SOFTWARE TOOL ARTICLE

MSF: Modulated Sub-graph Finder [version 1; referees: 1 approved, 1 approved with reservations, 1 not approved]

Mariam R. Farman¹, Ivo L. Hofacker¹, Fabian Amman²

¹Institute for Theoretical Chemistry, Theoretical Biochemistry Group, University of Vienna, Vienna, 1090, Austria
²Department of Chromosome Biology, Max F. Perutz Laboratories, University of Vienna, Vienna, 1030, Austria

Abstract
High throughput techniques such as RNA-seq or microarray analysis have proven to be invaluable for the characterization of global transcriptional gene activity changes due to external stimuli or diseases. Differential gene expression analysis (DGEA) is the first step in the course of data interpretation, typically producing lists of dozens to thousands of differentially expressed genes. To further guide the interpretation of these lists, different pathway analysis approaches have been developed. These tools typically rely on the classification of genes into sets of genes, such as pathways, based on the interactions between the genes and their function in a common biological process. Regardless of technical differences, these methods do not properly account for cross talk between different pathways and rely on binary separation into differentially expressed gene and unaffected genes based on an arbitrarily set p-value cut-off. To overcome this limitation, we developed a novel approach to identify concordantly modulated sub-graphs in the global cell signaling network, based on the DGEA results of all genes tested. Thereby, expression patterns of genes are integrated according to the topology of their interactions and allow potentially to read the flow of information from the perturbation source to the effectors. The described software, named Modulated Sub-graph Finder (MSF) is freely available at https://github.com/Modulated-Subgraph-Finder/MSF.

Keywords
Differential gene expression analysis, pathway analysis, combining p-value, cell signalling network
Corresponding author: Ivo L. Hofacker (ivo@tbi.univie.ac.at)

Author roles: Farman MR: Data Curation, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation; Hofacker IL: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing; Amman F: Conceptualization, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by the FWF ("Fonds zur Förderung der wissenschaftlichen Forschung") within the project Internationalen Kooperationsprojektes - Inf cooperation Project (Joint Project - Lead Agency Verfahren) [I 1988-B22]. The grant was assigned to ILH. FA was funded by the Austrian Science Fund (FWF) [SFB F43]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2018 Farman MR et al. This is an open access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Farman MR, Hofacker IL and Amman F. MSF: Modulated Sub-graph Finder [version 1; referees: 1 approved, 1 approved with reservations, 1 not approved] F1000Research 2018, 7:1346 (https://doi.org/10.12688/f1000research.16005.1)

Introduction

High throughput sequencing techniques have been widely used to yield differentially expressed genes (DEG) (Malone & Oliver, 2011). To this end, changes in transcript abundance are measured, e.g. by next generation sequencing techniques, and interpreted as an indicator of differential expression of genes. DEGs can be used to gain insights into the mechanism underlying differences between conditions of samples, such as healthy versus diseased. Differential gene expression analysis (DGEA) informs about the magnitude of expression changes between the conditions which are often expressed as fold change, sign of fold change and the confidence level of observing an authentic change, often expressed as \( p \)-value. These DEGs information is further interpreted to extract meaningful biological insights. For example, genes that could be involved in the response to a particular stimuli or maybe the cause of a disease. To this end, pathway-based analysis has become an important tool to further interpret the results of a DGEA and to acquire understandings of the perturbations in a biological system. Biological pathways are sets of genes and their interactions forming a functional unit. DEGs help to identify pathways or networks that may be altered during a change of condition providing important information about diseases and its treatment process (Khatti et al., 2012).

Pathway-based methods use predefined pathways or networks such as KEGG (Kanehisa & Goto, 2000) and Reactome (Fabregat et al., 2018), the expression measurements of the genes obtained from DGEA in combination with statistical methods and algorithms to identify specifically modulated pathways and processes (García-Campos et al., 2015).

Well established resources for pathway annotation are KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa & Goto, 2000) and Reactome (Fabregat et al., 2018). KEGG pathways is a branch of KEGG database that hosts a collection of manually drawn pathway maps representing the molecular interaction, reaction and relation networks of cellular functions. Similarly, Reactome is an open-source, manually curated, peer-reviewed database for signaling and metabolic molecules with their interactions formed into different biological pathways for nineteen species (Fabregat et al., 2018). Both provide predefined pathways which are sets of genes and their interactions categorized into functional units. Starting from a gene interaction network, genes are labeled according to their role in a specific biological process. In this sense a particular gene can be assigned to different pathways. E.g., the human gene STAT1 is associated with 24 different pathways in the pathway annotation curated by KEGG and in 12 different pathways in the Reactome database. Although carefully produced, the assignment of genes to those predefined pathway units can be considered to be subjective to some degree and suffers from observational bias (Schnoes et al., 2013).

Existing pathway-based analysis approaches use different research designs, which can be categorized into ORA (Over-representation analysis), FCS (Functional class scoring) and pathway topology based methods. All of which aim to find a subset of genes, e.g., significantly differentially expressed genes, genes associated with a certain pathway more often than expected given the total set of examined genes, e.g. the whole genome back-ground. ORA is the first and the most basic method of pathway analysis (García-Campos et al., 2015). It uses a DEG list with user defined cut-off for the log-fold change and \( p \)-value (most commonly using absolute log-fold change \( \geq 2 \) and \( p \)-value \( \leq 0.05 \)). Subsequently, sets of genes associated with annotated pathways are tested for being over-represented in the set of DEGs. To this end, hypergeometric distribution, chi-square tests, binomial probability or the Fisher’s exact test are used. Thereby the information of the topology of the genes in the pathways is neglected (Bayerlová et al., 2015). Furthermore, ORA assumes that the biological pathways are independent of each other and ignores the fact that they cross-talk and overlap (García-Campos et al., 2015; Khatti et al., 2012).

Unlike ORA, FCS has no artificial cut-off to define a DEG list. FCS works in three steps. First it calculates the gene-level statistics including correlation of molecular measurements using differential expression of individual genes, ANOVA, t-test and Z-score. In the second step the statistics of individual genes in a pathway are transformed to an individual pathway-level statistic commonly using Kolmogorov-Smirnov statistic, mean or median. Finally the statistical significance of the pathwaylevel statistics is assessed. Although FCS covers some of the limitations from ORA, it still ignores the topology of genes in a pathway, cross-talk and overlap of the pathways (García-Campos et al., 2015; Khatti et al., 2012). Pathway topology based methods are similar to FCS except that they consider the topology of each gene during the gene-level statistics but still don’t aim to link different functional pathways (Khatti et al., 2012).

On these grounds we propose a novel approach to make use of the rich gene and protein interaction annotation resources available to gain additional functional insights from basic DGEA. To this, we start with the presupposition that expression of neighboring genes within a functional pathway are not independent from each other. Rather, they are often regulating each others expression or are part of the same regulon (Michalak, 2008). We understand that the categorization of links between genes into labeled pathways is often an arbitrary one, given the extensive cross talk between different pathways. Although these categories have proven to be useful in many situations, they force a certain perspective onto the interpretation of novel data. Based on these two principles, we aim to find sub-graphs of connected genes within cell signaling network which exhibit as a whole significant differential expression changes. This approach differs in two main aspects from common pathway analysis tools. First, it does not aim to identify functional pathways enriched in differentially expressed genes, but detects sub-graphs or branches in a network graph (potentially spanning more than one functionally grouped pathway) which is coherently modulated. Second, it aims to improve the DGEA on the gene level, by collecting the information of neighboring genes, which as a whole might exhibit prominent enough signal to be called, again as a whole, significantly modulated.

As input, information on functional links between genes provided by e.g. KEGG or Reactome and information on the differential expression status of single genes resulting from a DGEA, are required. As a result the analysis returns sub-graphs and their joint
confident scores, reflecting how the perturbation is migrated through the network. Furthermore, the entry points of perturbation in the networks and overlap with conventional pathway categories are returned. The output is prepared in a directed adjacency file, convenient for visualization, e.g., with StringApp (Morris et al., 2018), available as a Cytoscape plug-in (Shannon et al., 2003).

All of this can be helpful to understand the cause and effect of a stimulus and might inform about potential points of intervention. The proposed algorithm was implemented as a java program, which was named Modulated Sub-graph Finder (abbreviated MSF). MSF can help transform the information obtained from DGEA into comprehensible knowledge of signal transduction of genes and thereby being a valuable complement to existing pathway based methods. MSF is freely accessible from GitHub https://github.com/Modulated-Subgraph-Finder/MSF.

Methods

Implementation

MSF is developed as a novel heuristic approach to find concertedly modulated sub-graphs in networks of biological interactions. MSF does not use predefined gene sets grouped into functional units, but rather relies purely on the network of interacting genes. The input network consists of nodes corresponding to genes and edges representing interactions. Furthermore, it utilizes comprehensive results from a differential gene expression analysis to discover the sub-graphs, or modules, which are as a whole modulated.

MSF uses the individual gene’s p-values generated from the DGEA. The p-value expresses the probability that the null hypothesis of unmodified gene expression can’t be rejected for a given statistical model. To find significantly modulated sub-graphs individual p-values of the vicinal genes in the global network are combined into a single combined p-value, using a statistical method for combining dependent p-values described by Hartung (Hartung, 1998). Hartung’s method uses the inverse of standard normal distribution function, individual gene p-values are first transformed to their corresponding normal score. Then using these normal scores, the correlation between genes is calculated, the normal scores and correlation are applied to the inverse normal function to calculate the combined p-value for all genes examined, namely the examined sub-graph. The combined p-value of a sub-graph will express the significance of the sub-graph being modulated together. Thereby, the information from the different genes are used as, although not independent, replicated measurement of the behavior of the whole sub-graph. This potentially increases the power to detect also significant sub-graphs consisting of genes which are not significant on their own.

Overview of our method

To reduce the complexity to score all possible connected sub-graphs MSF applies a four step heuristic as described in the following. The proceeding identification of modulated sub-graphs from a network by MSF is presented as a flowchart diagram (Figure 1).

Initial modulated sub-graphs. MSF constructs the first sub-graph starting with the genes associated with the lowest (most significant) p-value deduced from the DGEA. From this seed it tries to extend the sub-graph by adding directly neighboring genes, starting with the most significant one. A single combined p-value is calculated for the extended sub-graph. If the combined p-value is smaller than the p-value of the original sub-graph, the extended sub-graph is accepted. This step is iteratively repeated until no further extension is accepted. In this case the process starts over with all remaining genes not yet in a significantly modulated sub-graph. This step identifies all simple sub-graphs that are modulated in the whole network.

Extending modulated sub-graphs. In the next step, we check if any of the initial modulated sub-graphs could further be extended by adding more than one gene at a time. This is done by testing all possible extension paths up to N (default 2) genes at all nodes in the sub-graph. Again, this step is iteratively repeated until no further genes are added to the significant differentially expressed sub-graphs. This step bridges small gaps of genes without a clear differential signal in the DGEA.

Merging modulated sub-graphs. After detection and extension of the modulated sub-graphs, they are tested if combined sub-graphs score better than on their own. The merging of the sub-graphs is done by depth first search traversal from the first sub-graph to the second sub-graph. If the two sub-graphs merge with the connector of at most N genes (default 1 gene) and the combined p-value of the merged sub-graph including the bridging genes in between is less than the individual p-values of the two sub-graphs, the two sub-graphs are merged together to one big modulated sub-graph. This step is repeated iteratively until no sub-graphs could be merged.

Finding sources & sinks. In a last post processing step MSF identifies the trigger points of the modulated sub-graphs. These trigger genes are the sources of the sub-graphs with only outgoing edges. These genes can be interpreted as the possible entry points of perturbation from where the stimulus causes downstream effects. In the same spirit the most downstream genes of the modulated sub-graph are identified and defined as sinks. Sinks can be interpreted as the effectors where the integrated information within the signal transduction network is set to action. Due to circular loops not all sub-graphs are guaranteed to have sources or sinks.

Operation. MSF requires Java version 8 and JDK 1.8. The few package dependencies are already been added to the release. MSF runs fast on a standard laptop computer and so it has normal system requirements. To run MSF, the user must provide two text files, one containing the DGEA and the second one containing the interactions in a adjacency format file. Example files and a detailed tutorial to use MSF has been provided on github https://github.com/Modulated-Subgraph-Finder/MSF.

Results

Case Study

To demonstrate its usefulness, MSF is applied to an RNA-seq data set of primary human monocytederived macrophages (MDMs) infected with Ebola virus (GSE84188) (Olejnik et al., 2017). Ebola
Virus (EBOV) belongs to the Filoviridae family; filamentous, enveloped and single stranded RNA viruses. EBOV causes hemorrhagic fever in humans, inducing the host innate and adaptive immune response to be unable to control virus infection (Prins et al., 2009). Until now, there are no approved antiviral drugs for the treatment of Ebola virus infection (Konde et al., 2017; Rhein et al., 2015). The initial targets of EBOV are the macrophages and dendritic immune cells (Falasca et al., 2015; Rhein et al., 2015). EBOV inhibits the critical innate immune response of the host, which includes the activation of alpha/beta interferon (IFN-α/β) (Cárdenas et al., 2006; Konde et al., 2017; Prins et al., 2009). It has been proposed that IFNα/β should be tested against Ebola for its antiviral activity through clinical trials (Konde et al., 2017). Ebola infection data was selected to test the approach because it has been well recognized for the last several decades, and vast literature is available for the pathogenesis of Ebola. Thereby facilitating the verification of the results of MSF with the vast literature present on Ebola infection. Especially, the detection of IFN-α/β as point of action for the virus, could be considered as a basic indicator of the correctness and usefulness of the approach.

EBOV infection count data was downloaded from Gene Expression Omnibus (GEO) (GSE84188). Differential gene expression analysis was performed on the count data with edgeR package (version 3.4.2) (Robinson et al., 2010). The DEG analysis results generated by edgeR were used as input for MSF. Directed cell signaling interactions were filtered from Reactome Functional interactions (FIs) Version 2016 (Wu et al., 2016), which was used as a second input for MSF.

The EBOV infection experiments describe the course of infection at three time-points 6, 24 and 48 hour post infection (hpi). For the earliest time point at 6 hpi, three large modulated sub-graphs were identified with 41, 107, and 18 genes were identified, as well as two with less than 4 genes. The modulated sub-graphs consist predominantly of cytokines, chemokines (CXCL10, CCL8) and Interleukin genes (IL6, IL27, IL23). IFNB1 and IFNA1 were both
identified as two of the possible sources in one of the sub-graph identified with 18 genes. Most of the sources identified by MSF were type I interferon induced genes (Supplementary material 6H). At 24 hpi eight modulated sub-graphs were identified with four main sub-graphs consisting of 27, 101, 134 and 194 genes, others being smaller than 10 genes. IFNβ1 and IFNA1 were identified as the two sources out of the total sources. For the last time-point 48 hpi, six modulated sub-graphs were identified. Three of the sub-graphs were less than nine genes and main sub-graphs had 122, 191 and 202 genes. IFNβ1 was identified as one of the sources in the most significantly modulated sub-graph (Supplementary).

As stated earlier IFN-α/β was reported to be one of the target genes of Ebola infection. We were able to successfully identify IFNA1 as a source in all three time-points and INFB1 in two of the time-points. Although IFNA1 and IFNB1 were already among the most significant gene in the DGEA during the later time points, MSF was also able to detect them as a source in the very early time-point when the genes were not significant based on the individual DGEA alone. Identifying the possible sources will reduce the search space for potential target genes and can help the biologist as the starting point of clinical testing for drugs and vaccines against an infection.

Table 1 compares the results of MSF, namely the number of detected sub-modules and their total genes numbers, to a simple analysis of mapping significantly modulated genes from the DGEA to the network and joining neighbors to modules. The numbers indicate that MSF detects less but larger sub-modules, applying its statistical test. Furthermore, the dependency of the results from the p-value cutoff choice is demonstrated for the DGEA, which is avoided for MSF altogether.

**Table 1. Comparison of connected sub-graphs of modulated genes in the global network between the analysis results of Modulated Sub-graph Finder (MSF) and mapping the raw list of differentially expressed genes from the standard edgeR analysis, applying different p-value cut-offs, onto the network.** hpi - hours post infection.

<table>
<thead>
<tr>
<th></th>
<th>Total number of genes in network</th>
<th>Number of connected sub-graph in network</th>
</tr>
</thead>
<tbody>
<tr>
<td>6hpi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>edgeR</td>
<td>166</td>
<td>3</td>
</tr>
<tr>
<td>edgeR (p-value ≤ 0.1)</td>
<td>123</td>
<td>39</td>
</tr>
<tr>
<td>edgeR (p-value ≤ 0.05)</td>
<td>123</td>
<td>47</td>
</tr>
<tr>
<td>edgeR (p-value ≤ 0.01)</td>
<td>112</td>
<td>53</td>
</tr>
<tr>
<td>24hpi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>edgeR</td>
<td>456</td>
<td>4</td>
</tr>
<tr>
<td>edgeR (p-value ≤ 0.1)</td>
<td>325</td>
<td>76</td>
</tr>
<tr>
<td>edgeR (p-value ≤ 0.05)</td>
<td>305</td>
<td>102</td>
</tr>
<tr>
<td>edgeR (p-value ≤ 0.01)</td>
<td>262</td>
<td>110</td>
</tr>
<tr>
<td>48hpi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>edgeR</td>
<td>515</td>
<td>3</td>
</tr>
<tr>
<td>edgeR (p-value ≤ 0.1)</td>
<td>392</td>
<td>43</td>
</tr>
<tr>
<td>edgeR (p-value ≤ 0.05)</td>
<td>366</td>
<td>74</td>
</tr>
<tr>
<td>edgeR (p-value ≤ 0.01)</td>
<td>306</td>
<td>106</td>
</tr>
</tbody>
</table>

In more detail, sub-graph 1 (top) shows how the activation of toll-like receptor, cytokine, chemokine and jakstat genes lead via TNF into apoptosis. The next significant sub-graph (sub-graph 2: middle) reveals how information from the Extra-cellular matrix (ECM) receptor, which are reported to interact with Ebola glycoprotein (GP) (Veljkovic et al., 2015), chemokines and cytokines, and cytosolic DNA sensing, is integrated into again modulation of apoptosis pathway. Eventually, sub-graph 3 (bottom) demonstrate how INFA1 and INFB1 modulate once more, via only a few intermediate steps, the apoptotic response of the cell.

This show cast example demonstrates with how little effort complex data can be interpreted, help to apprehend the dynamics of the underlying processes and suggest testable hypothesis and potential points of intervention.

**Robustness**

A potential concern is how noise in the gene expression measurements affects our analysis. To assess the robustness and stability of our method, we therefore added Poisson distributed noise to the read counts of the three time-points data set, used above. Then DGEA was carried out on the disturbed data with the same parameters as for the native data using edgeR, followed by analysis with MSF. This procedure was carried out 100 times. Every time the genes from the modulated sub-graphs identified from noisy data were compared to the genes of sub-graphs identified from the native data. This was also done for the DEG obtained for each run, using three different cutoffs of FDR 0.01, 0.05 and 0.1. The robustness of MSF and the DEG analysis for the time-point 6, 24, and 48 hpi are shown in Figure 3. The procedure for how data noise was modeled can be considered as rather stringent, which is already reflected by the limited recall rate in the edgeR based DGEA, between 68 % (6 hpi) and 93 % (24 hpi). For MSF-analysis the observed median recall rates lay between 71 % (6 hpi) and 84 % (48 hpi). The better performance of the pure DGEA can be explained by the fact that disturbed p-values do not change the results for DGEA as long as the p-value does not rise above the chosen cutoff value. In contrast, MSF is sensitive to p-value changes across the whole range of possible values.
Figure 2. Visualisation of the three modulated sub-graphs identified by Modulated Sub-graph Finder (MSF) at 6 h after Ebola Virus (EBOV) infection in gene detail and their epitomized representation to depict the flow of perturbation in the directed network. The node coloring is associated to KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways referring to the colors in the legend. The graph edges are from Reactome. Important genes to EBOV infection as from literature are enlarged in the graph.

Figure 3. Shows the percentage of differentially expressed genes (DEG) analysis (3 different cut-offs) genes and Modulated Sub-graph Finder (MSF) identified sub-graph genes recall rates for the three different time points of Ebola Virus (EBOV) infection data for 100 simulation where Poisson distributed noise was added to the experimental deduced reads per gene counts.
Comparison to Reactome pathway analysis tool

Gene enrichment analysis was performed using Reactome Analyze data tool (Fabregat et al., 2018). Reactome’s over-representation analysis tests whether certain Reactome pathways are enriched by the list of genes submitted. Genes from MSF identified sub-graphs for each time-point were analyzed for gene enrichment analysis. For comparison the DEG results from edgeR were filtered using three different cut-offs of adjusted p-value 0.01, 0.05 and 0.1. This three subsets of DEG list were used for gene enrichment analysis. The compression of MSF and the three subset DEG list is shown in Figure 4.

The comparison shows most of the pathways known from literature to be dis-regulated by Ebola infection are enriched in both the enrichment analysis. Toll-Like receptor signaling pathway when interacts with EBOV glycoprotein (GP), it triggers the activation of cytokines (Olejnik et al., 2017). Toll-like receptor pathway is expected to be dis-regulated in the early stage of infection, this pathway was not identified as significantly dis-regulated when p-value cut off DEG lists were analyzed for enrichment. Ten toll-like receptor cascades were identified as dis-regulated from gene enrichment analysis of MSF identified genes, not a single one of these pathways was shown to be dis-regulated in DEG cut off lists. Since MSF considers the complete DEG results, even the weak signal at the earliest time-point was detected; for example Toll-like receptor signaling.

Discussion

Classic pathway analysis tools aim to detect in lists of significantly deregulated genes enriched associations with pathway genes categorized by their biological function and their interactions. Thereby, depending on the tool, the internal pathway topology is considered or neglected all together. The here presented tool, MSF, employs a different approach, by aiming to detect sub-graphs in whole gene regulatory networks which are significantly deregulated in a concerted manner. To this end, neighboring genes in the user provided network are tested for jointly common regulation. Exploiting that each gene’s abundance, although not independent from its neighbors, is measured repeatedly on its own, sensitivity can be increased by our applied p-value meta-analysis, namely Hartung’s method. This potentially enables to call non-significant modulated genes based on the DGEA, to be

Figure 4. The Upset plot shows the number of shared pathways between different time-point and cut offs. All the different toll-like receptor signaling pathways identified from Modulated Sub-graph Finder (MSF) identified genes are marked.
convincingly part of a deregulated gene group. Furthermore, it allows to identify connected sub-graphs, representing the propagation of gene regulation perturbation in the input network. A better understanding of this propagation, especially the critical spots such as sensors, effectors, or intermediate bottlenecks and hubs, facilitates the projection of potential intervention points, e.g., for drug development. Since MSF only uses interaction information in gene regulation networks, but not the functional grouping of the genes into functional pathways, it is especially adapted to discover so called cross-talk between such pathways.

Conclusions

MSF is a fast, robust and easy to use tool to find concertedly modulated sub-graphs in a given network. Its implementation in JAVA enables its use across many operating systems. It requires as input the results from a differential gene expression analysis in the appropriate file format. So far the raw output from edgeR (Robinson et al., 2010) and DESeq2 (Love et al., 2014) are supported. Furthermore, a gene network has to be provided in the format of a directed adjacency list.

Data availability

The Ebola infection RNA-seq data set analyzed during the current study are available in the GEO repository (GSE84188) (Olejnık et al., 2017). The cell signaling network file used is from Reactome Functional interactions (FI) Version 2016 (Wu et al., 2016).

Software availability

Source code is available from GitHub: https://github.com/Modulated-Subgraph-Finder/MSF


Software license: MIT license.

Grant information

This work was funded by the FWF (“Fonds zur Förderung der wissenschaftlichen Forschung”) within the project Internationalen Kooperationsprojektes - Intl cooperation Project (Joint Project - Lead Agency Verfahren) [I 1988-B22]. The grant was assigned to ILH. FA was funded by the Austrian Science Fund (FWF) [SFB F43].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplementary material

Supplementary material is available form GitHub: https://github.com/Modulated-Subgraph-Finder/MSF

References


Malone JH, Oliver B: Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 2011; 9(1): 34. Published Abstract | Publisher Full Text | Free Full Text


Page 9 of 17


Reference Source
Open Peer Review

Current Referee Status: ❌ ✔️ ?

Version 1

Referee Report 29 November 2018

https://doi.org/10.5256/f1000research.17481.r40208

Haibo Liu
Department of Animal Science, Iowa State University, Ames, IA, USA

In this manuscript, the authors reported their newly developed tools, MSF, for interpreting gene lists from differential gene expression analysis. Their tool differs from existing pathway analysis tools: (1) it can identify concertedly modulated sub-graphs from user-provided gene networks, thus it can accounting for cross-talk between pathways; (2) it can potentially infer the flow of biological information in response to a perturbation from source to sink. Like the gene set enrichment analysis (GSEA), no arbitrary p-value threshold is set to dichotomize whole gene lists before applying MSF analysis, instead all genes from DGEA are ordered by p-values from the smallest to the largest. An algorithm similar to the widely used network propagation algorithm is used for subgraph initialization and extension. The authors applied their tools to analyze an RNA-seq dataset from an Ebola virus infection experiment and showed their tool outperform the other tools. They concluded this tool, fast, robust and easy-to-use, is a good supplement to existing pathway-based analysis tools. However, the overall writing is very problematic and there are quite a few issues needing to be fixed. See the details listed as follows.

1. First of all, the code for the tool is not available via Github. I carefully checked multiple times the Github repository: https://github.com/Modulated-Subgraph-Finder/MSF, however I can't find the ModulatedSubPathFinder.jar file, which is the Java implementation of the proposed tool, MSF.
2. There are too many grammar issues and writing issues. Just mentioned a few, in the first paragraph of Introduction, "mechanism" in Line 6 should be plural, while "stimuli" in Line 14 should be "stimulus", "maybe" in Line 15 should be "may be". Careful proof-reading is strongly recommended.
3. The authors have a few misconceptions. For example, they treated "effectors" and "sinks" equally. In my opinion, effectors include sources, intermediate genes and sinks, i.e., all genes responding to perturbations. The authors think of the significance of statistical tests in the form of p-values as a confidence level of observing an authentic expression change. This might not be correct. Besides small p-values, the magnitude of fold changes is also important metric of authentic expression change. By the way, the fold change of gene expression is always non-negative. The expression of "sign of fold change" is not meaningful. Only log-transformed fold change is signed.
4. The flow of information/idea is not fluent in many places. For example, at the end of the first paragraph of Introduction, the authors mentioned the KEGG and Reactome Pathways. Then at the beginning of the second paragraph, they gave a detailed description of the two pathway databases, which might be unnecessary and disrupted the flow to set up the stage to introduce why their tool is necessary and useful. Some information about how their tool was implemented given in the last paragraph of Introduction should be moved to the Implementation section of Methods.

Paragraph 3 under the section of "Case Study", the DEGA results might better
section of Methods. Paragraph 3 under the section of “Case Study”, the DEGA results might better be described immediately after the second sentence of Paragraph 2. Similarly, some information in Conclusion should be move to the section of “Implementation”.

5. The title for the section of “Initial modulated sub-graphs” should be “Initializing modulated sub-graphs”. Under this section, “starting with the most significant one” should be “starting with the next most significant one”, “… not yet in a significantly modulated …” should be “…not yet in the significantly modulated …”.

6. Under the section of “Extending modulated sub-graphs”, it is not clear how the sub-graphs are extended by adding “MORE THAN ONE” gene at a time. If doing this way, there are infinite possibilities. The criterion to accept or reject added genes is not clear.

7. Under the section of “Merging modulated sub-graphs”, the authors mentioned that “After detection and extension of the modulated sub-graphs, they are tested if combined subgraphs SCORE better than on their own.” At this point, no merging has been done yet, how are the combined subgraphs tested? What is the SCORE used here? How can a depth-first search traverse from the FIRST sub-graph to the SECOND sub-graph before they are merged (Aren’t the subgraphs not necessarily connected?)?

8. Under the section of “Finding sources & sinks”, “circular loops” should be just “loop”.

9. Paragraph 2 Under “Case Study”, some details about edgeR-based DEGA are missing. How the directed cell signaling interactions are filtered from the Reaction Fls? Based on what?

10. The directions of edges in right panel of Figure 2 should be showed, because the MSF can generate directed subgraphs.

11. In the last paragraph of the section of “Modulated sub-graphs at 6 hpi”, “This show cast example…” should be “This show case example…”.

12. The Robustness test seemed to show that the MSF is not robust enough to extra noise. What is the authors’ conclusion and explanation?

13. The authors compared the results from their tool to those from the Reactome pathway analysis tool and demonstrated better performance of their tool. Can they also compare the results from their tool to those from the GSEA tools, which don’t set arbitrary cutoff beforehand? The latter comparison might be more convincing.

14. In Discussion, what are “intermediate bottlenecks”? In Conclusion, the authors claimed their tool is “fast, robust and easy-to-use”. However, they did provide evidence to show their tool is fast. The robustness of their tool is not apparent.

15. Issues with figures: all legend titles are too long. In Figure 1, Texts in the flow chart symbols are not well-written. There are inconsistent case issues; “initial” should be “initializing”; symbol for condition test (“checked all interaction?”) should be diamond, not hexagon. Question marks might be added to condition test for readability. In the legend to Figure 1, what does “without exhaustively testing all connected subgraphs” mean? Does it the output from MSF analysis might not be comprehensive? Legends to Figure 3 and 4 are poorly written. Is the Toll-like receptor signaling one of the 149 shared pathways in Figure 4. It is not clearly described in the legend.

16. Limitations of their tool: In their implementation, the fold changes and direction of changes are not taken into account. If a resulting subgraph contains both down-regulated genes and up-regulated genes, how should the users interpret it? The authors didn’t test the sensitivity and specificity of their tool in this manuscript.

**Is the rationale for developing the new software tool clearly explained?**
Yes

**Is the description of the software tool technically sound?**
Yes
Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

**Competing Interests:** No competing interests were disclosed.

**Referee Expertise:** bioinformatics and computational biology, transcriptomics and systems biology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Referee Report 12 November 2018

https://doi.org/10.5256/f1000research.17481.r40009

Stefanie Widder  
Medical University of Vienna, Department of Medicine 1, Research Lab of Infection Biology, Waehringer Guertel 18-20, 1090 Vienna, Austria

The authors present a novel method for finding groups of genes that concomitantly change their expression profile upon signals and conditional changes. As opposed to standing tools in the field that rely on predefined functional classification, this approach is based on topology of the global interaction network and enables an unbiased identification of larger functional blocks highlighting pathway cross-talk. It furthermore includes a topological method for identifying source and sink pathways that provides a prediction of process causality. The latter is particularly useful for hypothesis generation and add-on experimental validation far beyond the field of biomedicine.

The paper is structured into the presentation of the algorithm, a biomedical use-case with validating background information and a slim benchmarking against differential expression analysis with different cut-offs.

While the method clearly fills an existing gap in high-throughput gene expression analysis and is very elegantly reasoned, I would like to raise a number of comments with regards to writing and benchmarking.

**Introduction:**
1. The main line of argumentation gets lost sometimes, in particular in paragraph 1. The narrative would furthermore benefit from actual examples instead of repeating a general statement of 'changed conditions'. Sometimes, the same argument is repeated in differently phrased sentences.

2. Paragraph 2: Suppress 'be subjective to some degree'.

3. Paragraph 4: Better highlight and delimit the novelty of the presented approach.

4. Overall, shorter sentences benefit the reader.

Methods:

The context of Hartung's method is described nicely, yet the actual way how the individual p-values are combined to result in a single measure is omitted. Please include as this information provides instructive benefit to the reader.

Results:

Case Study:

1. Paragraph 1: Readability would profit from focusing the background information, e.g. 'Ebola infection data was (->were) selected (..)' - rephrase into a half-sentence.

2. Wording: Until now->currently.

Robustness:

1. The recall-comparison (MSF to differential expression analysis) is weakening the proposed method, because no increase in recall for MSF can be achieved. It would be informative to see also precision stats in comparison. Generally, one would expect more robust statistics on larger subgraphs (MSF vs. DEG groups).

2. Also, I am not entirely convinced by adding noise to real and thus – already - noisy data. A small, artificial mock example with detailed known outcome (+/- noise) might be more suitable and supportive.

Figure 4:

This figure is very difficult to follow. I suggest to i) enhance the label sizes and texts, with particular emphasis on delimiting MSF and DEG, ii) improve the figure caption text including fundamental information of what is shown.

Github material:

Importantly - provide an example directory that contains complete examples cases (+*all* files) including the presented Ebola case + options and outcome to enable a swift recapitulation of the tool for the new user.
Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

**Competing Interests:** No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

---

**Guanming Wu**
Department of Medical Informatics and Clinical Epidemiology (DMICE), Oregon Health and Science University (OHSU), Portland, OR, USA

In this manuscript, Farman et al described an approach to search for network modules based on p-values collected from differential gene expression analysis to address pathway crosstalk, which cannot be addressed in conventional pathway enrichment analysis approaches. During the past decade, many network module-based approaches have been developed to understand the functions of genes collected from differential gene expression analysis and other omics approaches (for a review, see Mitra et al, 2013). Though the network approach described here has some innovative ideas (e.g. searching for sources and sinks in subgraphs), however, the authors introduce their approach in the context of pathway analysis, without mentioning these previously published similar approaches, let alone comparing their approach to others. Also, it is worthy of mentioning that the described approach in this manuscript is very similar to jActiveModule (Ideker et al, 2012), the first of this kind, widely used for network-based data analysis.

The manuscript used the Ebola virus (EBOV) time course gene expression data set to show case MSF, trying to demonstrate the validity and usefulness of the approach. Indeed, the authors found that IFNA1 and IFNB1 are source genes in subgraphs across multiple time points, as reported by literature references. However, the authors have not discussed other genes in the found subgraphs, though the whole lists of them are provided in their GitHub site. IFNA1 and IFNB1 are among other source genes.
The authors should develop a statistic approach to evaluate p-values and FDRs for subgraphs and individual source genes, therefore, providing users a way to choose the most significant genes for unknown phenotypes or biological processes. The current way to showcase the usefulness of the approach is not stringent and may not be useful if too many genes are collected in the subgraphs.

The authors compared MSF results with raw gene lists based on p-value cutoffs (Table 1). However, Table 1 is not a fair comparison. Only the largest subgraphs are listed for edgeR + MSF, while all subgraphs are listed for raw gene lists (e.g., for 6 hpi, 3 for edgeR + MSF, 39 for edgeR (p-value <= 0.1)). For a fair comparison, all subgraphs should be listed for edgeR + MSF too.

Section “Comparison to Reactome pathway analysis tool” and Figure 4 compare results produced by Reactome pathway enrichment analysis for different gene lists. The whole section is confusing. First, the section title is misleading: the comparison is for results generated from Reactome pathway analysis tool, not to that tool. Second, I cannot see too much value in this setting using different adjusted p-value cutoffs for gene lists, probably one (e.g. 0.05) should be enough, to reduce the clutter in Figure 4. Third, the authors want to point out MSF can enrich Toll-like receptor signaling pathway, but not the raw gene lists. However, such a comparison has not clearly indicated in Figure 4. The set comparisons include too many gene lists. Fourth, the authors should point out what p-value or FDR cutoff value used to choose pathways from the Reactome analysis tool. It is not correct to choose all pathways listed in that tool for comparison. Finally, where “Ten” toll-like receptor cascade pathways come from?

Searching for sources/targets in individual MSF subgraphs based on the directions in the Reactome FI network and then drawing the schematic diagrams as illustrated in Figure 2 are interesting. It will be better to show directions in the Cytoscape network view (the right-side networks in Figure 2). The schematic diagrams in Figure 2 are interesting, but may dramatically simplify things occurring inside cells. The Reactome FI network provides functional relationships among genes or proteins, which are not necessary gene regulatory relationships. The authors should point this out in the manuscript.

Finally, the writing of this manuscript is under question. The authors should really read their manuscript much more carefully. There are far too many typos, wrong uses of punctuations, and grammar errors. For examples, “the Filoviridea family; filamentous” should be “the Filoviridea family: filamentous”; “pathogenesis of Ebola. Thereby facilitating” should be “pathogenesis of Ebola, thereby, facilitating”; “among the most significant gene in the DGEA” should be “among the most significant genes in the DGEA”, and many others.

References

Is the rationale for developing the new software tool clearly explained?  
Partly

Is the description of the software tool technically sound?  
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**
Partly

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**
Partly

*Competing Interests:* No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

---

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com