**Computation Biology**

**Ghadi El Hasbani**

**Bioinformatics BS & Psychology Minor, Lebanese American University.** [**ghadi.elhasbani@lau.edu**](mailto:ghadi.elhasbani@lau.edu)

**Bilal W. Hamdanieh**

**Bioinformatics BS, Lebanese American University.** [**bilal.hamdanieh@lau.edu**](mailto:lara.issa02@lau.edu)

**Frederick Abi Chahine**

**Bioinformatics BS, Lebanese American University.** [**frederick.abichahine@lau.edu**](mailto:frederick.abichahine@lau.edu)

**ABSTRACT**

**INTRODUCTION**

1. **The Effect of Genes Over their Surrounding Area**

Differential Gene Expression Analysis (DEA) is the most widely used method for detecting significant gene-associations based on their mean expression 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 leading to limits in mechanistic insights. Hence, the analysis might miss crucial multi-gene interactions that underlie complex phenotypes. As a result. DEA can exhibit poor consensus with studies of the same conditions [2, 3]. This necessitates the emergence of new methods that take into consideration the network structure of genes.

Recently proposed methods use networks to measure the effect of genes over their surrounding area of other genes, identifying modules: a group of dysregulated genes that contribute to the disease/phenotype of study. One method, Local Enrichment Analysis (LEAN) identifies dysregulated subnetworks from genome-wide omics datasets by substituting common subnetworks with local subnetwork models that consist of only the direct neighboring genes (radiuslocal\_subnetwork = 1). The method is also parameter free and exhaustive over all the genes. Another method, PathFindR, extends LEAN by taking advantage of user input to specify the radius of the local subnetwork to be enriched using three different algorithms of choice: Greedy Algorithm (GD), Simulated Annealing, (SA) and Genetic Algorithm (GA). Further, it has been shown that the Greedy Algorithm performs better than the SA & GA; SA and GA are heuristic methods that do not make biologically-relevant assumptions on the active subnetwork model, which lets insignificant genes between two clusters of significant genes become a single connected active subnetwork, which results in high scoring active subnetworks with the remaining subnetworks becoming fewer and less informative. In short, there is a tendency towards large subnetworks which is attributed to a statistical bias prevalent in many tools [4]. Another recently developed algorithm is Genesurrounder that proposes an exhaustive method to consider the decay of DE and the sphere of influence of a gene. The optimal radius that identifies the effect of the gene on the neighbors is given from the combination of the two p-values using the Fisher method (pfisher from pdecay and psphere). However, the time complexity and implementation of GS requires further development and optimization to be favorable for common use.

Other research proposed network diffusion (ND) – also referred to as network propagation – for the development of integrative methods to analyze multiple gene-centered datasets while considering known or inferred relations between genes [5, 6]. ND can be of several types: Random Walk, Random Walk with Restart (RWR), Insulated Heat Diffusion (IHD) and Diffusion Kernel. The latter two methods are characterized by differences in the normalization of the adjacency matrix which implies dissimilar behaviors of information flow, mainly in relation to network hubs: at infinite time in the RWR hubs tend to naturally gather relatively more information than IHD which is characterized by an intrinsic hub penalization. Therefore, despite RWR and IHD are conceptually similar, they may present sensibly different results, especially when applied to complex biological networks. For this reason, RWR is more common in biological analysis. One of the previous methods, DMFIND [7], uses RWR network diffusion to define network proximity and uses a smoothing index quantity that allows to jointly quantify the amount of omics information in genes and in their neighborhood (not only direct neighbors). Another method, which is of main interest, is mND that similarly to DMFIND uses RWR network diffusion but uses a different scoring method (compared to the smoothing index) which takes into consideration the top k direct neighbors of a gene. This method classifies genes as modules (M), Linkers (L), Isolated (I), and Not-Selected (NS) and takes into consideration the global topology of the network while maintaining modules to contain only a gene and the top k direct neighbors (radius = 1).

Since the previous methods either consider the direct neighbors, global network topology or the decay of the gene expression, but not all together, our proposed method suggested a new protocol for considering network topology, and indirect-neighbors-based gene expression decay while still maintaining modules to contain only a gene and its direct neighbors (radius = 1).

**PROTOCOL GS & mND**

1. **Materials and Methods**
2. **Data Collection**

The data used was extracted from the pipeline of mND which was originally sourced from TCGA [8] using R package TCGAbiolinks [9]. It consists of two layers: somatic mutations frequency and differential expression scores from 11,796 genes expression data from matched tumor-normal samples (blood samples for SM and solid tissue samples for GE) for breast cancer patients and considering the human genome version 38 (hg38). The same adjacency network was also used from mND, which was originally obtained from STRING [10], to maintain comparable gene-gene interactions.

The reasoning behind choosing the same datasets used by mND is to allow us to assess and compare the results obtained with previous methods to validate any changes. This ensures the protocol for preprocessing, DE score calculation, and postprocessing is constant.

1. **Methodology**

The proposed method suggests using Genesurrounder (GS) to adjust the gene expressions prior to performing mND, then examining if the suggested workflow successfully identifies genes that have potential in the network due to the expression decay effect of other genes not necessarily in direct neighboring distance of the gene (considers radius > 1).

First, GS functions and methods were downloaded from GitHub repository to allow us to measure the decay of gene expressions exhaustively (radius > 1). Only pdecay was obtained from the GS pipeline as the data/sample was not readily available in the mND package. Then, the mND package was used to perform network diffusion using RWR and scoring the genes according to the top k direct neighbors’ network diffusion scores.

1. **Pipeline Analysis**

The pipeline starts by performing Genesurrounder, with parallelized implementation to cut down the running time of the method (ending up with an average of 4.5h/1000genes), to quantify the effect of genes on all other genes. If a gene is a source of disease-associated dysregulation, we may expect its neighbors to also exhibit a dysregulation. The quantity measured is called Decay of Differential Expression which tests whether the magnitude of the differential expression of a gene is inversely related to the distance to other genes in the interaction network. The latter allows us to obtain a pdecay value which can then be used to adjust values of expression layers prior to performing multi-network diffusion. Ideally, pfisher would be obtained by also calculating the Sphere of Influence of a gene using data/sample.

Therefore, we adjust the Differential Gene Expression layer by multiplying its values by -log10(pdecay) to increase the effect of a gene that is highly expressed and has a high decay score over its neighbors, followed by multi-network diffusion using k=2 (as indicated by k optimization results; see below) and k=3. After that, we classified the genes into M, L, I, and NS based on their differential gene expression and their neighboring information collected from network diffusion. The results were then saved for comparison with mND pipeline run alone without any modifications.

**RESULTS & DISCUSSION**

1. **Classification Improvement**

The mND classification results were compared before (R1) and after (R2) applying adjustments to the differential expression scores. mND performed alone (R1) obtained 2,297 selected genes as I, L or M, while our protocol (R2) obtained fewer selected genes (2,109 genes). The confusion matrix of the classification percentage showed that the difference in the genes is mainly in the I (76% of R1), L (69% of R1), or M (88% of R1) and not in the NS genes (96% of R1) (Table1Aå). In addition, 20% of the genes previously selected as Isolated and 27% of the those previously selected as Linkers are now not being selected. This might indicate a decrease in false positive genes selected, and raises the question of whether the decay score is the reason some genes are not selected anymore, and the reason for newly selected modules to appear. The reason could be that I and L classes are topology-dependent and could change with manipulation of the M class through score adjustment. We evaluate our protocol: is classification affected by the decay scores?

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Our Protocol Classification** | | | | |
| **mND Only Protocol** | Isolated | Linkers | Modules | Not Selected |
| Isolated | 76.70 | 0.00 | 3.57 | 19.71 |
| Linkers | 0.27 | 69.73 | 2.55 | 27.4 |
| Modules | 4.69 | 1.80 | 88.44 | 5.05 |
| Not Selected | 1.70 | 1.46 | 0.084 | 96.75 |

Table 1A. Confusion Matrix Percentages Over Original mND Scores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Our Protocol Classification** | | | | |
| **mND Only Protocol** | Isolated | Linkers | Modules | Not Selected |
| Isolated | 79.91 | 0.00 | 10.51 | 1.88 |
| Linkers | 0.34 | 84.15 | 8.92 | 3.11 |
| Modules | 1.47 | 0.55 | 78.03 | 0.14 |
| Not Selected | 18.28 | 15.29 | 2.55 | 94.87 |

Table 1B. Confusion Matrix Percentages Over Adjusted mND Scores

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Our Protocol Classification** | | | | |
| **mND Only Protocol** | Isolated | Linkers | Modules | Not Selected |
| Isolated | 6.00 | 0.00 | 0.28 | 1.54 |
| Linkers | 0.03 | 6.49 | 0.24 | 2.55 |
| Modules | 0.11 | 0.04 | 2.08 | 0.12 |
| Not Selected | 1.37 | 1.18 | 0.07 | 77.91 |

Table 1C. Confusion Matrix Percentages Over All Genes

1. **Percentage Variation in Classification**

When each class is analyzed, we notice that although we have a decrease in the number of selected genes, 7.7% of genes were additionally significant with an mND(p) value threshold of 0.05. The additional significant genes were distributed amongst Isolated, Linkers and Modules; with 69 new modules and 32 discarded modules, and 544 new Linkers and Isolated genes with 319 discarded genes of the latter type. To assess the reasoning behind the appearance of these new genes or the discarding of others, we introduce a Cumulative Decay Score and perform significance tests.

1. **Cumulative Decay Score & Significance Test**

After obtaining the classification of every gene, we can now check the Cumulative Decay Score on target genes. This score is used to assess the decay in relation to 1/distance from genes and its formula is defined as: *Cumulative Decay Score = GE\_neighbor/d\_neighbor\_target.* To reason, we hypothesize that distinct genes would show a significant pattern of high scores as compared to the background distribution of discarded genes. The algorithm’s pseudocode is as follows:

ALGORITHM 1: Cumulative Decay Score

for each gene i, do

if(old\_classi == Module) # If the gene is an old module then we add its affect to neighboring (d<r) nodes

for each gene j in neighborhood, do # Traverse other genes in the surrounding radius and # update their cumulative\_decay\_score

if(distanceij =< radiusi), do

scorej = scorej - log10(pdecay i)/distanceij

return scores

end

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart

Description automatically generatedChart, histogram

Description automatically generated

Figure .B: Discarded Module Genes in Green

Figure 1.D: Discarded I & L Genes in Green

Figure 1.C: Distinct I & L Genes in Green

Figure 1.A: Distinct Module Genes in Green

The density plots of the cumulative decay score (Figure 1), shows a visual difference between the distinct genes (modules in Fig1.A, and I&L in Fig1.C). To assess this difference, we perform a significance test by comparing variances (F-test) and means (T-test) of distinct and discarded genes. For distinct modules, we find a statistically significant (p-value = 5.896e-09) difference in the means whereby the distinct genes had a significantly higher mean (168.97) than that of the background distribution (126.33) with a confidence interval [29.82, 55.46]. Also, we obtain a p-value of 0.002 and ratio of 0.55 in the F-test which means that the variance of the distinctive genes that were selected as new modules was significantly lower than the old modules for the calculated cumulative decay score with a confidence interval [0.41, 0.80]. Next, for the discarded modules, we find a statistically significant (p-value = 5.055e-06) difference in the means whereby the discarded modules had a significantly higher mean (180.41) than that of the background distribution (126.44) with a confidence interval [33.95, 74.00]. Also, we obtain a p-value of 0.083 and ratio of 0.61 in the F-test which means his means that the variance of the discarded genes that were selected as new modules was not significantly lower than the old modules for the calculated cumulative decay score with a confidence interval [0.39, 1.07].

As for the linkers and the isolated genes, we obtain a similar pattern with the distinct I & L genes having a significant difference in mean (p-value of 0.0006) compared to the background mean of all other genes, and a significant difference in variance (p-value of 5.432e-18). On the other hand, the variance of the I&L discarded genes showed to be higher than the background but insignificantly (p-value = 0.053), while the mean showed the same pattern as the distinct selection of genes.

This means that the since the same pattern is exhibited in both cases (distinct vs discarded), we realize that the variance might be an important factor in deciding the classification of genes into M or I & L. Although both M and I & L showed a significantly higher mean, this might be due to neighbors’ decay score of M genes and network topology of I&L which requires further research and validation.

1. **Coverage & Enrichment**

We obtain the list of breast cancer (BC) related genes from three databases (eDGAR, drugbank, & MalaCards) using the AutoSeed R package. The intersection between this list and genes included in the study was taken. Coverage was then calculated as the percentage of genes shared between BC list and full sorted gene lists at different mND(p) cutoffs.

**Chart, line chart

Description automatically generated**

Figure

As observed in Figure 2, we can see that the lowest percentage of cumulative coverage was that of mND (k=3) and the highest was that of GS-adjusted mND (k=3). Besides, GS-adjusted mND (k=2) appeared to be consistent with GS-adjusted mND (k=3) for the first few mND(p) cutoffs but the difference begins to show as the cutoff increases, whereas the difference is evident in all cutoffs between mND (k=3) and the other two. In this sense, our algorithm outperforms mND in prioritizing genes no matter the value of k for the first few cutoffs. Nonetheless, k=3 for our algorithm performs the best in the task of prioritizing genes at all cutoffs of mND(p).

1. Connectivity

To account for the connectivity of genes, we use a scoring method proposed by Besanelli et, al. (2016), and used for mND; checks for significantly connected, functionally enriched nodes at different cutoffs:

**Ω**

In our protocol we were able to permute over only the top 1000 genes, which produced results different from mND protocol: due to time constraints. Regardless, the DE scores were insignificant for lower cutoff of sorted genes. The distribution of enriched connectivity score showed a consistent variation (decrease) from our protocol with k=2, to k=3 then to mND protocol alone. These results need to be reassessed with permuting for more than 1,000 genes. The difference realized in enriched connectivity score for original DE scores is most probably due to initial difference in the scores.

Chart, box and whisker chart

Description automatically generated

Figure 3

Chart, box and whisker chart

Description automatically generated

Figure 4

1. Optimizing K

OPTIMIZE…..

**LIMITATIONS & FURTHER RESEARCH**

The limitations of our work include the absence of samples data which would allow us to obtain p-sphere of influence in using GS which can be merged with p-decay using Fisher Method to calculate p-fisher. GS shows p-fisher can capture the decay and influence of a gene more significantly than only one of the latter p-values. Further, GS can be optimized through heuristics or Genetic approach, to cut down the running time to make it feasible for other research to replicate our work. In addition, Cumulative Decay Score Variance showed the potential to use Cumulative Decay Score as a score to classify genes or as a cutoff for classification. Also, Omega values require more validation and optimization as running the workflow on 1000 genes was extremely time consuming. Other methods to validate the best K and calculation of Omega could be taken into consideration in further research. Finally, comparing our results to DMFIND and other ND algorithms might reveal new insights about our protocol.

**REFERENCES**

1. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43(7):e47.
2. Manoli T, Gretz N, Gröne HJ, Kenzelmann M, Eils R, Brors B. Group testing for pathway analysis improves comparability of different microarray datasets. 2006; 22(20):2500.
3. Braun R, Shah S. Network methods for pathway analysis of gene expression data. 2014. arXiv [q-bio.QM]. arXiv. <http://arxiv.org/abs/1411.1993>.
4. Nikolayeva, I., Guitart Pla, O., Schwikowski, B. (2018). Network module identification—a widespread theoretical bias and best practices. *Methods* 132, 19–25. doi: 10.1016/j.ymeth.2017.08.008
5. Cowen, L., Ideker, T., Raphael, B. J., & Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, *18*(9), 551-562.
6. Di Nanni, N., Bersanelli, M., Milanesi, L., & Mosca, E. (2020). Network diffusion promotes the integrative analysis of multiple omics. *Frontiers in Genetics*, *11*, 106.
7. Bersanelli, M., Mosca, E., Remondini, D., Castellani, G., & Milanesi, L. (2016). Network diffusion-based analysis of high-throughput data for the detection of differentially enriched modules. *Scientific reports*, *6*(1), 1-12.
8. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology (Poznan, Poland)*, *19*(1A), A68–A77. <https://doi.org/10.5114/wo.2014.47136>
9. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, Sabedot TS, Malta TM, Pagnotta SM, Castiglioni I, Ceccarelli M, Bontempi G, Noushmehr H. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016 May 5;44(8):e71. doi: 10.1093/nar/gkv1507. Epub 2015 Dec 23. PMID: 26704973; PMCID: PMC4856967.
10. Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., ... & Mering, C. V. (2010). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, *39*(suppl\_1), D561-D568.