

# Detecting Perturbed Subpathways towards Mouse Lung Regeneration Following H1N1 Influenza Infection

Aristidis G. Vrahatis <sup>1,\*</sup>, Konstantina Dimitrakopoulou <sup>2</sup>, Andreas Kanavos <sup>1</sup>, Spyros Sioutas <sup>3</sup> and Athanasios Tsakalidis <sup>1</sup>

<sup>1</sup> Department of Computer Engineering and Informatics, University of Patras, Patras 26500, Greece; kanavos@ceid.upatras.gr (A.K.); tsak@ceid.upatras.gr (A.T.)

<sup>2</sup> Centre for Cancer Biomarkers CCBIO and Computational Biology Unit, Department of Informatics, University of Bergen, Bergen 5020, Norway; Konstantina.Dimitrakopoulou@uib.no

<sup>3</sup> Department of Informatics, Ionian University Corfu, Corfu 49100, Greece; sioutas@ionio.gr

\* Correspondence: agvrahatis@upatras.gr; Tel.: +30-694-767-7069

Academic Editor: Demos T. Tsahalis

Received: 31 December 2016; Accepted: 29 March 2017; Published: 3 April 2017

**Abstract:** It has already been established by the systems-level approaches that the future of predictive disease biomarkers will not be sketched by plain lists of genes or proteins or other biological entities but rather integrated entities that consider all underlying component relationships. Towards this orientation, early pathway-based approaches coupled expression data with whole pathway interaction topologies but it was the recent approaches that zoomed into subpathways (local areas of the entire biological pathway) that provided more targeted and context-specific candidate disease biomarkers. Here, we explore the application potential of PerSubs, a graph-based algorithm which identifies differentially activated disease-specific subpathways. PerSubs is applicable both for microarray and RNA-Seq data and utilizes the Kyoto Encyclopedia of Genes and Genomes (KEGG) database as reference for biological pathways. PerSubs operates in two stages: first, identifies differentially expressed genes (or uses any list of disease-related genes) and in second stage, treating each gene of the list as start point, it scans the pathway topology around to build meaningful subpathway topologies. Here, we apply PerSubs to investigate which pathways are perturbed towards mouse lung regeneration following H1N1 influenza infection.

**Keywords:** lung regeneration; systems biology; computation on networks and graphs

## 1. Introduction

We are going through the “Network Medicine” era, an emerging research field which has the potential to capture more realistically the molecular complexity of human diseases and provide -computational methodologies that can discern more efficiently how such complexity controls disease manifestations, prognosis, and therapy. It integrates “Systems Medicine” and “Network Science” fields to formulate unbiased large-scale network-based analyses in order to uncover this complexity -. However, the current high-throughput molecular technologies produce an unprecedented amount of biological data, posing a growing need for new “Network Medicine” tools to manage the complexity of “Big Data” and “Big Graphs” that are generated [1].

There is growing consensus that the advances in analysis methods fall behind relative to the massive amounts of omics data produced nowadays. In recent years, there was a paradigm shift that successfully moved the research focus from coupling diseases with single genes or single-nucleotide polymorphism (SNPs) to disease signatures or gene sets [2]. More recently, more sophisticated

systems-level approaches gained ground and pushed forward the transition from gene-to-gene analysis to signaling pathways and complex interaction networks, thereby gaining a more realistic and holistic insight into disease mechanisms [3].

Towards this orientation, pathway-based analysis has been proved to be efficient for comprehending biological mechanisms and disease etiology [4,5]. The main concept is a simplified analysis that groups single genes into sets of functionally related and interacting proteins. In this way, the complexity is reduced to a numerically feasible number at the magnitude of hundreds and, moreover, identifying “differential” pathways between two conditions has more explanatory power than gene lists. The first works in this field ignored the pathway interacting topology and used over-representation to compare the number of interesting genes that hit a given pathway with the number of genes expected to hit the given pathway by chance [4]. Later studies used Functional class scoring (FCS) to identify coordinated changes in the expression of genes in the same pathway [6]. Other approaches focused on the effect of the upstream genes relative to the downstream genes and coupled classical enrichment analysis along with the perturbation of a specific pathway to quantify the impact of upstream genes [7,8].

More recently, pathway analysis evolved to subpathway analysis, which searches for sub-areas on the topology to interpret the related biological phenomena and provides more targeted and context-specific molecular candidate signatures for disease etiology [9–16]. Subpathways are local subnetworks in the pathway topology which can be associated with small scale biological functions, within the boundaries of the pathway, and whose deregulation can give rise to a disease. Subpathway-based analysis has dealt with various challenges and signifies rightfully the next generation in pathway analysis [12]. Examining the entire pathway topology as one unit, hinders the detection of the small scale perturbations which might reflect a pathophysiological state or response to treatment [9]. Also, different pathway subnetworks may perform the same function in the same pathway and different pathways due the high overlap may use the same subnetworks in similar roles [9]. Subpathway-based tools, with their capacity to scan the entire pathway network and zoom into the specific subareas that are deregulated, can explore deeper the biological significance of disease-associated mutations identified by genome-wide association studies and full-genome sequencing. Hence, in the recent years several tools have been published under this perspective, offering new horizons in the Network Medicine field [9–16].

In previous work [17], we developed - Perturbed Subpathways (PerSubs) tool to extract - perturbed disease-specific subpathways from pathway networks. An important feature of the algorithm is that it identifies perturbed subpathways from KEGG pathway maps by using as starting point, prior to scanning pathway topology, a set of interesting gene-nodes (i.e., differentially expressed genes, disease-specific genes etc.). PerSubs utilizes a measure based on two multivariate logistic functions to set the co-expression status between the members of an interacting pair - as highly positive or negative. We applied PerSubs on a microarray experiment that included colony samples from control and H1N1 influenza treated lungs (12 days post infection) to study mechanisms towards lung regeneration following catastrophic damage [18]. Our results show that PerSubs can provide subpathways that reflect well processes related to tissue repair and development.

## 2. Materials and Methods

PerSubs algorithm [17] extracts perturbed subpathways from pathways taking into account graph topology and differential expression of the corresponding gene-nodes (Algorithm 1). Differential expression is used based on gene expressions from transcriptomics data. PerSubs extracts subpathways perturbed by a condition (disease or biological process) under study. Subpathways are extracted in the form of - densely connected subgraphs around nodes of interest based on topological criteria. For this, we follow a “seed growing” approach similarly to [19], where we start from an initial Node of Interest (NoI) and we identify the perturbation caused by this node in the entire pathway network. Users can provide a list of genes of interest, but here we selected as nodes of interest the significantly differentially expressed genes.

Node (gene/protein) differential expression intensity is calculated based on a geometrical multivariate effective approach, called the Characteristic Direction (chdir) [20]. It uses a linear classification scheme, which defines a separating hyper-plane, the orientation of which can be interpreted to identify differentially expressed genes (DEG). More specifically, it incorporates a regularization scheme to deal with the problem of dimensionality, and also provides an intuitive geometrical picture of differential expression in terms of a single direction. This geometrical picture reliably characterizes the differential expression and also leads to some natural extensions of the approach such as improved gene-set enrichment analysis.

In the computation of the characteristic direction, in order to identify differentially expressed genes, initially the steps below are followed:

1. Gene expression data have  $N$  samples, in which the expression of  $p$  genes is measured.
2. Each sample's expression profile forms a row of the matrix  $X$  ( $N \times p$ ) (each of sample's expression comes from one of  $K$  classes (e.g., disease or normal state) belonging to the set  $G$ ).
3. Bayes rule provides an expression for the class posteriors  $P(G|X)$ ,

$$P(G = k|X = x) = \frac{f_k(x)\pi_k}{\sum_{i=1}^K f_i(x)\pi_i},$$

where  $f_k(x)$  is the class-conditional density of  $X$ ,  $x$  is a particular instance of the values in a gene expression profile,  $\pi_k$  is the prior probability of class  $k$ .

4. The class-conditional density can be modeled as a multivariate Gaussian:

$$f_k(x) = \frac{1}{2\pi^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^{-1}\Sigma_k^{-1}(x-\mu_k)},$$

where  $\mu_k$  and  $\Sigma_k$  is mean and covariance respectively.

5. Then, linear discriminant analysis (LDA) is applied based on the assumption that the covariance matrix is the same for each class ( $\Sigma_k = \Sigma \forall k$ ). The log-ratio of class posteriors  $P(G|X)$ , provides a measure of the relative likelihood of classifying to those classes. Hence, the log ratio of classifying to classes  $k$  and  $l$  is formulated as:

$$\log \frac{Pr(G = k|X = x)}{Pr(G = l|X = x)} = \log \frac{\pi_k}{\pi_l} - \frac{1}{2}\gamma^T \Sigma^{-1} \gamma + x^T \Sigma^{-1} \gamma,$$

where,  $\gamma$  is  $(\mu_k - \mu_l)$ ,  $\pi_k$  is the class mean, and it is assumed that both classes have the same covariance matrix,  $\Sigma$  ( $\Sigma_k = \Sigma \forall k$ ),  $\mu_k = \sum_{g_i=k} \frac{x_i}{N_k}$ ,  $\Sigma = \sum_{k=1}^K \sum_{g_i=k} (x_i - \mu_k)^T / (N - K)$ , where  $x_i$  is a row from the data matrix  $X$ .

6. Finally, the orientation of the separating hyper-plane (between classes  $k$  and  $l$ ) is defined by the normal  $p$ -vector, in the third term on the right hand side, that we label  $b$ ,

$$b = \Sigma^{-1}(\mu_k - \mu_l)$$

The Characteristic Direction method is significantly more sensitive than existing methods for identifying DEGs. In our methodology, the chdir value is used as weight for the corresponding node and the final pathway graph is weighted with respect to edges with the mean chdir value of the corresponding gene values. This weight promotes the interconnecting nodes with high differential expression.

Subsequently, in order to extract perturbed subpathways from pathways, we use some graph theoretical properties to determine the densely connected neighborhood of a node. Let  $G = (V, E)$  a weighted directed graph, where  $V$  is the node set and  $E$  the edge set, with  $w_{vu}$  denoting the edge weight from node  $v$  to node  $u$ . With  $N(v)$  we represent the neighbors of node  $v$ . For a subgraph  $S \subseteq G$ , the internal degree  $N_{INT}(v, S)$  of a node  $v \in S$  is defined as the number of edges connecting  $v$  with nodes within  $S$  and the external degree as the number of edges connecting  $v$  with nodes not

belonging to  $S$ . The weighted internal degree is defined as the sum of weights of internal edges divided by internal degree:

$$NW_{INT}(v, S) = \frac{1}{N_{INT}(v, S)} \sum_{u \in N(v) \cap S} w_{vu}$$

Similarly, we define external weighted degree. The density of a graph is defined as the number of edges divided by the number of all possible edges. The weighted density of a (sub)graph is defined as the sum of all edge weights over the number of all possible edges:

$$DW(G) = \frac{1}{|V|(|V|-1)} \sum_{(v, u) \in E} w_{vu}$$

The algorithm operates on two phases, firstly the node set is expanded by selecting some of the external neighbors and secondly the selected node set is pruned. Initially, we start with a set  $S$  including only the NoI node  $s$ . Then, for each NoI's neighbor  $v \in N(s)$ , we compute the internal and external unweighted and weighted degree. In order to select a highly connected subset, a node  $v$  is included in the set  $S$ , if it satisfies the following two criteria:

$$\text{Criterion1} : \frac{N_{INT}(v, S)}{N_{INT}(v, S) + N_{EXT}(v, S)} > \alpha$$

$$\text{Criterion2} : NW_{INT}(v, S) > NW_{EXT}(v, S)$$

where  $\alpha$  is a parameter set for direct neighbors of NoI and for other nodes. After a fine tuning with repetitive trials, the optimal parameter value of  $\alpha$  was set to 0.55 and 0.85 respectively.

In the second phase we aim to obtain a more compact set by maximizing the weighted density. For this, we remove one by one nodes until we reach to a maximum value. The order of nodes is determined by the magnitude of the first criterion, with the less significant nodes examined first for removal. The algorithm is iterated in terms of the external neighbors of the selected nodes, until no more nodes can be added to the set  $S$ .

---

### Algorithm 1. Pseudocode of PerSubs Algorithm

---

Input: NoI,  $G$ ,  $\alpha_1$ ,  $\alpha_2$

Output: final subpathway  $S$

I.  $S = \{\text{NoI}\}$  // initialize

II. For each  $v$  in  $S$  // inclusion step

a. Find neighbors  $N(v)$

b. Keep not included neighbors:  $N(v) = N(v) - S$

c. For every  $u$  in  $N(v)$

i. Calculate  $N_{INT}$ ,  $N_{EXT}$ ,  $NW_{INT}$ ,  $NW_{EXT}$

ii. If  $u \in N(\text{NoI})$

1. Evaluate if Criterion1  $> \alpha_1$

iii. Else

1. Evaluate if Criterion1  $> \alpha_2$

iv. Evaluate Criterion2

v. if Criterion1 = true AND Criterion2 = true

1. Include  $u$ :  $S = S \cup u$

III. For each  $v$  in  $S$  ordered by increasing Criterion1 // pruning step

a. if  $DW(S - v) > DW(S)$

i. Remove  $v$ :  $S = S - v$

IV. Repeat steps II and III until no new nodes added

---

The output of PerSubs is a list of subpathways that can serve as potential network biomarkers for the case under study. Further, we evaluate statistically the resulted subpathways in order to keep the most reliable ones based on a permutation strategy. The gene labels in the RNA-Seq dataset are randomly shuffled 1000 times and each time PerSubs is re-applied. The subpathways starting from the same gene are compared based on their average weight. For each subpathway, the  $p$ -value is the percentage of cases where the average weight is lower than the respective value in the real condition ( $p$ -value < 0.05).

### 3. Results

We applied PerSubs on mouse microarray data [18] that explore the extent of lung regeneration following catastrophic damage after infection with H1N1 virus. In particular, the experiment contains samples from 3 colonies from control and H1N1 influenza treated lungs (12 days post infection). The complete dataset is deposited in NCBI's Gene Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) and is accessible through the accession number GSE32600. By applying PerSubs, we detected subpathways (Figure 1) which contain both differentially expressed and co-expressed associated genes, as to their expression change between control and infection state. All non-metabolic pathway maps of *Mus musculus* (mmu) were downloaded from KEGG [21] and were converted to gene-gene networks based on the CHRONOS R Bioconductor package [11].

Influenza infection in the lungs causes severe inflammatory damage to the lung through a respiratory outbreak of the innate immune response and the resulting lung injury can lead to other complications or chronic damage if not treated [22]. Zooming into H1N1 influenza A strain, it has been shown to induce acute respiratory distress syndrome (ARDS), pneumonia, alveolar damage, hypoxemia, and massive increase in inflammatory cytokines [18]. Influenza is a very common respiratory pathogen and as such it has been extensively studied to reveal its infection kinetics and pathogenicity [18]. Comprehending the influenza infection phases and especially repair stage will be an enabling step towards preventing these complications by assisting the lung to recover properly [22].

In this work we explore the (sub)pathways perturbed after H1N1 viral infection of mouse lungs at a specific time point (12 days post infection (dpi)). In the original work of [18], the tissue damage based on immune cell infiltration displayed peak at 11 dpi, declined at 21 dpi and mostly cleared in the lung at 60 dpi. Also, in the interval 10–12 dpi the weight loss of animals reached a peak and recovered at 20 dpi. In this work we first identified a set of differentially expressed genes (DEGs) between control and infected samples and then applied PerSubs with each DEG as starting point to detect the perturbed sub-topologies. In Table 1, we present some representative identified KEGG pathway terms. The pathway “ECM-receptor interaction” was found significantly enriched in two Influenza A related studies [23,24]. It has been reported that cellular processes such as adhesion, dynamic behaviors and apoptosis, regulated by ECM-receptor interaction, influence the entry or replication of influenza viruses [23]. Regarding “TGF- $\beta$  signaling”, it has been shown that respiratory viral infections offset secretion of TGF- $\beta$  which in turn is implicated in decreasing pulmonary inflammation and extending host survival [25,26]. Also, TGF- $\beta$  is involved in tissue repair and respiratory tract re-modeling of by stimulating matrix protein production, epithelial proliferation and differentiation. Moving forward, “Cytokine-cytokine receptor interaction” pathway has been shown to participate into activating the immune and inflammatory response to prevent from virus infections [24]. Moreover, “PPAR signaling” and “complement and coagulation” cascades have been suggested to repair excessive tissue damage by exhibiting anti-inflammatory functions [27]. Finally, with respect to “leukocyte transendothelial migration”, it has been suggested that circulating blood leukocytes migrate to tissue injury and infection site to terminate the primary inflammatory trigger and thus assist tissue repair [28,29].



3. Khatri, P.; Sirota, M.; Butte, A.J. Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.* **2012**, *8*, e1002375.
4. Jin, L.; et al. Pathway-based analysis tools for complex diseases: A review. *Genom. Proteom. Bioinform.* **2014**, *12*, 210–220.
5. Wang, K.; Li, M.; Bucan, M. Pathway-based approaches for analysis of genomewide association studies. *Am. J. Hum. Genet.* **2007**, *81*, 1278–1283.
6. Shi, J.; Walker, M.G. Gene set enrichment analysis (GSEA) for interpreting gene expression profiles. *Curr. Bioinform.* **2007**, *2*, 133–137.
7. Tarca, A.L.; Draghici, S.; Khatri, P.; Hassan, S.S.; Mittal, P.; Kim, J.S.; Kim, C.J.; Kusanovic, J.P.; Romero, R. A novel signaling pathway impact analysis. *Bioinformatics* **2009**, *25*, 75–82.
8. Rahnenfuhrer, J.; Domingues, F.S.; Maydt, J.; Lengauer, T. Calculating the statistical significance of changes in pathway activity from gene expression data. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, 1055.
9. Chen, X.; Xu, J.; Huang, B.; Li, J.; Wu, X.; Ma, L.; Jia, X.; Bian, X.; Tan, F.; Liu, L.; Chen, S.; Li, X. A sub-pathway-based approach for identifying drug response principal network. *Bioinformatics* **2011**, *27*, 649–654.
10. Judeh, T.; Johnson, C.; Kumar, A.; Zhu, D. TEAK: Topology enrichment analysis framework for detecting activated biological subpathways. *Nucleic Acids Res.* **2013**, *41*, 1425–1437.
11. Vrahatis, A.G.; Dimitrakopoulou, K.; Balomenos, P.; Tsakalidis, A.K.; Bezerianos, A. CHRONOS: A time-varying method for microRNA-mediated sub-pathway enrichment analysis. *Bioinformatics* **2016**, *32*, 884–892.
12. Vrahatis, A.G.; Balomenos, P.; Tsakalidis, A.K.; Bezerianos, A. DEsubs: An R package for flexible identification of differentially expressed subpathways using RNA-seq experiments. *Bioinformatics* **2016**, *32*, 3844–3846.
13. Dimitrakopoulos, G.N.; Balomenos, P.; Vrahatis, A.G.; Sgarbas, K.; Bezerianos, A. Identifying disease network perturbations through regression on gene expression and pathway topology analysis. In Proceedings of the 2016 IEEE 38th Annual International Conference of the Engineering in Medicine and Biology Society (EMBC), Lake Buena Vista (Orlando), FL, USA, 17–20 August 2016; pp. 5969–5972.
14. Vrahatis, A.G.; Dimitrakopoulos, G.N.; Tsakalidis, A.K.; Bezerianos, A. Identifying miRNA-mediated signaling subpathways by integrating paired miRNA/mRNA expression data with pathway topology. In Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Milan, Italy, 25–29 August 2015; pp. 3997–4000.
15. Nam, S.; Chang, H.R.; Kim, K.T.; Kook, M.C.; Hong, D.; Kwon, C.H.; Jung, H.R.; Park, H.S.; Powis, G.; Liang, H.; Park, T.; Kim, Y.H. PATHOME: An algorithm for accurately detecting differentially expressed subpathways. *Oncogene* **2014**, *33*, 4941–4951.
16. Li, C.; Li, X.; Miao, Y.; Wang, Q.; Jiang, W.; Xu, C.; Li, J.; Han, J.; Zhang, F.; Gong, B.; Xu, L. SubpathwayMiner: A software package for flexible identification of pathways. *Nucleic Acids Res.* **2009**, *37*, e131.
17. Vrahatis, A.G.; Rapti, A.; Sioutas, S.; Tsakalidis, A.K. PerSubs: A graph-based algorithm for the identification of perturbed subpathways caused by complex diseases. In Proceedings of the Genetics, Geriatrics and Neurodegenerative Diseases Research, Sparta, Greece, 20–23 October 2016.
18. Kumar, P.A.; Hu, Y.; Yamamoto, Y.; Hoe, N.B.; Wei, T.S.; Mu, D.; Sun, Y.; Joo, L.S.; Dagher, R.; Zielonka, E.M.; et al. Distal airway stem cells yield alveoli in vitro and during lung regeneration following H1N1 influenza infection. *Cell* **2011**, *147*, 525–538.
19. Maraziotis, I.A.; Dimitrakopoulou, K.; Bezerianos, A. Growing functional modules from a seed protein via integration of protein interaction and gene expression data. *BMC Bioinform.* **2007**, *8*, 408.
20. Clark, N.R.; Hu, K.S.; Feldmann, A.S.; Kou, Y.; Chen, E.Y.; Duan, Q.; Ma’ayan, A. The characteristic direction: A geometrical approach to identify differentially expressed genes. *BMC Bioinform.* **2014**, *15*, 79.
21. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30.
22. Tan, K.S.; Choi, H.; Jiang, X.; Yin, L.; Seet, J.E.; Patzel, V.; Engelward, B.P.; Chow, V.T. Micro-RNAs in regenerating lungs: An integrative systems biology analysis of murine influenza pneumonia. *BMC Genom.* **2014**, *15*, 587.
23. Chen, Y.; Zhou, J.; Cheng, Z.; Yang, S.; Chu, H.; Fan, Y.; Li, C.; Wong, B.H.; Zheng, S.; Zhu, Y.; et al. Functional variants regulating LGALS1 (Galectin 1) expression affect human susceptibility to influenza A (H7N9). *Sci. Rep.* **2015**, *5*, 8517.

24. Li, Y.; Zhou, H.; Wen, Z.; Wu, S.; Huang, C.; Jia, G.; Chen, H.; Jin, M. Transcription analysis on response of swine lung to H1N1 swine influenza virus. *BMC Genom.* **2011**, *12*, 398.
25. Furuya, Y.; Furuya, A.K.; Roberts, S.; Sanfilippo, A.M.; Salmon, S.L.; Metzger, D.W. Prevention of Influenza Virus-Induced Immunopathology by TGF- $\beta$  Produced during Allergic Asthma. *PLoS Pathog.* **2015**, *11*, e1005180.
26. Carlson, C.M.; Turpin, E.A.; Moser, L.A.; O'Brien, K.B.; Cline, T.D.; Jones, J.C.; Tumpey, T.M.; Katz, J.M.; Kelley, L.A.; Gauldie, J.; Schultz-Cherr, S.; et al. Schultz-Cherry, S. Transforming growth factor- $\beta$ : Activation by neuraminidase and role in highly pathogenic H5N1 influenza pathogenesis. *PLoS Pathog.* **2010**, *6*, e1001136.
27. Croasdell, A.; Duffney, P.F.; Kim, N.; Lacy, S.H.; Sime, P.J.; Phipps, R.P. PPAR $\gamma$  and the Innate Immune System Mediate the Resolution of Inflammation. *PPAR Res.* **2015**, *2015*, 549691.
28. Pociask, D.A.; Scheller, E.V.; Mandalapu, S.; McHugh, K.J.; Enelow, R.I.; Fattman, C.L.; Kolls, J.K.; Alcorn, J.F. IL-22 is essential for lung epithelial repair following influenza infection. *Am. J. Pathol.* **2013**, *182*, 1286–1296.
29. Nourshargh, S.; Alon, R. Leukocyte migration into inflamed tissues. *Immunity* **2014**, *41*, 694–707.



© 2017 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).