https://doi.org/10.1186/s12859-024-05871-x

BMC Bioinformatics

RESEARCH

Open Access

Ant colony optimization for the identification of dysregulated gene subnetworks from expression data

Eileen Marie Hanna[1*], Ghadi El Hasbani[1] and Danielle Azar[1]

*Correspondence: eileenmarie.hanna@lau.edu.lb

[1] Department of Computer Science and Mathematics, Lebanese American University, Byblos, Lebanon

Abstract

Background: High-throughput experimental technologies can provide deeper insights into pathway perturbations in biomedical studies. Accordingly, their usage is central to the identification of molecular targets and the subsequent development of suitable treatments for various diseases. Classical interpretations of generated data, such as differential gene expression and pathway analyses, disregard interconnections between studied genes when looking for gene-disease associations. Given that these interconnections are central to cellular processes, there has been a recent interest in incorporating them in such studies. The latter allows the detection of gene modules that underlie complex phenotypes in gene interaction networks. Existing methods either impose radius-based restrictions or freely grow modules at the expense of a statistical bias towards large modules. We propose a heuristic method, inspired by Ant Colony Optimization, to apply gene-level scoring and module identification with distance-based search constraints and penalties, rather than radius-based constraints.

Results: We test and compare our results to other approaches using three datasets of different neurodegenerative diseases, namely Alzheimer's, Parkinson's, and Huntington's, over three independent experiments. We report the outcomes of enrichment analyses and concordance of gene-level scores for each disease. Results indicate that the proposed approach generally shows superior stability in comparison to existing methods. It produces stable and meaningful enrichment results in all three datasets which have different case to control proportions and sample sizes.

Conclusion: The presented network-based gene expression analysis approach successfully identifies dysregulated gene modules associated with a certain disease. Using a heuristic based on Ant Colony Optimization, we perform a distance-based search with no radius constraints. Experimental results support the effectiveness and stability of our method in prioritizing modules of high relevance. Our tool is publicly available at github.com/GhadiElHasbani/ACOxGS.git.

Keywords: Gene expression, Enrichment analysis, Gene interaction network, Ant Colony Optimization

Background

Differential Gene Expression Analysis (DEA), typically performed using tools such as *limma*, is the most widely used method for detecting significant gene-disease associations based on mean expression variations between phenotypes [1]. However, while DEA can identify specific disease-associated genes, it does not take into consideration the network of interactions that govern the studied set of genes. The latter might lead to missing crucial mechanistic insights about multi-gene connections that underlie complex diseases. As a result, DEA can exhibit poor consistency between analyses of similar studies [2, 3]. To address this drawback, several methods that take into consideration the structure of gene networks were introduced. They measure the effect of nodes over their direct and indirect neighbors. Their output typically consists of modules which are groups of dysregulated genes that contribute to a disease or phenotype of study.

One group of methods makes use of interaction information to perform pathway analyses, yielding a *group* of genes underlying a phenotype of interest. These pathway analyses can be functional such as GSEA [4] which aims to identify functionally enriched groups of genes in relation to a phenotype; topological such as SPIA [5] and CePa [6, 7] which enhance functional scoring analyses with network information; or active module tools such as *jActiveModules* [8], HotNet [9], and COSINE [10], which combine expression and network information to identify disease-relevant subnetworks within pathways. Although pathway-level analyses can identify mechanistically interpretable multi-gene interactions [2, 11, 12], insights can be difficult to explore experimentally due to the lack of a precise gene target. Moreover, artificial pathway boundaries might limit the set of considered interactions. Therefore, there has been an increasing interest in developing gene-level analyses that incorporate network information.

Examples of such network-based methods include ENDEAVOUR [13] which relies on gene similarity with known disease genes, and GeneWanderer [14] which relies on gene distance from disease-relevant genes. Since information on such genes is sometimes not available, other methods were developed to overcome the need for gene-disease relevance information. For instance, the method in [15] performs a Laplacian kernel to transform the original network distances. It then uses this indirect distance measure along with differential expression of Laplace neighbors to identify disease genes. Similarly, DiSNEP [16] enhances the network with a diffusion process using similarity information. Since both methods effectively alter the distance between nodes in a network, interpreting the resulting modules and gene-level scores becomes complicated. Another method extends SPIA to produce a gene-level score that reflects changes in the expression of a given gene and its upstream neighbors [17]. Nevertheless, the suggested analysis is performed on each pathway separately rather than on the global network. Hence, later methods were designed to avoid these disadvantages by using direct interactions in a global network to produce gene-level scores.

One method, Local Enrichment Analysis (LEAN) [3], identifies dysregulated subnetworks from genome-wide omics datasets by focusing on local subnetworks of radius one which consist of only the direct neighbors of genes. The method is parameter free and exhaustive over all genes in the network. Another method, *pathfindR* [18], extends LEAN by letting the user specify the radius of local subnetworks to be enriched using three possible algorithms: Greedy Algorithm (GD), Simulated Annealing (SA), and

Genetic Algorithm (GA). The authors show that GD performs better than SA and GA. SA and GA are heuristic methods that do not make biologically-relevant assumptions on the active subnetwork model. Insignificant genes from two modules of significant genes may thus be combined to form a larger connected active subnetwork. This results in few large and high-scoring active subnetworks with the remaining subnetworks being small and less informative. In short, these heuristic methods exhibit a tendency towards large subnetworks which is attributed to a statistical bias that is prevalent in many tools [19]. A method attempting to solve this issue, *jActiveModules*, uses user-defined parameters to control the number of subnetworks that are maintained throughout simulated annealing as well as the behavior of the method when adding nodes with a degree above threshold to a subnetwork [8]. Nevertheless, this method does not produce gene-level scores. It uses static heuristics as user-specified thresholds, and optimizes subnetworks based on a calibrated average measure of differential expression which might not necessarily reflect nonlinear patterns of differential expression. In addition, MultiNEP is a recent network-based approach that analyzes muti-omics datasets in order to identify disease-related subnetworks [20]. It considers gene-metabolite interactions in order to identify disease-related genes. Similarly, GMIGAGO is a gene module identification method that is based on gene ontology and Genetic Algorithm [21]. It starts by clustering gene expression data and then detects gene modules by optimizing functional similarity based on gene ontology.

Another recent approach is GeneSurrounder (GS) [22]. It is an exhaustive method in the sense that it considers the decay of differential expression (DE) and the sphere of influence of a gene. The sphere of influence measures the correlation between the behavior of a gene of interest and that of its direct surrounding neighbor genes, regardless of the phenotype. On the other hand, the decay of DE measures a pattern of decrease in magnitude of disease-specific disruption up to a certain distance from the gene. The optimal radius R that identifies the effect of the gene on its neighbors is given by the combination of the two p values which are based on the Fisher method (p^{fisher} from p^D and p^S). p^S represents the p value for the sphere of influence which reflects the correlation between a center gene's expression intensity and its neighbors. On the other hand, p^D represents the p value for the decay of DE which reflects the discordance between DE scores of the genes included in the module and their distances from the center gene. GS achieves meaningful DEA and pathway enrichment results, and exhibits scores that are more concordant than both *limma* and LEAN across three studies of the same disease, namely ovarian cancer. However, the search complexity and implementation of GS requires further development and optimization to be favorable for common use as well as to yield refined results. In the original implementation, GS checks for the optimal radius R of each gene's module by choosing the radius with the lowest adjusted p^{fisher}. Nevertheless, genes could significantly influence just some of the neighboring genes on a particular radius, and this information is not always available in the form of weighted networks. Therefore, finding the optimal module could be seen as a combinatorial problem, as done in the heuristic methods of *pathfindR*. In this direction, we design and implement a biologically-informed heuristic method based on the Ant Colony Optimization (ACO) algorithm. The method takes as input the network and gene-expression data. Genes are sequentially considered and parallelized heuristic searches are then performed, on the basis of decay of differential expression, to identify candidate modules centered around each seed gene. The sphere of influence of each resulting module is then assessed to produce a combined *p* value. The best-scoring module is selected for further downstream analyses. Unlike GS, our implementation does not impose radius-based restrictions on the search space but rather distance penalties to avoid large modules. Moreover, this method, which, unlike *pathfindR*'s SA and GA, functions at the gene-level, is performed on each gene individually, similarly to GS, where each gene is assessed as the center of an optimized module solution. Accordingly, if a gene is part of a small insignificant module centered around a different gene, its own module will be independently assessed for significance, i.e., it will still be considered as a seed gene. That is since a module's score is directly related to its designated center, i.e., seed, gene, similarly to GS.

We test the method on three publicly available benchmark microarray datasets of different diseases, and show that it generally results in more meaningful and stable disease-relevant enrichment outcomes. Since the optimization described here is performed on the basis of pre-calculated differential expression scores, the method can be easily adapted to other gene expression data sources, such as RNAseq, by simply replacing the DEA tool in the built-in preprocessing with another appropriate choice. A comparison of the characteristics of the proposed approach with GS, LEAN, and *limma* is presented in Table 1. The compared aspects cover whether the method is network-based, uses radius while searching for modules, includes a gene-level score, and contains randomness in its implementation. Similarly to GS, we also check the concordance of gene-level *p* values, given by the output of each method, between pairs of three different subsets of an Alzheimer's disease dataset, GSE5281, whereby each subset is collected from a different brain region. Nevertheless, we focus our analysis on datasets with a case–control design instead of differing types of the same disease.

Methods

In this section, we present the methods used in the proposed approach. First, we start with a description of the general framework of Ant Colony Optimization. Next, we define the measures of module significance, and we define the quantification of biological influence of dysregulated genes. Then, we explain the workflow steps, starting with the gene expression and interconnection datasets as input, up till the generation of candidate dysregulated modules. A diagram depicting these steps is shown in Fig. 1. The structures of the corresponding input and output matrices are visualized in Fig. 2.

Table 1 Tested methods and some basic characteristics

Method	Network-based	Radius	Gene-level score	Contains randomness
Proposed method	Yes	Not radius-based	Yes	Yes
GeneSurrounder	Yes	Gene-specific R	Yes	Yes
LEAN	Yes	$R = 1$	Yes	Yes
limma	No	N/A	Yes	No

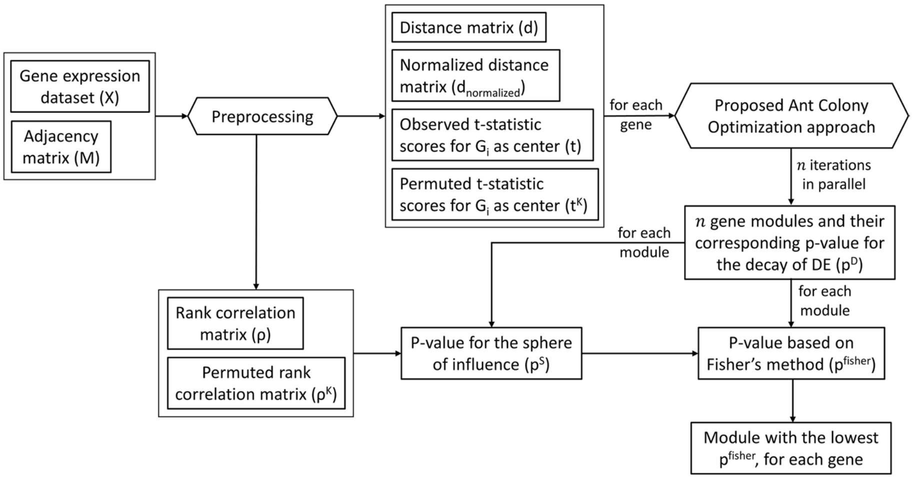
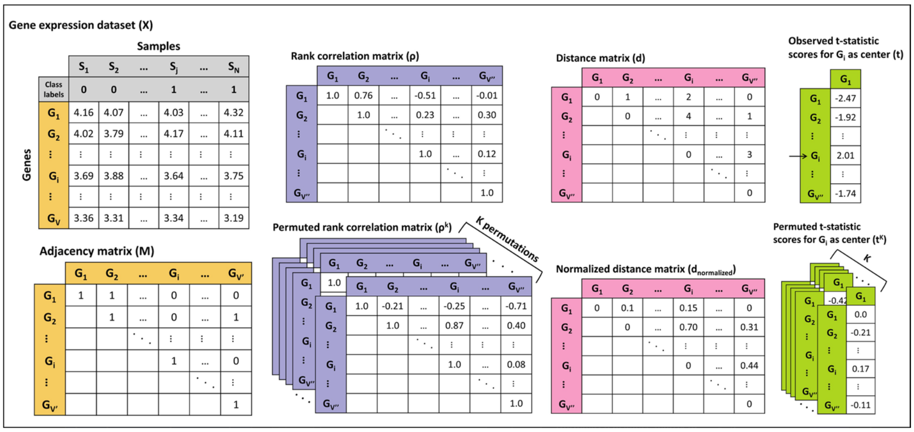

Gene expression dataset (X)
Adjacency matrix (M)
Preprocessing
Distance matrix (d)
Normalized distance matrix ($d_{normalized}$)
Observed t-statistic scores for G_i as center (t)
Permuted t-statistic scores for G_i as center (t^K)
for each gene
Proposed Ant Colony Optimization approach
n iterations in parallel
n gene modules and their corresponding p-value for the decay of DE (p^D)
for each module
for each module
Rank correlation matrix (ρ)
Permuted rank correlation matrix ($ρ^K$)
P-value for the sphere of influence (p^S)
P-value based on Fisher's method (p^{fisher})
Module with the lowest p^{fisher}, for each gene

Fig. 1 An overview of the proposed workflow which takes as input a gene expression dataset and an adjacency matrix representing gene interconnections. A built-in preprocessing step generates distance and rank correlation matrices, in addition to t-statistic scores reflecting the magnitude of differential expressions of gene in each module. Then, for each gene, the proposed Ant Colony Optimization approach identifies the most significantly dysregulated module, based on p^{fisher} which is derived from p^S, the p value for the sphere of influence of a center gene on its neighbors in a module, and p^D, the p value for the decay of differential expression within that module



Gene expression dataset (X)

	S_1	S_2	...	S_j	...	S_N
Class labels	0	0	...	1	...	1
G_1	4.16	4.07	...	4.03	...	4.32
G_2	4.02	3.79	...	4.17	...	4.11
⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_i	3.69	3.88	...	3.64	...	3.75
⋮	⋮	⋮	⋮	⋮	⋮	⋮
G_V	3.36	3.31	...	3.34	...	3.19

Rank correlation matrix (ρ)

	G_1	G_2	...	G_i	...	$G_{V''}$
G_1	1.0	0.76	...	-0.51	...	-0.01
G_2		1.0	...	0.23	...	0.30
⋮			⋱	⋮	⋮	⋮
G_i				1.0	...	0.12
⋮					⋱	⋮
$G_{V''}$						1.0

Distance matrix (d)

	G_1	G_2	...	G_i	...	$G_{V''}$
G_1	0	1	...	2	...	0
G_2		0	...	4	...	1
⋮			⋱	⋮	⋮	⋮
G_i				0	...	3
⋮					⋱	⋮
$G_{V''}$						0

Observed t-statistic scores for G_i as center (t)

	G_1
G_1	-2.47
G_2	-1.92
⋮	⋮
→ G_i	2.01
⋮	⋮
$G_{V''}$	-1.74

Adjacency matrix (M)

	G_1	G_2	...	G_i	...	$G_{V'}$
G_1	1	1	...	0	...	0
G_2		1	...	0	...	1
⋮			⋱	⋮	⋮	⋮
G_i				1	...	0
⋮					⋱	⋮
$G_{V'}$						1

Permuted rank correlation matrix ($ρ^k$)

	G_1	G_2	...	G_i	...	$G_{V''}$
G_1	1.0	-0.21	...	-0.25	...	-0.71
G_2		1.0	...	0.87	...	0.40
⋮			⋱	⋮	⋮	⋮
G_i				1.0	...	0.08
⋮					⋱	⋮
$G_{V''}$						1.0

Normalized distance matrix ($d_{normalized}$)

	G_1	G_2	...	G_i	...	$G_{V''}$
G_1	0	0.1	...	0.15	...	0
G_2		0	...	0.70	...	0.31
⋮			⋱	⋮	⋮	⋮
G_i				0	...	0.44
⋮					⋱	⋮
$G_{V''}$						0

Permuted t-statistic scores for G_i as center (t^K)

	G_1
G_1	0.0
G_2	-0.21
⋮	⋮
G_i	0.17
⋮	⋮
$G_{V''}$	-0.11

Fig. 2 The structures of input and output matrices in the proposed approach

Ant colony optimization

Ant Colony Optimization (ACO) is one of numerous meta-heuristic algorithms inspired by swarm intelligence, in this case the foraging behavior of ants. It is widely used to tackle combinatorial problems [23]. The inspiration stems from the observed efficiency in which ants conduct their search for food starting from the nest. Biological ants display a type of communication known as stigmergy. The main characteristics of stigmergy arise from the medium employed in this type of communication: pheromones. Pheromone deposits reflect the quality of the achievement and are then a means of indirect communication between ants. Moreover, pheromone deposits are local and transmit information between ants within a locus. Pheromones not only indirectly reflect the end result of an explored path but also its length. Although random fluctuations exist early on during the search, ants usually deposit pheromones faster after returning to the nest from the shorter path. In this way, ants converge to the shortest path. In ACO algorithms (Algorithm 1) [23], a model $P = (S, \Omega, f)$ consists of the search space S whereby a feasible solution $s \in S$ satisfies all constraints in a set Ω of constraints over the finite set of discrete variables defining S. A globally optimal solution $s^* \in S$ additionally minimizes a given objective function $f : S \rightarrow \mathbb{R}_0^+$ (i.e. $f(s^*) \leq f(s) \forall s \in S$). The set of all possible solution components (i.e. all possible variable assignments in S) is denoted by C. Each component of a solution $s \in S$ is associated with a pheromone value that varies with quality and evaporates at every iteration. A component is represented by either vertices from a set of vertices V or edges from a set of edges E of a construction graph $G_C(V, E)$.

Algorithm 1 The Ant Colony Optimization Metaheuristic

```
Set parameters, initialize pheromone trails
while termination condition not met do
    Construct ant solutions
    Apply local search                                    ▷ Optional
    Update pheromones
end while
```

At the start of an iteration, a set of A ants construct solutions by traversing the graph in a manner that satisfies constrains in Ω followed by an optional local search. In Ant Systems, the decisions made by a given ant during its construction walk are governed by a stochastic process influenced by the pheromones allocated to possible components. ACO systems, however, use a pseudorandom mechanism that encourages elitism by deterministically picking the most probable component if a random number ranging from 0 to 1 falls under a user-specified threshold. Otherwise, decisions are made similarly to Ant Systems. Pheromones could be deposited in different ways. In Ant Systems, an offline pheromone update occurs at the end of each iteration after solutions have been constructed. ACO systems additionally employ a local pheromone update that decreases pheromone concentration on the last visited edge after each construction step performed by the ant with the current best solution. This serves to offset the offline pheromone update and encourage diversity. Evaporation is also performed using a user-specified or dynamic parameter to guide ants towards shorter and more frequently explored paths. In our formulation, each ant produces a candidate solution in each iteration, but these are independent from solutions generated at other iterations. In other words, information is not carried over between iterations, and only local pheromone updates are performed. Given the nature of the problem whereby the objective function is module-centric, an offline pheromone update would encourage convergence towards high-quality paths, which could bias the topology of generated modules. For the same reason, the final solution, or module, generated at each iteration is the union of the individual ant solutions (i.e. sets of vertices) explored in this iteration. Finally, we introduce an alloted capacity to each ant at every iteration. This capacity diminishes as the ant makes increasingly unfavorable moves, causing the ant to stop moving when its capacity is too low. Therefore, an iteration automatically terminates when all ants have no more capacity for movement.

Module significance quantification

The Order Statistic Correlation Coefficient (OSCC) [24] is a measure that can be used to detect linear and monotone nonlinear associations. It possesses the same basic characteristics as Pearson's linear, Spearman's ρ, and Kendall's τ coefficients. OSCC also exhibits robustness to noise and efficient time complexity ($O(n \log n)$), when compared to Kendall's τ ($O(n^2)$). The calculation of OSCC is given in equation 1 using two paired input arrays x and y.

$$OSCC(x, y) \triangleq \frac{\sum_{i=1}^{N}(x_{(i)} - x_{(N-i+1)})y_{[i]}}{\sum_{i=1}^{N}(x_{(i)} - x_{(N-i+1)})y_{(i)}} \tag{1}$$

The paired input arrays x and y are first ordered such that the order statistics $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$ have respective concomitants being $y_{[1]}, y_{[2]}, \cdots, y_{[N]}$. The order statistics and concomitants of the input y array are similarly defined. As $N \to \infty$, $E\{OSCC(x, y)\} = 0$ under the assumption that x and y are mutually independent and both are independently identically distributed.

In this application, similarly to how GS uses Kendall's τ to score the decay of DE of modules [22], OSCC is used to score a module centered at a given gene G_i and denoted by the set module as shown in equation 2. The geodesic distance is the number of edges in the shortest path to the center gene. OSCC outputs a score ranging from -1 to 1. A module having a score of -1 is optimal, since the discordance between absolute moderated t-statistics, representing the magnitude of DE of genes belonging to a given set *module*, and the geodesic distances of those genes from the center gene G_i is maximal according to the OSCC.

$$OSCC(module) = OSCC(\{|t_j| : G_j \in module\}, \{d(G_i, G_j) : G_j \in module\}) \tag{2}$$

We denote the change in OSCC with and without the inclusion of a given node G_p to a set *module* by $\triangle OSCC(module \cup \{G_p\})$ which is calculated in equation 3. $\triangle OSCC$ outputs a value ranging from -2 to 2, with -2 being the greatest possible change in OSCC in the favorable direction, that is from 1 to -1.

$$\triangle OSCC(module \cup \{G_p\}) = OSCC(module \cup \{G_p\}) - OSCC(module) \tag{3}$$

Biological influence quantification

In this part, we explain how GS [22] quantifies biological influence at the gene-level through the incorporation of system-level network information. Then, we present our approach to derive such measures in the next section. A gene score is defined as a combination of two scores representing the decay of DE and the sphere of influence detected in neighboring genes that are selected on the basis of a variable radius R. The sphere of influence score indicates that a gene influences its neighbors such that their expression intensities are correlated. The decay of the DE score indicates how the dysregulation of a disease-relevant gene is propagated to its neighbors in a decreasing pattern whereby the level of dysregulation is inversely proportional to the distance from the given gene. Since the extent to which a gene influences its surrounding neighbors with respect to both scores is unknown, all possible values for *R* are considered. The pseudocode for GS is given in Algorithm 2.

Algorithm 2 GeneSurrounder

```
**Input:** dataset $X$ of size $V$x$N$, adjacency matrix $M$ of size $V'$x$V'$, $class\_labels$, $K \geq 1$, $seed$
**Output:** $p^{GS}, R^{GS}$
$\rho \Leftarrow spearmanRank(X)$
$\rho^{1:K} \Leftarrow spearmanRank(permutate(X, K, seed)$
$t \Leftarrow limma(X, class\_labels)$
$t^{1:K} \Leftarrow limma(X, permutate(class\_labels, K, seed))$
$d \Leftarrow distance(M)$
$diam \Leftarrow diameter(M)$
$p^{GS}, R^{GS} \Leftarrow$ array of size $= \min(V, V')$
for each gene $G_i$ present in both $X$ and $M$ do
    $p^{fisher} \Leftarrow$ array of size $= diam$
    for each possible value for radius $R$ do
        $p^{fisher}[R] \Leftarrow p_i^{fisher}(R, \rho_i, \rho_i^{1:K}, t, t^{1:K}, d)$
    end for
    $p^{GS}[i] \Leftarrow p_i^{GS}(p^{fisher}, diam)$
    $R^{GS}[i] \Leftarrow which.min(p^{fisher})$
end for
return $p^{GS}, R^{GS}$
```

Given a gene G_i, for each possible value for *R*, all neighboring genes having a minimum distance from G_i less than or equal to *R* are selected for the module centered at G_i. The combined score is then calculated for the candidate modules, and the optimal value for *R* is chosen as the one with the highest statistical significance.

Proposed method

The choice to base our implementation on ACO is mainly due to the problem being a local search [23]. The main modification applied to ACO was the introduction of a limited capacity for movement assigned to each ant to automatically terminate iterations. This capacity diminishes as the ant moves to nodes further from the center gene and nodes that unfavorably impact the module's score. Another important modification to ACO is that although the search is carried out over a specified number of iterations, each iteration represents an independent search that yields a candidate result. In other words, there is no exchange of information between different iterations. This is because the aim is not for ants to converge to similar paths but to spread and explore different paths outwards from the designated center node. Moreover, since each gene's score is a combination of both scores for the decay of DE and the sphere of influence, we optimize the scores for the decay of DE rather than optimize both scores. This is because the module of neighboring genes that exhibit the highest correlation of gene expression intensities with those of the center gene are expected to be relevant if they first show that they are influenced by the center gene in a disease-relevant context through the decay effect. The score for the sphere of influence is calculated for the module resulting from each iteration, and the optimal module is selected as the one having the combined score with the highest statistical significance, similarly to GS.

Given N samples $S_1, S_2, \cdots, S_j, S_{j+1}, \cdots, S_N$ having N class labels as well as a graph of V' nodes and E edges represented as an adjacency matrix M, we define a microarray dataset X of V genes $G_1, G_2, \cdots, G_i, G_{i+1}, \cdots, G_V$ and N samples. The input to the algorithm are M, X and its class labels, a seed for reproducibility (different than the seed gene), as well as several configuration parameters. The pseudocode is given in Algorithm 3.

Algorithm 3 Proposed Method

```
Input: dataset $X$ of size $V$x$N$, adjacency matrix $M$ of size $V'$x$V'$, $class\_labels$, $K \geq 1$, $n \geq 1$, $A \geq 1$, $\alpha > 0$, $\beta > 0$, $n\_cores \geq 1$, $starting\_capacity > 0$, $seed$
Output: $p^a, modules^a$
$\rho \Leftarrow spearmanRank(X)$
$\rho^{1:K} \Leftarrow spearmanRank(permutate(X, K, seed)$
$t \Leftarrow limma(X, class\_labels)$
$t^{1:K} \Leftarrow limma(X, permutate(class\_labels, K, seed))$
$d \Leftarrow distance(M)$
$d_{normalized} \Leftarrow normalizeRowWise(d)$
$p^a, modules^a \Leftarrow$ array of size $= \min(V, V')$
for each gene $G_i$ present in both $X$ and $M$ where $is.isolated(G_i) = FALSE$ do
    $module, p^{fisher} \Leftarrow$ vector of size $= n$
    for $iter$ in $1 : n$ iterations doParellel($n\_cores$)
        $Pher \Leftarrow 1.0$ matrix of size $= V'$x$V'$
        for each ant $a$ in $1 : A$ ants do
            $Capacity(a) \Leftarrow starting\_capacity$
            $position(a) \Leftarrow G_i$
        end for
        while any of $1 : A$ ants is moveable do
            Shuffle ant order if all ants picked this cycle
            pick first moveable ant $a$
            assume $position(a) = G_j$
            $possibilities \Leftarrow possibilities(a, d, d_{normalized}, t)$
            if $possibilities \neq \phi$ and $Capacity(a) > 0$ then
                if $iter == 1$ then
                    $G_p \Leftarrow sample(possibilities, seed)$
                else
                    $G_p \Leftarrow sample(P_j(a, possibilities, Pher, \alpha, \beta), seed)$        ▷ eq. (6)
                end if
                $position(a) \Leftarrow G_p$
                $Pher_{jp} \Leftarrow update(Pher_{jp}, a, d, d_{normalized}, t)$        ▷ eq. (5)
                $Capacity(a) \Leftarrow update(Capacity(a), d, d_{normalized}, t)$        ▷ eq. (7)
            else
                ant $a$ no longer moveable
            end if
        end while
        $module[iter] \Leftarrow$ set of nodes visited by any of $1 : A$ ants
        $p^{fisher}[iter] \Leftarrow p_i^{fisher}(module, \rho_i, \rho_i^{1:K}, t, t^{1:K}, d)$        ▷ eq. (8)-(11)
    end for
    $p^a[i] \Leftarrow BenjaminiHochberg(\min(p^{fisher}))$
    $modules^a[i] \Leftarrow module[which.min(p^{fisher})]$
end for
return $p^a, modules^a$
```

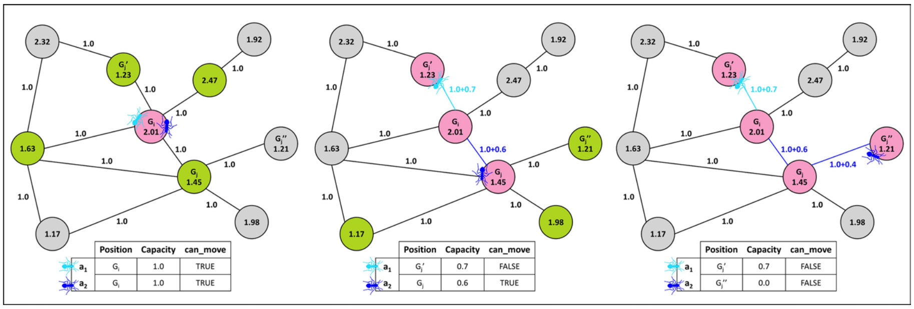
An built-in preprocessing step, as depicted in Fig. 1, filters dataset X and M to only include genes present in both X and M, calculates a Spearman rank correlation matrix ρ and ρ^k from dataset X as well as a user-specified number K of random permutations of genes in X, calculates a distance matrix d from matrix M, conducts a DEA over dataset X using true class labels as well as a user-specified number of random permutations K of class labels in X, and designates the resulting moderated t-statistic t_i and t_i^k as the observed and permuted scores, respectively, for each vertex G_i.

The algorithm is repeated a maximum of $\min(V, V')$ times, each time given a specified gene G_i and number of iterations n, and outputs n modules centered at G_i with each module being assigned a combined p value adjusted for n iterations similarly to p_i^{GS} except using Benjamini-Hochberg (BH) adjustment for a less stringent correction than the Bonferroni adjustment. A gene G_i is skipped if it is detected as being an isolated node (i.e. if G_i has no edge connected to another gene G_j where $i \neq j$).

As in the classic ACO algorithm [23], a designated number of ants A are given a starting point. In this case, the starting point is the designated center gene G_i. Figure 3 depicts the start of an iteration in the algorithm for a gene G_i as the starting point in a network whereby t indicates the score of each vertex.

In our method, at the start of every iteration, the ants are placed at the starting point. In the first iteration, the ants move randomly. Although iterations in this implementation are independent, the iteration marked as first is still randomized and considered as a viable candidate solution. This iteration could also be used as a reference in future analyses that might investigate the topology of resulting modules. This is because modules in the first iteration are mainly shaped by capacity rather than pheromone, both of which are discussed below. Nevertheless, since investigating how the topology of resulting modules is formed is not the focus of the study, the random iteration is not investigated further and is treated like any other iteration.

At each move, an ant a picks a possible node G_p directly neighboring (i.e., one edge away from) its current position G_j and assign a value to the explored edge as a representation of pheromone. For example, Fig. 3 depicts a snapshot mid-iteration where ant



	Position	Capacity	can_move
a_1	G_i	1.0	TRUE
a_2	G_i	1.0	TRUE

	Position	Capacity	can_move
a_1	G_j'	0.7	FALSE
a_2	G_j	0.6	TRUE

	Position	Capacity	can_move
a_1	G_j'	0.7	FALSE
a_2	G_j''	0.0	FALSE

Fig. 3 An example of two ants searching for gene modules in the proposed Ant Colony Optimization approach. The networks from left to right depict consecutive iterations with gene G_i as the starting point. Below every network, a table shows the position, capacity, and moving ability of ants. At each step, an ant moves to an adjacent gene, and updates the pheromone value on the traversed edge based on the favorability of the move. Visited nodes, colored in pink, constitute the module centered around G_i. Possible adjacent nodes to visit are colored in green nodes, while the rest of the nodes in the network are colored in grey. Note that given the proposed algorithm constraints, ants a_1 and a_2 become unable to move in the middle and the right networks, respectively

a_1, positioned on $G_j{}'$ is no longer able to move and ant a_2 positioned on G_j is assessing its possible movements after local pheromone updates are performed for the previous movement. The pheromone update value reflects the favorability of the move with respect to the objective function and is calculated using Eq. 4 and 5.

$$EL(module(a) \cup \{G_p\}) = \frac{d_{normalized}(G_i, G_p)}{4}(2 + \triangle OSCC(module(a) \cup \{G_p\})) \tag{4}$$

$$update(Pher_{jp}, a) = Pher_{jp} + (1 - EL(module(a) \cup \{G_p\})) \tag{5}$$

The nodes currently marked as visited by ant a constitute the set $module(a)$ centered at G_i. The values for pheromones present on each edge are recorded in a pheromone matrix $Pher$ that resets at every iteration to prevent information exchange between iterations. Nevertheless, pheromones transfer information between ants within the same iteration. An ant deposits pheromone mid-iteration as it crosses an edge with a local pheromone update similarly to the ACO system [23]. Rather than decreasing pheromones locally, ants exchange decision-quality information locally. In this way, diversification of trajectories explored by ants *between* iterations is encouraged with no offline information guiding ants across iterations. This is because the goal is to produce many candidate modules that are qualitatively different before choosing the best one. Moreover, another difference is that all ants are capable of depositing pheromone, not just the best one. This modification is put in place to encourage diversification of trajectories taken by the ants *within* a given iteration. In other words, within an iteration, the ants are encouraged to take different paths which are combined to result in a module of varying topology. The pheromone map is initialized to have all values equal to 1.0 at the start of every iteration. Pheromone deposited on the same edge accumulates, and no evaporation is incorporated since the pheromone map resets every iteration. The energy lost or required by ant a to make a specific move from G_j to G_p is represented by $EL(module(a) \cup \{G_p\})$. This value is calculated using a row-wise normalized version $d_{normalized}$ of the distance matrix d. EL returns a value in the interval [0,1], and the favorability of the move is expressed as $1 - EL$.

For each iteration except the first, the ants move probabilistically using Eq. 6 to calculate the probability of making a certain move from G_j to G_p. This is done similarly to Ant systems rather than an ACO system in order to avoid increasing elitism in the probability function [23].

$$P_{jp}(a) = \frac{(Pher_{jp})^{\alpha}(2 - EL(module(a) \cup \{G_p\}))^{\beta}}{\sum_{\{G_q : G_q \in possibilities(a)\}}(Pher_{jq})^{\alpha}(2 - EL(module(a) \cup \{G_q\}))^{\beta}} \tag{6}$$

The parameters α and β are user-specified and reflect the weight given to pheromone and attractiveness of the move, respectively. In other words, in a given iteration, α reflects how similarly the ants will move, and β reflects how much importance an ant will give to more favorable moves rather than moves similar to other ants. As α is increased, the ants are expected to produce modules that increasingly converge to a single path. Increasing β will allow the ants to explore different paths and decrease the constraints over the topology of the network. If the pattern of decay can be observed over numerous paths going away from the center (seed) gene, the ants are more likely to explore them. Moreover, the more ants are available, the more different favorable paths are likely to be explored. This could also be achieved by increasing the number of iterations to produce more candidate modules, but this could be more time-intensive than increasing the number of ants. Moreover, increasing the number of ants would impose less restrictions over how many different paths can be explored.

The probabilities are calculated for each possible movement. Possible movements are defined as nodes that are one edge away from the current position of an ant a and are represented by $possibilities(a)$. Nodes that have already been visited by a during the current iteration, nodes that would diminish the capacity of a, $Capacity(a)$, below 0 if visited, as well as nodes that keep a at the same distance from or bring a closer to G_i (i.e. $d(G_i, G_p) \leq d(G_i, G_j)$) are excluded from the set $possibilities(a)$. This helps short-list possible movements as well as guide ants to spread outwards from the center.

The main difference between ACO and our implementation is that a capacity measure is introduced. Each ant is specified a starting capacity that diminishes on each move the ant makes to a node G_p according to Eq. 7.

$$update(Capacity(a)) = Capacity(a) - EL(module(a) \cup \{G_p\}) \tag{7}$$

When an ant's capacity reaches 0 or it has no more possible movements to consider (i.e. $possibilities(a) = \phi$), the ant stops moving. An iteration ends when all ants are no longer able to move. The resulting module centered around G_i for this iteration are all nodes visited by at least one ant and is denoted as *module*. Using Eqs. 8–11, *module* is scored and recorded at each iteration. p_i^{fisher} is extracted similarly to GS. Figure 3 also depicts the end of a single iteration n whereby both ants are unable to move, and the module produced includes all nodes visited by any of the two ants. Significance is then assessed, but not depicted, using ρ^k and t^k.

$$C_i(module) = \sum_{\{G_j : G_j \in module\}} |\rho_{ij}| \tag{8}$$

$$p_i^S(module) = \frac{1}{K} |\{k \in [1, K] \cap N^* : C_i^k(module) > C_i(module)\}| \tag{9}$$

$$p_i^D(module) = \frac{1}{K} |\{k \in [1, K] \cap N^* : OSCC_i^k(module) < OSCC_i(module)\}| \tag{10}$$

$$\chi^2 = -2(\ln p_i^S(module) + \ln p_i^D(module)) \tag{11}$$

To further offset the added computational time of the proposed method as compared to GS, for each gene G_i, the n iterations are parallelized, as depicted in Fig. 1, using the R package parallel's *mclapply* [25]. Therefore, parallelization is only supported for Mac users. No parallelization is implemented between genes. We also use the order statistic correlation coefficient (OSCC) [24] which is more time-efficient ($O(n \log n)$) than Kendall's tau-b ($O(n^2)$), that is used in GS, in addition to numerous distance penalties and search restrictions, as described above, to constrain the search-space.

Experimentation

Datasets

To test our approach, we rely on the R package *GSEABenchmarkeR* [26] of Bioconductor [27]. This package includes curated expression datasets that are related to various human diseases. It was developed to facilitate the assessment and comparison of enrichment analysis methods. We consider a total of four datasets that have different case to control proportions and sample sizes. The characteristics of these datasets are shown in Table 2 which covers the corresponding GEO dataset ID, disease, total number of samples, total number of control samples, total number of mapped genes in the expression dataset, number of common genes between the expression dataset and the KEGG network, the diameter of the resulting largest component of the KEGG network, and the total number of isolated genes which are not part of the connected component. Three of them are related to neurodegenerative diseases, namely Parkinson's (GSE20291) [28], Alzheimer's (GSE5281) [29], and Huntington's (GSE8762) [30] disease. The latter allows testing the stability of our approach with respect to different dataset characteristics. All three neurodegeneration datasets are already preprocessed by removing outlier arrays, applying a log transform, applying RMA normalization [31] from the *affy* [32] R package, resolving duplicate probe to Entrez ID mappings by keeping the probe with the most significant *limma* [1] moderated t-statistic, and filtering out genes that could not be mapped to any KEGG [33] pathway [26, 34, 35]. The fourth dataset is the p53 mutation dataset from the NCI-60 cell lines originally available through the GSEA Broad Institute website [4, 36]. This dataset is used to ensure the method performs comparably better than other tested methods on a dataset unrelated to neurodegenerative diseases. The dataset is imported through the R package *GSAR* [37]. The p53 mutation dataset is also already preprocessed by quantile normalizing and log-transforming probe intensities, discarding probes without Entrez ID mappings, and resolving duplicate probe to Entrez ID mappings by keeping the probe with the largest absolute t-statistic between cases and controls [37].

Gene-gene interaction network

An initial step in our approach consists of generating the gene-gene interaction network that corresponds to the input gene expression dataset. Accordingly, we use the R

Table 2 Microarray gene expression datasets considered in this study along with their basic characteristics

GEO dataset ID	Disease	Total # of samples	# of control samples	Total # of mapped genes	# of genes common with KEGG network	KEGG network diameter	# of Isolated Genes
GSE20291	Parkinson's	33	19	13039	3718	18	23
GSE5281_VCX	Alzheimer's	31	12	21367	4302	20	5
GSE5281_HIP	Alzheimer's	23	13	21367	4302	20	5
GSE5281_EC	Alzheimer's	21	12	21367	4302	20	5
GSE8762	Huntington's	22	10	21405	4297	20	5
–	p53 mutation	50	17	8655	2913	18	64

The network used in the table corresponds to the largest connected component extracted from KEGG package *KEGGREST* [38] of Bioconductor [27] to get the set of unweighted and undirected KEGG [33] pathways. Next, we combine those individual pathways into one network, using a graph union operation, and we extract the largest connected component using the Python package *NetworkX* [39]. Then, this component is filtered to only cover that genes assayed in the expression dataset.

Experimental setup

To assess our proposed method, each dataset is used for the proposed heuristic, GS [22], and LEAN [3] using the respective KEGG [33] network. *limma* [1] is also run as the baseline tool. The parameter values used for the proposed heuristic are shown in Table 3. They consist of the number of ants per iteration, the number of iterations per gene, the starting capacity of each ant, the number of cores for parallelization, alpha and beta parameters, in addition to the number of random permutations. Several values for the number of ants are tested (results are not included), and the value of 40 is chosen through trial-and-error. It is generally observed that increasing the number of ants improved results although no value above 40 is tried. Hence, it could be useful to further investigate the impact of this parameter, and others, on the algorithm's behavior in future studies. The starting capacity is always 1.0, the number of iterations n is always 5, the number of cores used for the parallelization of iterations is always 5, and α and β were fixed to 0.6 and 1.2, respectively. In line with the formulation of ant colony proposed here which encourages diversification of paths taken by ants rather than their convergence within a given iteration, we set a higher value for hyperparamter β (1.2) as compared to α (0.6) so that, within an iteration, ants are influenced more by the favorability of their own route rather than that of the other ants. After several experimentations, we found that these parameter values gave the best results. The random seed is set to 1, 2, or 3 in each experiment, respectively, for reproducibility. For GS, the random seed is set to 121, 122, and 123 in each experiment, respectively. *limma* and LEAN are both employed using default configurations and random seed being set to 1, 2, or 3 in each experiment, respectively, for LEAN. The number of random permutations used for non-parametric statistical significance assessment within the tools is set to 100. The proposed method is observed to be stable for these datasets across the experiments. The methods are also only run for a single experiment on the p53 mutation dataset using the same corresponding random seeds as those of Experiment 3. The statistical significance cutoff is set to 0.05 for all experiments.

Table 3 Parameter settings used for the proposed heuristic method

Parameter	Description	Value
n_ants	# of ants per iteration per gene	40
n_iter	# of iterations per gene	5
starting_capacity	Starting capacity allotted per ant, resets every iteration per gene	1.0
n_cores	# of cores used for parallelization of iterations per gene (only for Mac users)	5
alpha	α parameter of probability function (Eq. 6)	0.6
beta	β parameter of probability function (Eq. 6)	1.2
n_resamples	# of random permutations used for p_i^S and p_i^D non-parametric estimations	100

An enrichment analysis was then conducted by testing whether genes of a given KEGG pathway exhibit p values that are lower than the background genes represented by all other genes. A total of 353 KEGG pathways were extracted using the R package *KEGGREST* [38]. In order to infer the statistical significance of our results, the Wilcoxon rank-sum tests were used along with BH adjustment to correct for multiple testing. Accordingly, we are able to examine whether the underlying distributions and findings would still hold when the data is rearranged and shuffled. The significance and ranking of disease-relevant pathways are then compared between tested methods for each dataset. Disease-relevant pathways were selected according to KEGG pathways listed as relevant under the KEGG [33] entry for each disease. This analysis constitutes a single experiment. This process was repeated ten times for each neurodegeneration dataset with different seed values to check the relative stability of the tested methods since all except *limma* include one or more random components. Therefore, a total of ten experiments per neurodegeneration dataset were conducted as compared to a single experiment for the p53 mutation dataset. Note that for the enrichment analysis, only the GSE5281_VCX subset was used to represent Alzheimer's disease since it is the one with the most samples (33) out of the three subsets considered.

Finally, concordance was assessed between pairs of 3 different subsets of the Alzheimer's dataset (GSE5281) [28]. Each subset consists of samples taken from different brain regions relevant to this disease. In this case, the visual cortex subset (GSE5281_VCX), the hippocampus subset (GSE5281_HIP), and the entorhinal cortex subset (GSE5281_EC), as divided in the *GSEABenchmarkeR* [26] package, were considered. As in GS [22], Spearman's ρ was used to assess concordance of gene-level p values between different dataset pairs for each of the tested methods.

Results

For each dataset, we run ten independent experiments for the proposed method, GS, and LEAN, in addition to one experiment for *limma* as a baseline. We report the ranks and significance levels of disease-related pathways which are identified based on the KEGG database. We focus on the main disease pathways, which are "Parkinson's disease", "Alzheimer's disease", and "Huntington's Disease", in Tables 4, 5, and 6 respectively. These tables show whether each method was able to detect the main corresponding disease pathway as significant, the mean rank of that pathway across the ten experiments, the mean number of significant pathways that are found across the ten experiments, and the mean proportion of significant pathways that are related to the disease across the experiments. The standard deviation of each entry is reported between parentheses. To explicitly examine the results, we also include detailed tables showing the names, ranks, and p values of all disease-relevant pathways across three out these ten experiments (Tables 9, 10, 11, 12). Each relevant pathway, regardless of its statistical significance, has a reported p value for each tested method. All p values are adjusted for multiple testing, as described in previous sections. Pathways that do not satisfy the significance threshold of 0.05 are indicated as having the rank "NA". Finally, to measure the performance of the compared approaches, we use the Receiver-Operator- Characteristic (ROC) analysis to show the true positive rates and the false positive rates over all possible detection p value cutoffs. These curves are shown in Fig. 4 and are further discussed in the next sections.

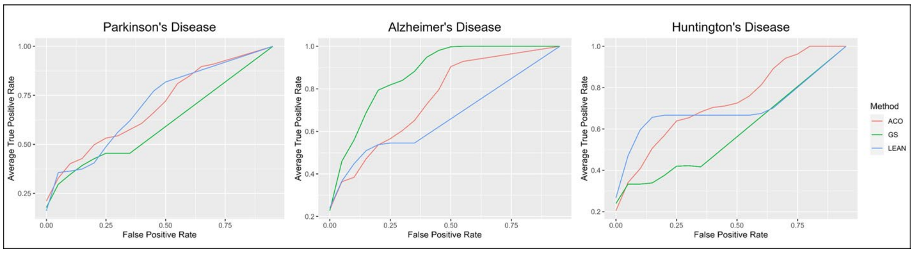


Fig. 4 ROC analysis to compare the performance of the proposed approach, GeneSurrounder, and LEAN

Table 4 Summarized results for the Parkinson's disease (PD) dataset across 10 experiments

Method	Found PD pathway	Mean ranks of PD pathway	Mean # of significant pathways	Mean proportion of significant pathways related to PD
Proposed Method	Yes	3.33 (0.52)	40.33 (3.14)	3.17 (0.41)
GeneSurrounder	Yes	6 (2.19)	56.17 (9.06)	3.83 (0.41)
LEAN	Yes	5 (2.37)	110 (5.33)	5.17 (0.75)

A total of 11 KEGG pathways related to PD are included in this analysis

The standard deviation of each entry is reported between parentheses

Table 5 Summarized results for the Alzheimer's disease (AD) dataset across 10 experiments

Method	Found AD pathway	Mean ranks of AD pathway	Mean # of significant pathways	Mean proportion of significant pathways related to AD
Proposed Method	Yes	7 (0)	83 (7.92)	5.33 (0.82)
GeneSurrounder	Yes	6.83 (0.41)	92.83 (11.89)	8.50 (1.22)
LEAN	Yes	8.33 (0.82)	83.17 (6.01)	6 (0)

A total of 11 KEGG pathways related to AD are included in this analysis

The standard deviation of each entry is reported between parentheses

Table 6 Summarized results for the Huntington's disease (HD) dataset across 10 experiments

Method	Found HD pathway	Mean ranks of HD pathway	Mean # of significant pathways	Mean proportion of significant pathways related to HD
Proposed Method	Yes	5 (0)	121.80 (9.09)	7.8 (0.45)
GeneSurrounder	Yes	1.60 (0.89)	61.60 (3.13)	4 (0)
LEAN	Yes	1.80 (1.79)	121.40 (6.19)	8 (0)

A total of 13 KEGG pathways related to HD are included in this analysis

The standard deviation of each entry is reported between parentheses

Parkinson's disease

For dataset GSE20291 [28], a total of 11 KEGG pathways that are related to Parkinson's disease are included in this analysis. All tested methods except *limma* consistently find the "Parkinson's disease" KEGG pathway (KEGG ID hsa05012) as statistically significant. As was reported in [26], *limma* fails to output even a single significantly DE gene for some datasets. Table 4 shows the summarized results of ten experiments on this dataset.

It reports the average and standard deviation values of the ranks of the “Parkinson’s disease” KEGG pathway, of the number of significant pathways, and of the mean proportion of significant pathways related to the disease. Our method achieves the highest mean rank of the “Parkinson’s disease” pathway and the lowest mean number of significant pathways across the ten experiments. LEAN achieves the highest mean proportion of significant pathways related to the disease but with the highest mean number of significant pathways. The Parkinson’s disease graph in Fig. 4 corresponds to areas under the curve (AUC) of 0.6592, 0.5707, and 0.6655 for the proposed method, GS, and LEAN, respectively. Although our approach scores a slightly lower AUC than LEAN, it tends to favorably detect a lower number of significant pathways (on average 40.33 vs 110) and scores the highest mean rank of the main disease pathway.

As reported in Tables 9, 10, 11, 12, no method finds hsa04141, hsa00350, hsa04120, and hsa04137 as significant in any of the experiments for this dataset. This is also the case for hsa04210 except that this pathway attains a *p* value that is slightly higher than the significance threshold for the proposed method in Experiment 1 (Table 9). The proposed method consistently ranks hsa05012, hsa05022, and hsa00190 in the top 10 across the experiments. Similarly, GS consistently ranks hsa05012 and hsa00190 in the top 10 across the experiments. Nevertheless, hsa05022 is assigned by GS a rank that ranges from 11th to 28th. Pathway hsa04020 is also significant for GS in 2 out of 3 experiments and ranks from 25th to 27th. Similarly to the proposed method, LEAN consistently ranks hsa05012 and hsa05022 in the top 10 across the experiments but ranks hsa00190 11th to 15th. Pathway hsa04020 is consistently significant for LEAN across runs and ranks from 12th to 18th. Moreover, hsa04728 is significant for LEAN in 2 out of 3 experiments and ranks 79th to 88th. LEAN is also the only method to find this pathway significant. Finally, hsa03050 appears ranked as 19th for GS in Experiment 1 (Table 10), ranked as 59th for LEAN in Experiment 2 (Table 11), but not significant for any tested method in Experiment 3 (Tables 9, 10, 11).

Alzheimer’s disease

For dataset GSE5281_VCX [29], a total of 11 KEGG pathways that are related to Alzheimer’s disease are included in this analysis. All tested methods consistently find the “Alzheimer’s disease” KEGG pathway (KEGG ID hsa05010) as statistically significant. With a similar structure as Tables 4, 5 shows the summarized results of ten experiments on this dataset. Although our method achieves the second highest mean rank of the main disease pathway, it scores a rank standard deviation of zero which corresponds to the highest stability. It also results in the lowest mean number of significant pathways across the ten experiments. We note that GS achieves the highest mean proportion of significant pathways that are related to the disease but with the highest mean number of significant pathways. The Alzheimer’s disease graph in Fig. 4 corresponds to areas under the curve (AUC) of 0.7150, 0.8291, and 0.6309 for the proposed method, GS, and LEAN, respectively.

Tables 9, 10, 11, 12 show that GS generally exhibits slightly higher total number of significant pathways than the proposed method. Out of the total significant pathways, the proportion that are included in the list of related pathways for Alzheimer’s disease are 6/78, 5/65, and 6/73 for the proposed method, 9/86, 9/92, 9/75 for GS, 6/63, 6/73, and

6/71 for LEAN, and 5/26 for *limma*. GS achieves the highest proportion of significant pathways out of the total related pathways, whereas *limma* achieves the highest proportion of significant pathways out of the total significant pathways for this dataset. The proposed method finds hsa05010, hsa05022, hsa00190, hsa03050, and hsa04140 as significant consistently across the experiments. The proposed method consistently ranks hsa05010, hsa05022, and hsa00190 in the top 10 whereas hsa03050 and hsa04140 have a rank of 14th-15th and 53rd-59th respectively. The remaining related pathway found significant by this method is hsa04020 but only in 2 out of 3 experiments and ranks as 66th in Experiment 1 and 46th in Experiment 3 (Table 9). Nevertheless, it is the only method to find hsa04020 as significant for this dataset. GS consistently finds hsa05010, hsa05022, hsa00190, hsa03050, hsa04141, hsa04210, hsa04140, hsa04910, and hsa04933 as significant across the experiments. Similarly to the proposed method, hsa05010, hsa05022, and hsa00190 are the only pathways that consistently rank in the top 10 for GS. hsa03050 achieves a rank ranging from 18th to 23rd for GS. The ranks for the remaining pathways range from 47th to 58th for hsa04141, from 15th to 22nd for hsa04210, from 40th to 58th for hsa04140, from 41st to 70th for hsa04910, and from 61st to 72nd for hsa04933. GS is the only tested method that finds hsa04210, hsa4910, and hsa04933 as significant. LEAN consistently finds hsa05010, hsa05022, hsa00190, hsa03050, hsa04141, and hsa04140 as significant across the experiments. Similarly to GS and the proposed method, hsa05010, hsa05022, and hsa00190 are pathways that consistently rank in the top 10 for LEAN. Nevertheless, LEAN also consistently ranks hsa03050 in the top 10. The ranks for the remaining pathways range from 43rd to 57th for hsa04141 and from 26th to 68th for hsa04140. *limma* ranks hsa05010, hsa05022, hsa00190, and hsa03050 in the top 10 and hsa04141 as 14th.

Concordance was also reported between different subsets of GSE5281 for each of the tested methods (Table 7). The concordance between GSE5281_VCX and GSE5281_HIP is highest for LEAN (rho = 0.47). Between GSE5281_VCX and GSE5281_EC, the concordance is highest for GS (rho = 0.45). Finally, between GSE5281_HIP and GSE5281_EC, the concordance is highest for the proposed method. The proposed method outperforms GS, LEAN, and *limma* for two out of three dataset pairs each. LEAN outperforms GS with two dataset pairs each while *limma* outperforms GS and LEAN with two dataset pairs each.

Huntington's disease

For dataset GSE8762 [30], a total of 13 KEGG pathways that are related to Huntington's disease are included in this analysis. All tested methods except *limma* consistently find

Table 7 Spearman rank correlations for gene-level *p* values generated on pairs of subsets of the Alzheimer's dataset [29]

Method	VCX - HIP	VCX - EC	EC - HIP
Proposed method	0.31073225	0.18717926	0.24277068
GeneSurrounder	0.26129886	0.45191229	0.14672705
LEAN	0.47151218	0.11027717	0.16508144
limma	0.3576224	0.16245708	0.21930437

the "Huntington's disease" KEGG pathway (KEGG ID hsa05016) as statistically significant. Table 6 shows the summarized results of ten experiments on this dataset. Our method scores very close to the highest mean proportion of significant pathways related to the disease. GS scores the lowest mean number of significant pathways across the ten experiments, while our approach and LEAN scores similar values. The Huntington's disease graph in Fig. 4 corresponds to areas under the curve (AUC) of 0.6956, 0.5526, and 0.66061 for the proposed method, GS, and LEAN, respectively.

Tables 9, 10, 11, 12 show that GS generally has the least total number of significant pathways across experiments excluding *limma*. Out of the total significant pathways, the proportion that are included in the list of related pathways for Huntington's disease are 8/112, 8/92, and 8/113 for the proposed method, 4/44, 4/49, and 4/43 for GS, and 8/98, 8/106, and 8/103 for LEAN. The proposed method and LEAN achieve the highest proportion of significant pathways out of the total related pathways, whereas all tested methods excluding *limma* achieve a comparable proportion of significant pathways out of the total significant pathways for this dataset. The proposed method finds hsa05016, hsa05022, hsa00190, hsa03050, hsa04210, hsa04140, hsa04115, and hsa04144 as significant consistently across the experiments. The proposed method consistently ranks hsa05016, hsa05022, and hsa00190 in the top 10 whereas hsa03050 ranks 18th-20th. For the remaining pathways, the proposed method ranks hsa04210 from 46th to 111th, hsa04140 from 26th to 54th, hsa04115 from 54th to 61st, and hsa04144 from 81st to 92nd. The proposed method is the only one to find has04144 as significant. GS consistently finds hsa05016, hsa05022, hsa00190, and hsa03050 as significant across the experiments. GS also consistently ranks hsa05016, hsa05022, and hsa00190 in the top 10 whereas hsa03050 has a rank ranging from 10th to 12th. LEAN consistently finds hsa05016, hsa05022, hsa00190, hsa03050, hsa04210, hsa04140, hsa04115, and hsa04141 as significant across the experiments. Similarly to the proposed method and GS, LEAN consistently ranks hsa05016 and hsa00190 in the top 10. Nevertheless, its rank for hsa05022 ranges from 3rd to 12th. For the remaining pathways, LEAN ranks hsa03050 from 21st to 26th, hsa04141 from 34th to 50th, hsa04210 from 10th to 17th, hsa04140 from 12th to 41st, and hsa04115 from 47th to 55th. LEAN is the only method to find has04141 as significant. No method finds hsa04020, hsa04724, and hsa03022 as significant in any of the experiments. Pathway hsa03020 has no genes that pass the preprocessing for this dataset and therefore yields no results.

p53 mutation

We further assess the tested methods on the p53 dataset to ensure the method performs comparably or better than other tested methods on a dataset unrelated to neurodegenerative diseases. Table 8 summarizes the enrichment analysis results on the p53 dataset in terms of the IDs, names, ranks, and *p* values of the KEGG pathways that are identified as significant by each of the compared methods. These results show that the proposed method ranks all 4 relevant pathways, hsa04115, hsa04110, hsa04210, and hsa04218, in the top 10. GS only ranks hsa04115 and hsa04110 in the top 10 but ranks hsa04210 as 86th and hsa04218 as 26th. LEAN ranks hsa04115, similarly to GS and the proposed method, as well as hsa04210 in the top 10. LEAN also ranks hsa04110 as 23rd and

Table 8 Enrichment results for the p53 mutation dataset in a single experiment

		Proposed method	**GeneSurrounder**	**LEAN**	***limma***
Total # of significant pathways		**60**	**110**	**83**	**46**
KEGG Pathway ID	**KEGG Pathway Name**	**Ranking (*p* value)**			
hsa04115	p53 signaling pathway	2 (1.86e−09)	4 (9.32e−09)	3 (9.31e−14)	5 (3.49e−10)
hsa04110	Cell cycle	1 (5.47e−19)	1 (1.39e−20)	23 (1.99e−05)	NA (0.61)
hsa04210	Apoptosis	5 (1.73e−05)	86 (0.01)	9 (3.09e−10)	28 (6.54e−06)
hsa04218	Cellular senescence	7 (3.30e−05)	26 (1.06e−04)	15 (5.86e−08)	NA (0.61)

Note: seed = 3 for the proposed method and LEAN, seed = 123 for GS

hsa04218 as 15th. Finally, *limma* is the only method not to find all 4 relevant pathways significant, only ranking hsa04115 as 5th and hsa04210 as 28th.

Discussion

The usage of high-throughput experimental technologies is central to the identification of molecular targets and the development of suitable treatments for various diseases. Classical interpretations of generated data, such as differential gene expression and pathway analyses, disregard interconnections when looking for gene-disease associations. Given that interconnections between studied genes are central to cellular processes, there has been a recent interest in incorporating them in such studies to allow the detection of gene modules that underlie complex phenotypes in gene interaction networks. Existing methods either impose radius-based restrictions or grow modules freely at the expense of a statistical bias towards large modules. We propose a heuristic method, inspired by Ant Colony Optimization, to apply gene-level scoring and module identification with distance-based search constraints and penalties rather than radius-based restrictions. We test and compare our results to other approaches using three different neurodegenerative diseases, namely Alzheimer's, Parkinson's, and Huntington's, over three independent experiments. It can be seen that the method maintains relative stability compared to the other methods, especially for highly-ranked pathways (top 10), across the conducted experiments. This is evidenced by the fact that the proposed method is the only method to consistently showcase the same relevant pathways in the top 10 across the experiments. Moreover, if a relevant pathway consistently ranks in the top 10 across the experiments for any of the other methods, in most cases, it will appear in the top 10 for the proposed method. Nevertheless, the converse is not true. The only exception to this is pathway hsa03050 which appears in the top 10 consistently for LEAN using GSE5281_VCX (Table 11) but not the proposed method (Table 9). In addition, aside from *limma*, the proposed method exhibits the least or ties for the least total number of significant pathways in 2 out of 3 of the tested datasets. These results indicate that the method performs well in prioritizing genes with high relevance through their significance ranking as compared to other tested methods in these experiments.

Table 9 Enrichment results for the proposed method for dataset GSE20291, GSE5281_VCX, and GSE8762 in 3 experiments

Parkinson's disease (GSE20291)				
		Experiment 1 (seed= 1)	**Experiment 2 (seed= 2)**	**Experiment 3 (seed= 3)**
		Total # of significant pathways		
		22	**25**	**22**
KEGG Pathway ID	**KEGG Pathway Name**	**Ranking (*p* value)**		
hsa05012	Parkinson's Disease	3 (6.49e−10)	4 (1.27e−08)	3 (1.36e−10)
hsa05022	Pathways of neurodegeneration- multiple diseases	8 (5.34e−08)	6 (6.70e−07)	10 (4.03e−06)
hsa00190	Oxidative phosphorylation	2 (1.71e−11)	1 (2.29e−11)	1 (3.46e−14)
hsa03050	Proteasome	NA (0.78)	NA (0.99)	NA (0.97)
hsa04020	Calcium signaling pathway	NA (0.26)	NA (0.22)	NA (0.09)
hsa04141	Protein processing in endoplasmic reticulum	NA (0.99)	NA (0.99)	NA (0.99)
hsa04210	Apoptosis	NA (0.05)	NA (0.52)	NA (0.15)
hsa00350	Tyrosine metabolism	NA (0.99)	NA (0.99)	NA (0.99)
hsa04120	Ubiquitin mediated proteolysis	NA (0.35)	NA (0.76)	NA (0.58)
hsa04137	Mitophagy - animal	NA (0.98)	NA (0.99)	NA (0.99)
hsa04728	Dopaminergic Synapse	NA (0.80)	NA (0.53)	NA (0.99)
Alzheimer's disease (GSE5281_VCX)				
		Experiment 1 (seed= 1)	**Experiment 2 (seed= 2)**	**Experiment 3 (seed= 3)**
		Total # of significant pathways		
		78	**65**	**73**
KEGG pathway ID	**KEGG Pathway Name**	**Ranking (*p* value)**		
hsa05010	Alzheimer's disease	7 (1.86e−41)	7 (1.61e−41)	7 (4.74e−42)
hsa05022	Pathways of neurodegeneration- multiple diseases	5 (1.08e−55)	5 (1.06e−54)	4 (6.80e−57)
hsa00190	Oxidative phosphorylation	4 (4.68e−56)	4 (2.64e−55)	6 (2.35e−51)
hsa03050	Proteasome	14 (9.21e−13)	14 (3.64e−14)	15 (1.18e−13)
hsa04020	Calcium signaling pathway	66 (0.02)	NA (0.11)	46 (3.75e−03)
hsa04141	Protein processing in endoplasmic reticulum	NA (0.96)	NA (1.00)	NA (0.96)
hsa04210	Apoptosis	NA (0.18)	NA (0.92)	NA (0.25)
hsa04140	Autophagy - animal	59 (0.01)	55 (0.03)	53 (9.37e−03)
hsa04310	Wnt signaling pathway	NA (0.27)	NA (0.87)	NA (0.68)
hsa04910	Insulin signaling pathway	NA (0.63)	NA (1.00)	NA (0.67)

Table 9 (continued)

Alzheimer's disease (GSE5281_VCX)

		Experiment 1 (seed= 1)	**Experiment 2 (seed= 2)**	**Experiment 3 (seed= 3)**
		Total # of significant pathways		
		78	**65**	**73**
KEGG pathway ID	**KEGG Pathway Name**	**Ranking (*p* value)**		
hsa04933	AGE-RAGE signaling pathway in diabetic complications	NA (0.25)	NA (0.16)	NA (0.16)

Huntington's disease (GSE8762)

		Experiment 1 (seed= 1)	**Experiment 2 (seed= 2)**	**Experiment 3 (seed= 3)**
		Total # of significant pathways		
		112	**92**	**113**
KEGG pathway ID	**KEGG pathway name**	**Ranking (*p* value)**		
hsa05016	Huntington's disease	5 (1.25e−31)	5 (4.88e−29)	5 (7.00e−30)
hsa05022	Pathways of neurodegeneration- multiple diseases	7 (5.57e−24)	6 (8.94e−24)	7 (1.05e−23)
hsa00190	Oxidative phosphorylation	1 (1.91e−37)	1 (7.64e−34)	1 (5.81e−36)
hsa03050	Proteasome	18 (1.40e−06)	20 (1.60e−05)	20 (8.12e−06)
hsa04020	Calcium signaling pathway	NA (0.73)	NA (0.51)	NA (0.76)
hsa04141	Protein processing in endoplasmic reticulum	NA (0.71)	NA (0.31)	NA (0.43)
hsa04210	Apoptosis	111 (0.05)	46 (3.43e−03)	52 (2.56e−03)
hsa04140	Autophagy - animal	54 (2.14e−03)	26 (2.29e−04)	31 (2.80e−04)
hsa04115	p53 signaling pathway	57 (2.83e−03)	54 (9.84e−03)	61 (6.37e−03)
hsa04144	Endocytosis	84 (0.02)	81 (0.03)	92 (0.02)
hsa04724	Glutamatergic Synapse	NA (0.83)	NA (0.77)	NA (0.49)
hsa03022	Basal transcription factors	NA (1.00)	NA (0.38)	NA (0.65)
hsa03020	RNA polymerase	NA	NA	NA

The proposed method also shows high stability in which disease-relevant pathways it indicates as significant. The only inconsistency by the proposed method in this regard was for hsa04020 in Experiment 2 using GSE5281_VCX (Table 9). Nevertheless, the proposed method was the only one to find this pathway as significant for this dataset. The method's stability is also supported by the concordance analysis results which show that the proposed method is the only one to outperform all other tested methods in 2 pairs of GSE5281 subsets. All other tested methods outperform another method or more, but not all, for 2 or fewer pairs.

Table 10 Enrichment results for GeneSurrounder for dataset GSE20291, GSE5281_VCX, and GSE8762 in 3 experiments

Parkinson's disease (GSE20291)				
		Experiment 1 (seed= 121)	**Experiment 2 (seed= 122)**	**Experiment 3 (seed= 123)**
		Total # of significant pathways		
		32	**51**	**39**
KEGG Pathway ID	**KEGG Pathway Name**	**Ranking (*p* value)**		
hsa05012	Parkinson's Disease	4 (1.09e−11)	6 (2.26e−07)	9 (3.45e−07)
hsa05022	Pathways of neurodegeneration- multiple diseases	11 (4.19e−07)	28 (3.39e−03)	14 (3.02e−04)
hsa00190	Oxidative phosphorylation	3 (4.40e−12)	4 (8.43e−10)	2 (1.34e−14)
hsa03050	Proteasome	19 (6.43e−03)	NA (0.47)	NA (0.91)
hsa04020	Calcium signaling pathway	NA (0.15)	27 (2.21e−03)	25 (8.33e−03)
hsa04141	Protein processing in endoplasmic reticulum	NA (1.00)	NA (0.99)	NA (1.00)
hsa04210	Apoptosis	NA (1.00)	NA (0.99)	NA (1.00)
hsa00350	Tyrosine metabolism	NA (1.00)	NA (0.99)	NA (1.00)
hsa04120	Ubiquitin mediated proteolysis	NA (1.00)	NA (0.99)	NA (1.00)
hsa04137	Mitophagy - animal	NA (1.00)	NA (0.99)	NA (1.00)
hsa04728	Dopaminergic Synapse	NA (1.00)	NA (0.99)	NA (1.00)
Alzheimer's Disease (GSE5281_VCX)				
		Experiment 1 (seed= 121)	**Experiment 2 (seed= 122)**	**Experiment 3 (seed= 123)**
		Total # of significant pathways		
		86	**92**	**75**
KEGG Pathway ID	**KEGG Pathway Name**	**Ranking (*p* value)**		
hsa05010	Alzheimer's Disease	7 (3.33e−50)	5 (3.73e−59)	7 (8.28e−58)
hsa05022	Pathways of neurodegeneration- multiple diseases	6 (2.54e−56)	3 (2.43e−68)	4 (1.13e−63)
hsa00190	Oxidative phosphorylation	4 (1.13e−60)	7 (2.16e−55)	5 (4.02e−60)
hsa03050	Proteasome	18 (1.05e−11)	23 (8.19e−11)	21 (4.30e−12)
hsa04020	Calcium signaling pathway	NA (0.54)	NA (0.42)	NA (0.58)
hsa04141	Protein processing in endoplasmic reticulum	47 (1.59e−03)	51 (3.31e−04)	58 (7.39e−03)
hsa04210	Apoptosis	22 (7.43e−11)	17 (5.60e−14)	15 (1.37e−17)
hsa04140	Autophagy - animal	58 (4.28e−03)	42 (8.55e−05)	40 (5.75e−04)
hsa04310	Wnt signaling pathway	NA (0.93)	NA (0.31)	NA (0.32)
hsa04910	Insulin signaling pathway	50 (1.61e−03)	41 (4.23e−05)	70 (0.03)
hsa04933	AGE-RAGE signaling pathway in diabetic complications	72 (0.02)	69 (8.02e-03)	61 (0.01)

Table 10 (continued)

Huntington's Disease (GSE8762)

		Experiment 1 (seed= 121)	**Experiment 2 (seed= 122)**	**Experiment 3 (seed= 123)**
		Total # of significant pathways		
		44	**49**	**43**
KEGG pathway ID	**KEGG pathway name**	**Ranking (*p* value)**		
hsa05016	Huntington's Disease	1 (1.27e−65)	3 (1.93e−68)	2 (2.10e−66)
hsa05022	Pathways of neurodegeneration- multiple diseases	6 (5.01e−37)	6 (9.96e−42)	6 (8.44e−38)
hsa00190	Oxidative phosphorylation	5 (3.89e−51)	5 (8.14e−52)	5 (1.43e−50)
hsa03050	Proteasome	10 (1.92e−27)	10 (2.18e−27)	12 (1.38e−23)
hsa04020	Calcium signaling pathway	NA (0.99)	NA (0.99)	NA (0.99)
hsa04141	Protein processing in endoplasmic reticulum	NA (0.99)	NA (0.99)	NA (0.99)
hsa04210	Apoptosis	NA (0.33)	NA (0.68)	NA (0.72)
hsa04140	Autophagy - animal	NA (0.99)	NA (0.99)	NA (0.99)
hsa04115	p53 signaling pathway	NA (0.99)	NA (0.99)	NA (0.99)
hsa04144	Endocytosis	NA (0.99)	NA (0.98)	NA (0.99)
hsa04724	Glutamatergic Synapse	NA (0.99)	NA (0.99)	NA (0.99)
hsa03022	Basal transcription factors	NA (0.99)	NA (0.99)	NA (0.99)
hsa03020	RNA polymerase	NA	NA	NA

Conclusion

We propose an approach for gene-level subnetwork identification that produces noteworthy enrichment results for three gene expression datasets relating to neurodegenerative diseases and having different sample sizes and case to control proportions. The presented method shows superior stability in comparison to other approaches, namely GeneSurrounder [22], LEAN [3], and *limma* [1]. It also detects significantly dysregulated and disease-relevant gene modules when tested on the p53 mutation dataset. Our approach allows the detection of crucial mechanistic multi-gene connections that underlie complex diseases.

In terms of future work, we identify three main levels of improvements. First, we plan to consider other types and resources for gene expression as well as network interaction datasets. Such data could also target different types of diseases. In addition, we aim to work on further reducing the computational time required to perform the overall analysis. This currently is $O(g^2 + n * A * g)$ where g, n, and a are the number of genes, iterations, and ants, respectively. A better runtime could be achieved through the exploration of other types of optimization techniques such as GA and SA, as done in *pathfindR* [18]. Finally, it is important to note that although most tested methods rank the main KEGG pathway corresponding to a neurodegenerative disease in the top ten significant pathways, most enrichment analyses also highly rank pathways related to other neurodegenerative diseases. This indicates the need for future research to improve the sensitivity of these approaches to properly distinguish between related diseases while maximizing the rank of pathways related to the studied disease.

Table 11 Enrichment results for LEAN for dataset GSE20291, GSE5281_VCX, and GSE8762 in 3 experiments

Parkinson's disease (GSE20291)				
		Experiment 1 (seed= 1)	**Experiment 2 (seed= 2)**	**Experiment 3 (seed= 3)**
		Total # of significant pathways		
		88	**98**	**102**
KEGG pathway ID	**KEGG pathway name**	**Ranking (*p* value)**		
hsa05012	Parkinson's Disease	4 (4.79e−14)	2 (1.19e−15)	4 (2.36e−15)
hsa05022	Pathways of neurodegeneration- multiple diseases	8 (5.47e−12)	8 (5.45e−13)	8 (6.41e−13)
hsa00190	Oxidative phosphorylation	11 (2.17e−11)	15 (3.88e−10)	11 (5.46e−12)
hsa03050	Proteasome	NA (0.06)	59 (7.39e−04)	NA (0.19)
hsa04020	Calcium signaling pathway	12 (1.72e−10)	16 (7.60e−10)	18 (6.07e−10)
hsa04141	Protein processing in endoplasmic reticulum	NA (0.53)	NA (0.92)	NA (0.53)
hsa04210	Apoptosis	NA (1.00)	NA (1.00)	NA (1.00)
hsa00350	Tyrosine metabolism	NA (1.00)	NA (1.00)	NA (1.00)
hsa04120	Ubiquitin mediated proteolysis	NA (0.69)	NA (0.44)	NA (0.70)
hsa04137	Mitophagy - animal	NA (0.26)	NA (0.21)	NA (0.15)
hsa04728	Dopaminergic Synapse	NA (0.08)	88 (0.03)	79 (6.45e−03)
Alzheimer's disease (GSE5281_VCX)				
		Experiment 1 (seed= 1)	**Experiment 2 (seed= 2)**	**Experiment 3 (seed= 3)**
		Total # of significant pathways		
		63	**73**	**71**
KEGG pathway ID	**KEGG pathway name**	**Ranking (*p* value)**		
hsa05010	Alzheimer's Disease	10 (1.78e−15)	8 (8.78e−22)	8 (1.57e−22)
hsa05022	Pathways of neurodegeneration- multiple diseases	7 (8.25e−19)	5 (2.87e−32)	5 (3.24e−33)
hsa00190	Oxidative phosphorylation	8 (2.23e−17)	6 (6.17e−32)	6 (2.76e−32)
hsa03050	Proteasome	5 (4.77e−23)	10 (1.78e−12)	9 (1.05e−15)
hsa04020	Calcium signaling pathway	NA (1.00)	NA (1.00)	NA (1.00)
hsa04141	Protein processing in endoplasmic reticulum	57 (0.04)	44 (3.42e−03)	43 (1.34e−03)
hsa04210	Apoptosis	NA (1.00)	NA (1.00)	NA (1.00)
hsa04140	Autophagy - animal	26 (4.51e−05)	62 (0.03)	68 (0.04)
hsa04310	Wnt signaling pathway	NA (1.00)	NA (1.00)	NA (1.00)
hsa04910	Insulin signaling pathway	NA (1.00)	NA (1.00)	NA (1.00)
hsa04933	AGE-RAGE signaling pathway in diabetic complications	NA (1.00)	NA (1.00)	NA (1.00)

Table 11 (continued)

Huntington's disease (GSE8762)				
		Experiment 1 (seed = 1)	**Experiment 2 (seed = 2)**	**Experiment 3 (seed = 3)**
		Total # of significant pathways		
		44	**49**	**43**
KEGG Pathway ID	**KEGG Pathway Name**	**Ranking (*p* value)**		
hsa05016	Huntington's Disease	1 (1.99e−16)	1 (3.22e−26)	5 (3.22e−28)
hsa05022	Pathways of neurodegeneration- multiple diseases	3 (1.87e−14)	5 (1.03e−16)	12 (6.39e−19)
hsa00190	Oxidative phosphorylation	4 (1.83e−13)	3 (2.74e−20)	1 (7.24e−44)
hsa03050	Proteasome	26 (6.56e−05)	25 (1.02e−06)	21 (2.27e−11)
hsa04020	Calcium signaling pathway	NA (1.00)	NA (1.00)	NA (1.00)
hsa04141	Protein processing in endoplasmic reticulum	34 (2.47e−04)	50 (6.82e−04)	35 (3.84e−07)
hsa04210	Apoptosis	13 (4.43e−08)	10 (1.06e−11)	17 (5.69e−13)
hsa04140	Autophagy - animal	12 (4.30e−08)	17 (2.21e−08)	41 (7.89e−06)
hsa04115	p53 signaling pathway	55 (2.84e−03)	47 (5.21e−04)	47 (6.32e−05)
hsa04144	Endocytosis	NA (1.00)	NA (1.00)	NA (1.00)
hsa04724	Glutamatergic Synapse	NA (1.00)	NA (1.00)	NA (1.00)
hsa03022	Basal transcription factors	NA (1.00)	NA (1.00)	NA (1.00)
hsa03020	RNA polymerase	NA	NA	NA

Table 12 Enrichment results for *limma* for dataset GSE20291, GSE5281_VCX, and GSE8762

Parkinson's disease (GSE20291)

		Total # of significant pathways
		0
KEGG pathway ID	**KEGG pathway name**	**Ranking (*p* value)**
hsa05012	Parkinson's Disease	NA (1.00)
hsa05022	Pathways of neurodegeneration- multiple diseases	NA (1.00)
hsa00190	Oxidative phosphorylation	NA (1.00)
hsa03050	Proteasome	NA (1.00)
hsa04020	Calcium signaling pathway	NA (1.00)
hsa04141	Protein processing in endoplasmic reticulum	NA (1.00)
hsa04210	Apoptosis	NA (1.00)
hsa00350	Tyrosine metabolism	NA (1.00)
hsa04120	Ubiquitin mediated proteolysis	NA (1.00)
hsa04137	Mitophagy - animal	NA (1.00)
hsa04728	Dopaminergic Synapse	NA (1.00)

Alzheimer's disease (GSE5281_VCX)

		Total # of significant pathways
		26
KEGG pathway ID	**KEGG pathway name**	**Ranking (*p* value)**
hsa05010	Alzheimer's Disease	7 (5.57e—15)
hsa05022	Pathways of neurodegeneration- multiple diseases	6 (1.12e—15)
hsa00190	Oxidative phosphorylation	8 (2.81e—12)
hsa03050	Proteasome	5 (5.59e—16)
hsa04020	Calcium signaling pathway	NA (0.99)
hsa04141	Protein processing in endoplasmic reticulum	14 (5.77e–04)
hsa04210	Apoptosis	NA (0.99)
hsa04140	Autophagy-animal	NA (0.16)
hsa04310	Wnt signaling pathway	NA (0.99)
hsa04910	Insulin signaling pathway	NA (0.99)
hsa04933	AGE-RAGE signaling pathway in diabetic complications	NA (0.99)

Huntington's disease (GSE8762)

		Total # of significant pathways
		0
KEGG pathway ID	**KEGG pathway name**	**Ranking (*p* value)**
hsa05016	Huntington's Disease	NA (1.00)
hsa05022	Pathways of neurodegeneration- multiple diseases	NA (1.00)
hsa00190	Oxidative phosphorylation	NA (1.00)
hsa03050	Proteasome	NA (1.00)
hsa04020	Calcium signaling pathway	NA (1.00)
hsa04141	Protein processing in endoplasmic reticulum	NA (1.00)
hsa04210	Apoptosis	NA (1.00)
hsa04140	Autophagy - animal	NA (1.00)

Table 12 (continued)

Huntington's disease (GSE8762)

		Total # of significant pathways
		0
KEGG pathway ID	**KEGG pathway name**	**Ranking (*p* value)**
hsa04115	p53 signaling pathway	NA (1.00)
hsa04144	Endocytosis	NA (1.00)
hsa04724	Glutamatergic Synapse	NA (1.00)
hsa03022	Basal transcription factors	NA (1.00)
hsa03020	RNA polymerase	NA (1.00)

Acknowledgements

The authors would like to thank Yazid Hoblos for generating the ROC plots.

Authors' contributions

GEH and EMH conceived and designed the study. GEH implemented the approach and interpreted the experimental findings under the supervision of EMH. DA guided the design and implementation of the optimization part. GEH wrote the manuscript draft. EMH and DA reviewed and edited the draft. All authors read and approved the final manuscript.

Funding

Not Applicable

Availability of data and materials

The preprocessed versions of the Parkinson's (GSE20291) [28], Alzheimer's (GSE5281) [29], and Huntington's (GSE8762) [30] disease datasets that support the findings of this study are openly available via the R package ***GSEABenchmarkeR***. Similarly, the preprocessed version of the p53 mutation dataset from the NCI-60 cell lines, originally available through the GSEA Broad Institute website [4, 36], is openly available via the R package ***GSAR*** [37].

Code availability

All code necessary to replicate the findings of this study is available at github.com/GhadiElHasbani/ACOxGS.git.

Declarations

Ethics approval and consent to participate

Not Applicable

Consent for publication

Not Applicable

Competing interests

The authors declare that they have no Conflict of interest.

Received: 18 October 2023 Accepted: 12 July 2024

Published online: 01 August 2024

References

1. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47–e47. https://doi.org/10.1093/nar/gkv007.
2. Manoli T, Gretz N, Gröne HJ, Kenzelmann M, Eils R, Brors B. Group testing for pathway analysis improves comparability of different microarray datasets. Bioinformatics. 2006;22(20):2500–6. https://doi.org/10.1093/bioinformatics/btl424.
3. Gwinner F, Boulday G, Vandiedonck C, Arnould M, Cardoso C, Nikolayeva I, et al. Network-based analysis of omics data: the LEAN method. Bioinformatics. 2016;33(5):701–9. https://doi.org/10.1093/bioinformatics/btw676.
4. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102.
5. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS, et al. A novel signaling pathway impact analysis. Bioinformatics. 2008;25(1):75–82. https://doi.org/10.1093/bioinformatics/btn577.

6. Gu Z, Liu J, Cao K, Zhang J, Wang J. Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. BMC Syst Biol. 2012;6(1):56.
7. Gu Z, Wang J. CePa: an R package for finding significant pathways weighted by multiple network centralities. Bioinformatics. 2013;29(5):658–60. https://doi.org/10.1093/bioinformatics/btt008.
8. Ideker T, Ozier O, Schwikowski B, Siegel AF. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics. 2002;18(suppl1):S233–40. https://doi.org/10.1093/bioinformatics/18.suppl_1.S233.
9. Vandin F, Upfal E, Raphael BJ. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol. 2011;18(3):507–22.
10. Ma H, Schadt EE, Kaplan LM, Zhao H. COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. Bioinformatics. 2011;27(9):1290–8.
11. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2): e1002375.
12. Braun R, Shah S. Network methods for pathway analysis of genomic data. arXiv preprint. 2014 Nov; arXiv:1411.1993. [q-bio.QM].
13. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, et al. Gene prioritization through genomic data fusion. Nat Biotechnol. 2006;24(5):537–44.
14. Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82(4):949–58.
15. Nitsch D, Tranchevent LC, Thienpont B, Thorrez L, Van Esch H, Devriendt K, et al. Network analysis of differential expression for the identification of disease-causing genes. PLoS ONE. 2009;4(5): e5526.
16. Ruan P, Wang S. DiSNEP: a disease-specific gene network enhancement to improve Prioritizing candidate disease genes. Briefings Bioinf. 2020;22(4):bbaa241. https://doi.org/10.1093/bib/bbaa241.
17. Shafi A, Donato M, Draghici S. a systems biology approach for the identification of significantly perturbed genes. In: Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics. BCB '15. New York, NY, USA: Association for Computing Machinery; 2015. p. 423–432. Available from: https://doi.org/10.1145/2808719.2808763.
18. Ulgen E, Ozisik O, Sezerman OU. pathfindR: an R package for comprehensive identification of enriched pathways in omics data through active subnetworks. Front Genet. 2019;10:858. https://doi.org/10.3389/fgene.2019.00858.
19. Nikolayeva I, Guitart Pla O, Schwikowski B. Network module identification-A widespread theoretical bias and best practices. Methods. 2018;132:19–25. https://doi.org/10.1016/j.ymeth.2017.08.008.
20. Xu Z, Marchionni L, Wang S. MultiNEP: a multi-omics network enhancement framework for prioritizing disease genes and metabolites simultaneously. Bioinformatics. 2023;39(6):btad333.
21. Zhang Y, Shi W, Sun Y. A functional gene module identification algorithm in gene expression data based on genetic algorithm and gene ontology. BMC Genom. 2023;24(1):76.
22. Shah SD, Braun R. GeneSurrounder: network-based identification of disease genes in expression data. BMC Bioinf. 2019;20(1):229.
23. Dorigo M, Birattari M, Stutzle T. Ant colony optimization. IEEE Comput Intell Mag. 2006;1(4):28–39. https://doi.org/10.1109/MCI.2006.329691.
24. Xu W, Chang C, Hung YS, Kwan SK, Fung PCW. Order statistics correlation coefficient as a novel association measurement with applications to biosignal analysis. IEEE Trans Signal Process. 2007;55(12):5552–63. https://doi.org/10.1109/TSP.2007.899374.
25. R Core Team. R: A language and environment for statistical computing. Vienna, Austria. Available from: https://www.R-project.org/.
26. Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, et al. Toward a gold standard for benchmarking gene set enrichment analysis. Brief Bioinf. 2020;22(1):545–56. https://doi.org/10.1093/bib/bbz158.
27. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80.
28. Zhang Y, James M, Middleton FA, Davis RL. Transcriptional analysis of multiple brain regions in Parkinson's disease supports the involvement of specific protein processing, energy metabolism, and signaling pathways, and suggests novel disease mechanisms. Am J Med Genet B Neuropsychiatr Genet. 2005;137B(1):5–16.
29. Liang W, Dunckley T, Beach T, Grover A, Mastroeni D, Walker D, et al. Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. Physiol Genom. 2007;28(3):311–22. https://doi.org/10.1152/physiolgenomics.00208.2006.
30. Runne H, Kuhn A, Wild EJ, Pratyaksha W, Kristiansen M, Isaacs JD, et al. Analysis of potential transcriptomic biomarkers for Huntington's disease in peripheral blood. Proc Natl Acad Sci U S A. 2007;104(36):14424–9.
31. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England). 2003;4(2):249–64. https://doi.org/10.1093/biostatistics/4.2.249.
32. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy-analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307–15. https://doi.org/10.1093/bioinformatics/btg405.
33. Kanehisa M. The KEGG database. Novartis Found Symp. 2002;247:91–101; discussion 101–3, 119–28, 244–52.
34. Tarca AL, Draghici S, Bhatti G, Romero R. Down-weighting overlapping genes improves gene set analysis. BMC Bioinf. 2012;13(1):136.
35. Tarca AL, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLOS ONE. 2013;8(11):1–10. https://doi.org/10.1371/journal.pone.0079217.
36. Olivier M, Eeles R, Hollstein M, Khan MA, Harris CC, Hainaut P. The IARC TP53 database: new online mutation analysis and recommendations to users. Human Mutat. 2002;19(6):607–14. https://doi.org/10.1002/humu.10081.

37. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene sets net correlations analysis (GSNCA): a multivariate differential coexpression test for gene sets. Bioinformatics. 2013;30(3):360–8.
38. Tenenbaum D, Maintainer BP.: KEGGREST: client-side REST access to the Kyoto encyclopedia of genes and genomes (KEGG). R package version 1.40.0.
39. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. In: Varoquaux G, Vaught T, Millman J, editors. Proceedings of the 7th Python in Science Conference. Pasadena, CA USA; 2008. p. 11-15.

Publisher's Note