002. This is another knowledge base database.

[10] P. Khatri, S. Draghici, G. C. Ostermeier, and S. A. Krawetz. Profiling gene expression using Onto-Express. *Genomics*, 79(2):266–270, February 2002.

[11] P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375, 02 2012. offers an overview of the current methods of analysis: ORA, FCS,topology-based.

[12] E. Lehmann. *Testing Statistical Hypotheses: And Theory of Point Estimation*. Probability and Statistics Series. John Wiley & Sons Canada, Limited, 1986.

[13] M. Li, D. F. Carpio, Y. Zheng, P. Bruzzo, V. Singh, F. Ouaaz, R. M. Medzhitov, and A. A. Beg. An essential role of the nf-îºb/toll-like receptor pathway in induction of inflammatory and tissue-repair gene expression by necrotic cells. *The Journal of Immunology*, 166(12):7128–7135, 2001.

[14] P. Li, Z. Zhu, Y. Lu, and J. G. Granneman. Metabolic and cellular plasticity in white adipose tissue II: role of peroxisome proliferator-activated receptor-alpha. *American Journal Of Physiology-Endocrinology And Metabolism*, 289(4):E617–626, 2005.

[15] Y. Luan and H. Li. Group additive regression models for genomic data analysis. *Biostatistics*, 9(1):100–113, 2008. Response variable is a survival variable or censoring.Regression considering gene interaction.

[16] E. P. Mottillo, X. J. Shen, and J. G. Granneman. Role of hormone-sensitive lipase in beta-adrenergic remodeling of white adipose tissue. *Am J Physiol Endocrinol Metab*, 293(5):E1188–97, 2007.

[17] M. A. Sartor, G. D. Leikauf, and M. Medvedovic. Lrpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211–217, 2009.

[18] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[19] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-s. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero. A novel signaling pathway impact analysis.

Bioinformatics, 25(1):75–82, 2009. Mostly looked at http://bioinformatics.oxfordjournals.org/content/suppl/2008/11/05/btn577.DC1/bioinf-2008-1181-File002.pdf which shows a way to combine independent p-values.

[20] A. Tomoiaga, P. Westfall, M. Donato, S. Draghici, S. Hassan, R. Romero, and P. Tellaroli. Pathway crosstalk effects: shrinkage and disentanglement using a bayesian hierarchical model. *Statistics in Biosciences*, 8(2):374–394, 2016.

[21] Z. Wei and H. Li. Nonparametric pathway-based regression models for analysis of genomic data. *Biostatistics*, 8(2):265–284, 2007. This paper has a similar idea: regression on gene expression level, but it calculates an activity level for a pathway as a sum of expression levels of its genes; then it relates these levels to the phenotype through a regression model.

[22] F. Yates. Tests of significance for 2x2 contingency tables. *Journal of Royal Statistics Society*.