WestVirginiaUniversity
THE RESEARCH REPOSITORY @ WVU

Graduate Theses, Dissertations, and Problem Reports

2015

A novel computational system for identification of biological processes from multi-dimensional high-throughput genomic data

Julian Marshall Dymacek

Follow this and additional works at: https://researchrepository.wvu.edu/etd

Recommended Citation

Dymacek, Julian Marshall, "A novel computational system for identification of biological processes from multi-dimensional high-throughput genomic data" (2015). *Graduate Theses, Dissertations, and Problem Reports*. 5524.
https://researchrepository.wvu.edu/etd/5524

This Dissertation is protected by copyright and/or related rights. It has been brought to you by the The Research Repository @ WVU with permission from the rights-holder(s). You are free to use this Dissertation in any way that is permitted by the copyright and related rights legislation that applies to your use. For other uses you must obtain permission from the rights-holder(s) directly, unless additional rights are indicated by a Creative Commons license in the record and/ or on the work itself. This Dissertation has been accepted for inclusion in WVU Graduate Theses, Dissertations, and Problem Reports collection by an authorized administrator of The Research Repository @ WVU. For more information, please contact researchrepository@mail.wvu.edu.

A Novel Computational System for Identification of Biological Processes from Multi-dimensional High-throughput Genomic Data

Julian Marshall Dymacek

Dissertation submitted to the
Benjamin M. Statler College of Engineering and Mineral Resources
at West Virginia University
in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in
Computer Science

Lan Guo , Ph.D., Chair
Donald Adjeroh, Ph.D.
Guodong Guo, Ph.D.
Katerina Goseva-Popstojanova, Ph.D.
Michael Andrew, Ph.D.

Lane Department of Computer Science and Electrical Engineering
Morgantown, West Virginia
2015

Keywords: time series, dose response, high throughput, non-negative matrix factorization

Abstract

A Novel Computational System for Identification of Biological Processes from Multi-dimensional High-throughput Genomic Data

Julian Marshall Dymacek

Identifying potential toxicity signaling pathways could guide future animal studies and support human risk assessment and intervention efforts. This thesis describes a novel computational approach for identifying biological processes and pathways that are significantly associated with a disease pathology from time series, dose response, gene expression data.

Our system employs a novel constrained non-negative matrix factorization algorithm and Monte Carlo Markov chain simulation to identify underlying patterns in mRNA gene expression data. Quantitative pathology can be used as a pattern constraint. The found patterns can be thought of as functions that influence a gene's expression. Using a database of curated gene sets, we can identify biological processes that are significantly related to a pathology.

We also developed a computational model for integrating miRNA with mRNA time series microarray data along with disease pathology. The dynamic temporal regulatory effects of miRNA are not well known and a single miRNA may regulate many mRNA. The integrated analysis includes identifying both mRNA and miRNA that are significantly similar to the quantitative pathology. Potential regulatory miRNA/mRNA target pairs are then identified through databases of both predicted and validated pairs. Finally, potential target pairs are filtered, keeping only pairs that demonstrate regulatory effects in the expression data.

Multi-walled carbon nanotubes (MWCNT) are known for their transient inflammatory and progressive fibrotic pulmonary effects; however, the mechanisms underlying these pathologies are unknown. In this thesis, we used time series microarray data of global lung mRNA and miRNA expression isolated from 160 C57BL/6J mice exposed by pharyngeal aspiration to vehicle or 10, 20, 40, or 80 µg MWCNT at 1, 7, 28, or 56 days post-exposure. Quantitative pathology patterns of MWCNT-induced inflammation (bronchoalveolar lavage score) and fibrosis (Sirius Red staining, quantitative morphometric analysis) were obtained from separate studies.

Understanding the regulatory networks between mRNA and miRNA in different stages would be beneficial for understanding the complex path of disease development. These identified genes and pathways may be useful for determining biomarkers of MWCNT-induced lung inflammation and fibrosis for early detection of disease. Our computational approach detects biologically relevant processes with and without pathology information. The identified significant processes and genes are supported by evidence in the literature and with biological validation.

Acknowledgements

I would like to thank my advisor, Dr. Nancy Lan Guo, for her endless patience with me and her wealth of knowledge, wisdom, and time. I am grateful for the Lane Department of Computer Science and Electrical Engineering and the Mary Babb Randolph Cancer Center for providing for me a place to develop as a researcher and teacher.

I would also like to thank my committee members for their guidance through this process: Dr. Donald Adjeroh, Dr. Katerina Goseva-Popstojanovah, Dr. Guodong Guo, and Dr. Michael Andrews.

I am beholden to the members of the Guo Lab for their continued help and friendship: Joseph Putila, DaJie Luo, Rebecca Raese, Maricica Pacurari, Mehdi Iranmanesh, Michael Jude, Chunlin Dong and Ying-Wooi Wan. In particular, Dr. Brandi Snyder-Talkington has been a great colleague and I thank her for hours of support and biological expertise.

Lastly, I appreciate Drs. Michael Ochs and James Denvir for their input on the methodology, and Drs. Peter Gannet and Diandra Leslie-Pelecky for their training through the IGERT program. At NIOSH, I would like to thank Drs. Vincent Castranova, Dale Porter, Robert Mercer, and Yong Qian.

This work was supported in part by the West Virginia University NSF IGERT Research in Nanotoxicology training program for Julian Dymacek and in part by a grant from the West Virginia Graduate Student Fellowships in Science, Technology, Engineering and Math (STEM) program to Julian Dymacek. It was also funded by NIH grants R01 LM009500 (PI: Guo) and P20RR16440 and Stimulus supplement (PD: Guo).

This dissertation would not have been possible without the many people who have loved, cared, and supported me through these years: my friends George Willis and Steve Knudsen, with whom I have shared so many conversations; Scott and Ivy, who have given me constant support; and my parents, whose unending love, guidance, and belief have spurred me onward.

I dedicate this dissertation to my wife, Amanda.
"Weeping may tarry for the night, but joy comes with the morning."

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Due to limited time and money, biological research has traditionally utilized a compartmental approach, focusing on a few targeted genes. These genes may have been identified from literature, previous experience, or a hypothesis. As genome-wide analysis has matured, data can be collected from tens of thousands of genes simultaneously. This makes identifying genes and biological functions for closer study akin to finding a needle in a haystack. These identification problems are well suited for computational approaches and require the application of computer algorithms.

Nanotechnology is an emerging discipline in both industrial and medical fields. While various nanoparticles have been incorporated into diverse applications, the use of multi-walled carbon nanotubes (MWCNT) for both industrial and medical purposes is a quickly growing trend. The high surface area/mass and low density of MWCNT makes them easily aerosolized, thus a potential inhalation hazard during synthesis, product use and disposal. MWCNT have been widely used for various industrial applications [92] and exposure has been found to cause rapid onset lung inflammation, fibrosis and toxicity in treated mice [96, 83]. However, molecular mechanisms underlying MWCNT-induced pathogenesis are unknown.

Identifying biologically relevant processes and functions from time series dose-response toxicogenomics data is important to reveal toxicity mechanisms and assist in mechanistic studies by finding significantly changing genes [1, 52]. While microarray data is noisy and protein levels do not necessarily correspond to mRNA gene expression levels [47], our system finds novel hypotheses about involvement of diseases, processes, and functions from time series dose response microarray data.

Our system for identifying relevant pathways from time series microarray data includes several features. First, an analysis should be able to incorporate prior biological knowledge such as known pathways and relationships between genes. Second, the system should be able generate hypotheses about potential gene interactions. Finally, the system should be able to incorporate phenotypical data.

Analyzing time series microarray expression is a difficult task, as many time series experiments have few time points and are usually noisy [10]. Temporal information available from time series data is useful for discovering functional mechanisms and causal relationships. Many techniques have been developed to analyze time series microarray data [5, 74, 1]. Likewise, microarray data from multiple doses may reveal potential changes in toxicity and functional mechanisms.

Previously, benchmark dose (BMD)[39] methods have been used to identify a dose range for response in toxicity tests. By combining BMD with both Gene Ontology annotations and mRNA expression data, dose range estimates for the response of a biological processes can be found [119, 120, 121] Different models can be used, including power, linear, second degree polynomial, and third degree polynomial; with the best fit model being used to calculate the safe dose range. For a given biological process the response was calculated as the median BMD of the genes in the process. Additional parametric dose response models could be used including Gaussian, quadratic, and sigmoid [16, 70]. None of these studies included the use of histopathological data.

Clustering has widely been used for microarray analysis and many techniques have been tried [5, 4]. Clustering techniques group coexpressed genes based on a distance metric. Traditional distance based clustering methods can be extended by creating new metrics that include biological function information [57] and can use gene annotations [114] from the Gene Ontology annotation tree. However, genes are still placed into a single coexpression group even though the gene's expression might be influenced by multiple processes.

Non-negative matrix factorization(NMF) is a technique developed by Lee and Seung [71] that identifies underlying basis patterns used to reconstruct the original data. NMF has previously been used in microarray analysis [28], however prior information has not been used in generating the patterns. Like NMF, Bayesian Decomposition (BD) [88] attempts to find patterns and coefficients that can reconstruct the original data. However, BD allows for prior biological information to be encoded and influence the patterns found. BD has also been used to identify potentially activated pathways in drug treated time-series microarray data [90] but is not computationally efficient on the genome wide scale.

While dynamic temporal regulatory effects of microRNA (miRNA) are not well known, a single miRNA helps regulate many mRNA and therefore act on a multitude of proteins [8, 78]. The expression of over half of the genes in the human genome may be regulated by miRNA [41]. Compared to mRNA, miRNA are more stable and can be isolated from a wide variety of clinical samples while still being measured by microarray analysis and real-time PCR. These properties make miRNA useful as potential biomarkers [86].

Unfortunately, miRNA analysis is relatively new and curated annotation databases are still being created. Integration of well studied mRNA and regulatory miRNA provide a powerful analysis technique. Traditional integrated methods such as negative

correlation or simple up- or down-regulation may produce more potential target pairs especially when using predicted miRNA/mRNA pairs [19, 32, 60, 128]. Unfortunately, these additional pairs may have no relevance to the known disease pathology. Additionally, negative correlation rewards pairs that are consistently in opposite directions. The complex relationship between miRNA and mRNA is not well understood. A single miRNA may regulate many mRNA and not consistently negatively correlate with any single mRNA. An integrated system should allow for miRNA/mRNA pairs that demonstrate a negative relationship over a subset of the time points.

When dealing with exposure, response time is an important factor. Previously, non-negative matrix factorization algorithms have been applied to integrated analysis [139, 140] but did not include quantitative pathology data or focus on time series data. Some new systems are being developed to identify regulator networks from time series data [102]. These techniques do not focus on analyzing a specific pathology. Pathologies, such as exposure response, suggest that miRNA levels should be consistently changing with the pathology especially if regulating a consistent mRNA response. Our integrated method allows both divergence between miRNA/mRNA at individual time points and consistence with the pathology.

This document is organized as follows: Chapter 2 is a discussion of related work on identifying gene sets from time-series dose-response microarray data; Chapter 3 describes our methodology, the MEGPath system, and results from the application of our system on mRNA data; Chapter 4 includes an extension of the MEGPath system for integrating time series dose response miRNA with mRNA; Chapter 5 a description of *in vitro* validation and an Ingenuity Pathway Analysis evaluation of the integrated analysis; finally, Chapter 6 discusses the contributions of our research and potential future works.

Chapter 2

Related Work

Microarrays are a standard technique for measuring genome wide expression values. Many microarray experiments involve a static snapshot of the genome for a specific condition with multiple conditions being compared such as a normal sample versus a cancer sample. Time series experiments involve static samples at various intervals. These intervals help analyze the temporal process of gene regulation [10] and reveal underlying functional mechanisms. Various techniques have been used to analyze time series microarray data [74, 1].

A system for identifying relevant pathways from time series dose response microarray data needs to include several features. First, the analysis system should be able to incorporate prior biological knowledge linking genes to their functions. Second, the system should allow genes to be influenced by multiple underlying functions. Finally, the system should be able to incorporate phenotypical or pathological data.

This chapter provides a review of the methods and tools related to our study. An overview of methods for identifying functions and processes is in Section 2.1. Section 2.2 discusses distance-based clustering methods. A few additional techniques are discussed in Section 2.3 . A discussion of matrix decomposition methods is in

Section 2.4. Finally, Section 2.5 gives a summary of the chapter.

2.1 Pathway Analysis

Identifying the biological processes and functions related to a set of genes requires semantic information linking genes to their functions. This semantic information is often in the form of gene annotations. There are several databases of gene annotations including the Gene Ontology (GO) project [6], the Kyoto Encyclopedia of Genes and Genomes (KEGG) [64], and the Reactome project [25]. The Molecular Signatures Database (MSigDB) [76] contains curated sets of genes designed specifically for enrichment analysis and includes large portions of other databases including KEGG, GO, and Reactome.

Techniques for taking a set of genes and identifying significant annotations are called enrichment analysis techniques. Huang [58] surveyed 68 enrichment analysis techniques and identified three general classes: singular enrichment analysis, modular enrichment analysis, and gene set enrichment analysis. Khatri [66] uses slightly different nomenclature but identifies similar classes.

The class of singular enrichment techniques identifies significant genes and then the enriched annotation terms. GoMiner [138] is an example of this technique. Genes with significantly changing expression are identified, usually using $p < 0.05$ and a fold change greater than 1.5. Each annotation term is then checked for significance using Fisher's exact test by comparing the number of significant genes involved with the annotation term versus the number of significant genes in the genome. These techniques often lead to an over abundance of significant genes and terms.

Modular enrichment techniques incorporate singular enrichment analysis techniques while utilizing the relationships between annotation terms. An example of a

modular enrichment technique is the DAVID algorithm [59]. DAVID works by creating pairwise *kappa* statistics of gene-gene and term-term interactions. Cohen's *kappa* statistic is a measure for agreement of categorical terms while adjusting for chance. The categorical annotation terms come from the DAVID Knowledgebase [105] which provides an integration of various gene annotation databases while also handling various gene IDs.

Gene Set Enrichment Analysis (GSEA) [111] is a technique for identifying biological functions and pathways from both expression data and annotation data. GSEA works on the expression data for the genome and does not require identifying a set of interesting genes instead identifying a subset of genes which are over represented in a process or function. Through label permutation GSEA can identify processes that significantly differ over conditions. The standard GSEA algorithm is not well suited for time series data as the time points must be treated as separate conditions.

2.2 Point-Wise Distance-Based Clustering

Clustering has been widely used for microarray analysis for many years with multiple techniques having been tried [5, 4]. Point-wise distance-based clustering methods are used to describe the distance between two genes. Thresholds can be set to group similar genes into clusters. Many different distance metrics have been used, some of which are listed in Table 2.1. Some distance methods, such as linear correlation [95], do not work particularly well for time-series microarray data.

Once a distance metric has been decided upon, a clustering algorithm must be used. The traditional clustering algorithm that has been used for analyzing microarray data is *k-means* analysis [115]. In *k-means*, genes are partitioned into k clusters with each gene belonging to the cluster with the closest mean. Lloyd's algorithm [80]

Table 2.1: Point-wise distance metrics. Let E be a matrix where $E(i,t)$ is the expression value of gene i at time point t. Let there be N_g total genes and N_t total time points.

L1-norm	$d_{ij} = \sum_{t=1}^{N_t} \lvert E(i,t) - E(j,t) \rvert$
L2-norm	$d_{ij} = \sqrt{\sum_{t=1}^{N_t} (E(i,t) - E(j,t))^2}$
Mahalanoblis	$d_{ij}^2 = (E_i - E_j)^t \sum^{-1} (E_i - E_j)$
Correlation	$d_{ij} = 1 - r_{ij}$ $r_{ij} = \frac{\sum_{t=1}^{N_t} (E(i,t) - \bar{E}(i))((E(j,t) - \bar{E}(j))}{\sqrt{\sum_{t=1}^{N_t} (E(i,t) - \bar{E}(i))^2 \cdot \sum_{t=1}^{N_t} (E(j,t) - \bar{E}(j))^2}}$

is the standard iterative approach for implementation. In Lloyd's algorithm, genes are iteratively assigned to the closest cluster and then the cluster means are recalculated. A related algorithm is the *k-median* algorithm where a gene is assigned to the cluster with the closest median. Extensions to *k-means* have been explored, such as incorporating the temporal dimension of time-series data [110]. Point-wise techniques work with time-series data but do not allow genes to be influenced by multiple underlying functions.

2.2.1 Self-organizing Maps

Another clustering algorithm that has been used is called self-organizing maps (SOMs) [113]. A SOM is a set of interconnected nodes and a distance function on the nodes. The map nodes are initially placed at random, and then are iteratively adjusted. Genes are added to the SOM in a random order, with the distance between each gene and each map node being calculated. Then, the closest map node N is moved the most towards the gene. Other nodes are moved towards the gene based on the map distance to N. Recently SOMs have been applied to time series microarray data [17].

```
procedure CAST
    repeat
        Choose a gene not in a cluster and assign it to a cluster C
        repeat
            add to C unassigned genes
            remove and genes
        until C has converged
    until All genes are assigned to a cluster
end procedure

procedure PARTITIONING AROUND MEDOIDS
    assign each gene to the closest medoids
    for all medoid m do
        for all data point o do
            swap m and o and compute the total
        end for
    end for
end procedure

procedure K-MEANS
    repeat
        for all gene do
            associate gene with a cluster
        end for
        for all cluster do
            calculate the new centroid of the cluster
        end for
    until Convergence
end procedure
```

Figure 2.1: Clustering algorithms

Neither, annotation information or pathology constraints were used.

2.2.2 Biclustering

Biclustering [20] aims to cluster not only the rows (genes) but also the columns (conditions) of the microarray expression matrix. Biclustering algorithms typically work iteratively in two stages: a deletion stage and an addition stage. The deletion

stage removes rows or columns that do not decrease a fitness function. After deletion, previously excluded columns and rows may now lower the fitness function. After a cluster is found, the included conditions are replaced with random data. The *cc-Biclustering* algorithm [22] performs biclustering on genes divided by groups instead of columns. A group might consist of biological replicates or conditions on a given dose. In general biclustering is NP-hard [81].

One issue with traditional biclustering is that clusters are subsets of the data and potential temporal effects may be lost. For time-series data, restrictions could be made to keep contiguous columns [82]. However, expression values are often discretized into {DownRegulated, NoChange, and UpRegulated}, however pathology constraints would be difficult to add.

2.2.3 Fuzzy C-means

Fuzzy c-means clustering [27] is a clustering algorithm where each gene has a degree of belonging to a cluster. The centroid of a cluster is the average of all genes weighted by the gene's degree of belonging to that cluster. The degree of belonging allows genes to belong to multiple co-expression groups, defined by weights. Similar to k-means clustering, the initial cluster assignments strongly influence the final clusters. Prior biological knowledge can be incorporated [114] by deriving initial assignments from the Gene Ontology annotation tree. Additionally, the annotation information is utilized in distance calculation. These techniques allow for genes to be associated with multiple functions but pathology constraints could be difficult to incorporate.

2.2.4 Extensions to Distance Methods

Distance based clustering methods can be extended by creating new distance metrics based on biological function information. A shrinking distance metric is used by Huang [57] where a parameter r allows for distances to shrink if two genes share a biological annotation. The r parameter can range between 0 and 1 with only annotation information being used to no annotation information being used. As with the number of clusters, a value for r must be tuned to best describe the data.

The SICAGO system [65] utilizes a semantic distance between gene annotations in the GO project. The semantic distance measures used are based on how much information the genes share in common. Multiple information content theory metrics are included with the final distance incorporating the euclidean distance between expression. Unfortunately, SICAGO has no way of including known pathological information.

2.3 Additional Techniques

Additional techniques such as EPIG [23] and ASTRO [116] have been developed for finding patterns and co-expressed genes.

The EPIG[23] algorithm uses local clusters to help identify patterns for clustering. All gene profiles are initially considered as patterns. A pair-wise correlation is performed amongst all profiles, with highly correlated profiles being grouped together. All groups with fewer than six profiles are eliminated from consideration. Profiles are again filtered by a signal to noise ratio. Finally, patterns are formed from the averages of the profiles in the groups still remaining. The EPIG system eliminates the random behavior of many clustering algorithms and finds patterns which correlate closely with the data. The use of correlation limits a gene's ability to be influenced

by multiple functions.

ASTRO [116] is specifically designed to analyze time series expression data with limited time points. The algorithm converts the gene expression matrix to a rank matrix. A gene's expression is transformed from lowest expression to highest by ranks:

$$\{5, 15, 10, 20\} \Rightarrow \{1, 3, 2, 4\}\{0, 7, 6, 20\} \Rightarrow \{1, 3, 2, 4\}$$

The reduction of information enables ASTRO to have $O(nm)$ complexity were n is the number of genes and m is the number of samples. There is a loss of information and rankings lose information about relative differences between genes.

2.3.1 Model-based Clustering Methods

Model-based clustering methods differ from distance based methods. Model clustering is not performed on the similarity of gene expression but how closely the genes match underlying model functions. The model functions can be based on previously known distributions or prior biological information. Individual clusters of genes are assumed to match a unique underlying function [93, 46]. Model-based methods are easily extended to time-series data and are more robust in dealing with noisy data; however, they are limited to predetermined functions.

Hidden Markov models(HMM) have also been used to analyze time series microarray data. Unlike traditional clustering techniques, HMMs can be used to cluster genes [101] while including temporal information. HMMs can be described with the following parameters: the states S_i, π_i the probability of starting in state S_i, a_{ij} is the transition probability from state S_i to S_j, and $b_i(\omega)$ is the emission probability density of a symbol $\omega \in \Sigma$ in state S_i. To cluster the genes, k HMMs are found that maximize the likelihood. In order to find the parameters, an iterative algorithm can

be used where each gene is assigned to a HMM and then the parameters of each HMM are re-estimated using the genes assigned to it. Genes are limited to being described by only one function.

2.4 Matrix Decomposition Methods

2.4.1 Principal Component Analysis

Principal component analysis (PCA) is a well known data analysis technique and has been applied to gene expression data [33] and to time-series microarray data as well [136]. Given that X is a gene expression matrix, PCA finds the eigenvalues and eigenvectors of the covariance matrix. The covariance of X is $\frac{1}{n}XX^\top$ and the following decomposition is possible [106] with D being a diagonal matrix and W being the eigenvectors of X as columns.

$$XX^T = WDW^T$$

These eigenvectors form a set of orthonormal basis vectors representing underlying functions in the original data. By using only the k eigenvectors with largest eigenvalues, clustering can be performed.

Singular value decomposition (SVD) is a standard solution technique for PCA and has also been used in microarray analysis [2]. SVD works by factoring the original expression matrix X into three component matrices:

$$X = U\Sigma V^\top$$

The relationship between SVD and PCA can be seen as follows:

$$\begin{aligned} XX^T &= (U\Sigma V^T)(U\Sigma V^T)^T \\ &= ([U]\Sigma V^T)(V\Sigma U^T) \\ &= [U][\Sigma][I][\Sigma][U]^T \\ &= U\Sigma^2 U^T \end{aligned}$$

Implementations for SVD are included in many numerical packages including R, Matlab, and Numerical Recipes.

PCA and SVD remain standard techniques and identify underlying basis vectors without requiring additional parameters to tune. The basis vectors are required to be orthonormal, making pattern constraints impossible to add. There is also no way to steer the solution through prior domain knowledge.

2.4.2 Independent Component Analysis

Independent component analysis (ICA) is another technique that has been applied to time-series microarray data. Unlike PCA, the goal of ICA is to find a set of components which are as statistically independent of one another as possible [69, 33]. In signal processing, ICA has been used for blind source separation [63], the problem of identifying M sound signals from N microphones. Many biological signal processing problems have been cast as blind source separation including processing EEGs[126].

ICA has been applied to microarray data[42, 77, 72]. Unlike PCA the goal is to minimize the amount of mutual information [69], meaning that ICA finds functions that are as different as possible not functions that reconstruct the data. This constraint makes it difficult to incorporate pathological information or prior domain knowledge.

2.4.3 Network Component Analysis

Network component analysis (NCA) [75, 123] is a technique for identifying the bipartite connectivity between regulatory signals and output data. The network derived by NCA for microarray data, relates transcriptional factor activity (regulatory signals) to gene expression (output data). The relationship can be defined by the decomposition of a gene expression matrix G, consisting of m genes and n samples for each gene.

$$G_{m\times n} = CS_{m\times p} \times TF_{p\times n}$$

The connectivity matrix, CS, contains the strength of each edge. The matrix TF contains the regulatory nodes of the network. There are p transcriptional factor activities. NCA requires three additional constraints:

1. CS has full column rank
2. Each column of CS has at least $p-1$ zero values
3. TF must have full row rank

These constraints mean that G can be uniquely (up to a scaling factor) decomposed. The second constraint allows for a regulatory node and its related output nodes to be removed while still maintaining full column rank for CS.

2.4.4 Non-negative Matrix Factorization

Non-negative matrix factorization(NMF) is a technique developed by Lee and Seung [71] for use in facial recognition and text mining. NMF attempts to generate a set of basis vectors H and corresponding coefficients W that can be used to reconstruct

the original data. Given that X is the original gene expression:

$$[X]_{n\times m} \approx [W]_{n\times k} \times [H]_{k\times m}$$

There are n genes and m samples (time points). The variable k corresponds to the number of basis vectors and is usually set to the $m-1$. If $k = m$ then H becomes the trivial solution of the $m \times m$ identity matrix.

NMF algorithms impose the additional constraint that each entry in W, H, and V must be non-negative. This constraint creates an additive nature to the basis vectors and allows for intuitive functional meanings. Techniques such as PCA also find basis vectors, but allow negative values. Without the non-negative constraint the linear combinations of basis vectors in PCA can have complexities generated through cancellations making functional interpretation difficult. Unlike PCA, the basis vectors in NMF are not required to be orthonormal nor are they required to independent as in ICA.

NMF has been explored in microarray analysis, for a review of techniques see [28]. Algorithms that generate sparse basis vectors have also been explored [67] as sparse basis vectors identify specific functions particularly well.

A weakness of the NMF algorithms is the exclusion of prior information such as pathology data. The iterative techniques of most methods make additional constraint steps lead to numerical instability or loss of the central non-negative constraint. A limited amount of prior information (cell type) can be incorporated by identifying select marker genes and forcing their reconstruction to exactly one basis vector [45]. Similar techniques would be difficult to work for constraints on the basis vectors.

2.4.5 Bayesian Decomposition

A difficulty with traditional NMF techniques is that prior information cannot be used to help generate the patterns. Like NMF, Bayesian Decomposition (BD) [88] attempts to find patterns and coefficients that can reconstruct the original data. However, BD allows for prior biological information to be encoded and influence the patterns found. BD works as a Monte Carlo Markov Chain, which explores the state space of possible patterns and coefficients by updating probability density functions. By using Bayes' formula prior information can be used to guide the exploration. BD has also been used to identify potentially activated pathways in drug treated time-series microarray data [90]. Unfortunately, BD has proven to be computationally inefficient in general practice and provides no way of constraining an entire pattern to match pathology.

2.5 Summary

In this chapter we reviewed current methods and tools for analyzing time series, dose response microarray data. We have identified some requirements for our methodology and have identified that no existing system currently implements all of our requirements.

We propose a computational system, presented in Chapter 3, for identifying relevant biological functions and processes from time series, dose response microarray data. Our system incorporates prior biological knowledge linking genes to their functions, phenotypical or pathological data, and allows genes to be related to multiple functions.

Chapter 3

The MEGPath System

In this chapter, we present our methodology for analyzing time series dose response high throughput data. Our system for identifying relevant pathways from time series microarray data includes several features: first, the ability to incorporate prior biological knowledge such as known pathways and relationships between genes; second, the ability to generate hypotheses about potential gene interactions; finally, the ability to incorporate phenotypical data. As discussed in Chapter 2, there is a need for a system capable of combining all of these elements.

Identifying biologically relevant processes and functions from time series dose-response toxicogenomics data is important to reveal toxicity mechanisms and assist in mechanistic studies by finding significantly changing genes [1, 52]. While microarray data is noisy and protein levels do not necessarily correspond to mRNA gene expression levels [47], our system finds novel hypotheses about involvement of diseases, processes, and functions from time series, dose response microarray data. An overview of our system can be seen in Figure 3.1.

Clustering has widely been used for microarray analysis and many techniques have been tried [5, 4]. Clustering techniques group co-expressed genes based on a

distance metric. Traditional distance based clustering methods can be extended by creating new metrics which include biological function information [57] and can use gene annotations [114] from the Gene Ontology annotation tree. However, clustered genes are still placed into a single co-expression group even though a gene's expression might be influenced by multiple processes.

Non-negative matrix factorization(NMF) is a technique developed by Lee and Seung [71] that identifies underlying basis patterns used to reconstruct the original data. NMF has previously been used in microarray analysis [28] but prior information has not been used in generating the patterns. Like NMF, Bayesian Decomposition (BD) [88] attempts to find patterns and coefficients which can reconstruct the original data. However, BD allows for prior biological information to be encoded and influence the patterns found. BD has also been used to identify potentially activated pathways in drug treated time-series microarray data [90] but is not computationally efficient on the genome wide scale.

The MEGPath system identifies a set of non-negative patterns which represent underlying biological functions. The patterns are used to relate each gene to the underlying functions that direct its gene expression. The MEGPath system allows for constraints to be added on to the patterns such as representing a known pathology.

This chapter describes the MEGPath system and its application to mRNA data. Section 3.1 is a discussion of data sets and describes the four stages of our methodology. Section 3.2 is an algorithmic complexity analysis of the different parts of the system. System verification is covered in Section 3.3. Results of our system on mRNA data are described in Section 3.4 including the visualization and results with and without pathological information. Finally, Section 3.5 gives a discussion of the methodology and findings.

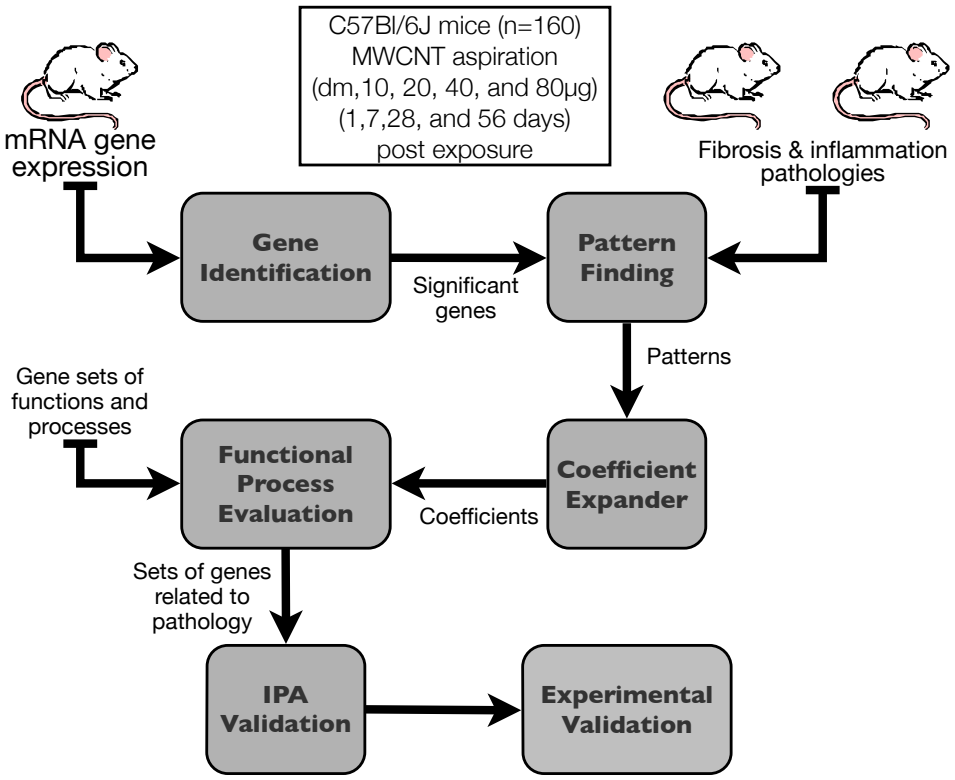

Figure 3.1: Overview of the four steps in the algorithm. Step 1: Identify significantly changing genes using SAM and linear model. Step 2: Find patterns and coefficients to reconstruct the gene expression data. Step 3: Find coefficients for the entire genome. Step 4: Using the patterns, coefficients, and pathways, we can identify significant pathways. The results can be verified using both Ingenuity Pathway Analysis and biological testing.

3.1 Materials and Methods

Our system, MEGPath, as shown in Figure 3.1, is divided into four main components: the Gene Identification component, the Pattern Finding component, the Coefficient Expander component, and finally the Functional Process Evaluation component. The MEGPath system is capable of generating novel hypotheses about genes by grouping genes to system identified patterns. In addition, the MEGPath system allows for constraints, such as a quantitative pathological data, to be added as prior information. The additive nature of the system enables genes to be associated with multiple patterns while maintaining understandable interactions. Finally, the Functional Process Evaluation step allows prior biological information, such as gene interactions and annotations, to be related to the expression data and tested for significance.

3.1.1 Data Sets

The data sets consisted of dose-dependent time series mRNA microarray expression data as well as quantitative pathology scores from an aspiration experiment. In total 480 mice were randomized into three groups: genome-wide mRNA expression, lung pathology scores, and fibrosis pathology scores. Scores were found for 1, 7, 28, and 56 days post-exposure with the mice exposed to 0, 10, 20, 40, or 80 µg of MWCNT.

Total RNA was extracted from the mice at each dose condition. Agilent Mouse Whole Genome Arrays were used for expression profiling. In total, our genome consisted of 41,059 probes and the data has been deposited to the NCBI Gene Expression Omnibus (GEO) repository using accession number GSE29042. The microarray data were log-transformed for analysis. The expression of each gene at each treatment condition can be visualized through a web-interface.[1]

[1] (http://www.mwcnttranscriptome.org)

In addition to mRNA expression data, 160 mice were used for inflammation scores. Inflammation scores (BAL polymorphonuclear leukocytes) were derived from the analysis of BAL fluid taken from the MWCNT exposed mice [96]. In total 4 mL of lavage fluid was collected and BAL cells were isolated by centrifugation. BAL cell counts were obtained using a Coulter Multisizer 3.

An additional 160 MWCNT exposed mice were used for fibrosis scores. Scores were found using morphometric analysis of Sirius Red staining for connective tissue [83]. Paraffin sections of the left lung were sliced and then rehydrated. After Sirius Red staining, the average thickness of connective tissue fibers in the alveolar region were obtained through quantitative morphometric methods.

3.1.2 Gene Identification

For each dose and time point, a set of differentially expressed genes were identified by performing a two-class unpaired Significance Analysis of Microarrays (SAM) [125] between the treated samples and the zero dose samples from the corresponding time point, using the Bioconductor package. A threshold delta value was chosen to produce a false discovery rate of 1% using the *findDelta* function from the same package. The list of significant probes was subsequently filtered, keeping probes that were at least 1.5 fold up- or down-regulated. Fold changes were computed from the data before imputation of missing values.

Additionally, a linear model was fit to the data, modeling the log expression of each gene as a function of time, dose, and the interaction of time with dose. We moderated the t-statistic associated with the dose and interaction parameters following the SAM algorithm and set a threshold to control for a false discovery rate of 0.1%. This generated a list of genes with expression values that were significantly dependent on

dose and a list of genes with expression values that were significantly dependent on dose in a time-dependent fashion.

3.1.3 Pattern Finding

The Pattern Finding algorithm (Algorithm 3.2) is a non-negative matrix factorization algorithm. Our algorithm identifies a set of non-orthogonal basis patterns (vectors) which can be linearly combined to reconstruct the original gene expression data. In addition to the patterns, the algorithm finds coefficients relating each gene to each pattern. The most important feature of the Pattern Finding algorithm is the ability for genes to be associated with multiple patterns.

Let D be the original data matrix, we wish to find the matrices P and C such that, $D = C \cdot P$. Gene expressions are normalized to be in the range $[0, 1)$ so that D contains the normalized fold change values.

The matrix P consists of the patterns that are used as the basis vectors. Each row contains one pattern as expressed across the experiment conditions. For example, a pattern is made of an expression value at each time point or an expression value for each dose across a time point. Patterns are the average response of similar genes and the patterns in P provide biologically relevant information.

The matrix C is the coefficient matrix, consisting of one row for each gene and one column for each pattern. The coefficients represent how strongly a gene is associated with a particular pattern.

The Pattern Finding algorithm works as a Monte Carlo Markov Chain. Each entry in the coefficient and pattern matrices has an associated probability density function (PDF). During each Monte Carlo Markov Chain step an entry is altered and the overall error is computed. If the change reduces the overall error, then the

```
procedure FindPatterns
    e ← TotalError(D, C, P)
    i ← 0
    while e > 0 and i < MaxIterations do
        for all v ∈ P and C do
            v ← Random(v's PDF)
            e1 ← TotalError(D, C, P)
            if e >= e1 then
                add r to the PDF
            end if
            e ← e1
            i ← i + 1
        end for
    end while
end procedure
```

Figure 3.2: The Pattern Finding algorithm.

PDF associated with the entry is updated. The pseudocode for the Pattern Finding algorithm is listed in Figure 3.2.

Prior information can be used to constrain a search. The constraints are formed by modifying an entry's PDF. The PDF of an entry could be constrained to always return the same value or pathological data could be encoded as a pattern constraint. For a pathology constraint, it is important that the pattern entries maintain their relative distances but be allowed to shift up or down as a group. This shifting can be accomplished with a single variable and an associated PDF.

After generating the PDFs, the Pattern Finding algorithm uses simulated annealing to minimize the overall error, using a standard annealing function [97]. The simulated annealing process utilizes each entrys's PDF and works in a similar manner to the original Monte Carlo Markov Chain.

3.1.4 Coefficient Expander

Coefficients are found using the Pattern Finding step. By using the Gene Identification step to identify significantly changing genes, not all genes will have coefficients. Genome-wide coefficients are insured by using the Coefficient Expander algorithm. The Coefficient Expander step attempts, through the use of simulated annealing, to find optimal coefficients for each gene in the genome. Each gene's coefficients should, when combined with the patterns, minimize the distance between the calculated expression and the actual expression.

3.1.5 Functional Pathway Evaluation

The final step is to calculate the Functional Process Evaluation (FPE) score for a given set of genes. The FPE score is based on the score from Gene Set Enrichment Analysis [111]. Each gene's coefficients are normalized to obtain the relative importance of a pattern to the gene. Genes that are not common to both the pathway and genome are ignored and not used in computing the score. If a gene has multiple probes, the probe with least reconstruction error is chosen.

The FPE score ranges from -1 to 1 with the scores being computed for each pattern. Scores are calculated by first sorting the genome by the pattern coefficient. After sorting, the genes with the highest coefficients for a pattern will be at the beginning. Scores are computed by starting at the highest ranking genes and either adding a value (for genes in the set) or subtracting a value (for genes not in the set). The highest score generated is used as the FPE score. Higher FPE scores correspond to sets of genes which are over-represented at the beginning. Leading sets are the genes that are used to generate the FPE score and are the genes considered most closely related to a pattern for the pathway. An overview of the FPE algorithm is

$p_{hit}, p_{miss}, score \leftarrow 0$
for all $g \in G$ **do**
  if $g \in P$ **then**
    $p_{hit} \leftarrow p_{hit} + C(g)/S$
  else
    $p_{miss} \leftarrow p_{miss} + 1/(N_g - N_p)$
  end if
  if $score < p_{hit} - p_{miss}$ **then**
    $score \leftarrow p_{hit} - p_{miss}$
  end if
end for
return $score$

Figure 3.3: The FPE score is based on the positive score from Gene Set Enrichment Analysis. Let the genome G contain N_g genes and P be gene set of N_p genes. Also let S be the sum of the coefficients of the genes in P and $C(g)$ be the coefficient of a gene g.

described in Figure 3.3.

A pathway's p-value is found by comparing its FPE score to the score of thousands of randomly generated gene sets of the same length and counting the number of random sets with higher scores. After p-values have been calculated for all the pathways, the Benjamini-Hochberg [11] method is used to adjust for multiple hypothesis testing. Gene sets with a p-value less than 0.05 are considered significant.

3.2 Algorithm Analysis

A complexity analysis of the MEGPath system involves looking at three components: Pattern Finding, Coefficient Expander, and the Functional Pathway Evaluation. The Gene Identification step, a single preprocess step utilizing the standard SAM algorithm, is not analyzed here.

3.2.1 Pattern Finding

The Pattern Finding algorithm tries to reduce the error difference between the original data D and the product of the the coefficient matrix C and pattern matrix P. Let n be the number of genes, m be the number of conditions, and $m-1$ be the number of patterns. A basic solution for finding the error involves a matrix multiplication, an entry by entry subtraction, and a sum of all the entries. A standard worst case complexity analysis of matrix multiplication involves counting the number of scalar multiplications [24]. The basic solution would have $n*m*(m-1)$ multiplications for each change to the P or C matrices. In total, there would be $n*m*(m-1)*(n*(m-1)+m*(m-1))$ multiplications, $O(n^2*m^3)$.

Now, we will consider our algorithm that limits the multiplications for each entry's update. First, note that when changing a gene's coefficient, only one row will change in the product of C and P. This one row will contain m values and will require $m*(m-1)$ multiplications. When changing an entry in P an entire column will be changed. The column change requires $n*(m-1)$ multiplications. The optimized error procedure takes $n*m*(m-1)*(1+(m-1))$ multiplications, $O(n*m^3)$. Our data, $n = 2,996$ and $m = 4$, is similar to most gene data where the number of genes is much larger than the number of conditions. The entry subtraction and sum can be reduced by caching entry error values and updating them during the multiplication process. The total number of iterations can vary but is independent of the number of genes or conditions.

3.2.2 Coefficient Expander

The Coefficient Expander algorithm attempts to find coefficients that minimize the error in a gene's reconstruction. Unlike the Pattern Finding algorithm, no changes

are needed to the pattern entries. The $m-1$ coefficients corresponding to the m patterns are optimized. Each optimization step requires $m*(m-1)$ multiplications. The total number of optimizations can vary but is independent of the number of genes or conditions.

3.2.3 Functional Pathway Evaluation

The Functional Pathway Evaluation algorithm attempts to find sets of genes significantly related to a pattern. The algorithm must compute the FPE score for each gene set and then compute FPE scores for 1,000 randomly generated sets. The FPE score calculation requires a one time sorting of the genome and then an iteration through the genome. Each gene is checked for inclusion in the set. If a hashtable is used to store the set, the algorithm will run in linear time according to the size of the genome.

3.3 Verification of the Algorithm

No "gold standard" data sets exist that incorporate all of the features used by the MEGPath system. Each of the components were individually evaluated to ensure that each component worked as intended.

3.3.1 Pattern Finding

The Pattern Finding component was evaluated to ensure that known patterns could be extracted from a generated data set. Expression values for three, randomly chosen probes at 40 µg across the four time points were used as the underlying patterns. These three patterns were then combined using randomly chosen weights to create 3,000 generated test expressions. By comparing the original patterns with the found

patterns, we can determine if MEGPath is capable of revealing known underlying patterns. A test set of 3,000 generated expressions were used with patterns found from 16 consecutive runs.

To check if patterns were better than random, three randomly selected probes were chosen as patterns and used for genome wide reconstruction. These probes were normalized and then the Coefficient Expander step was run to compute the total error. The results of this test suggest that the MEGPath system found patterns that are better than randomly chosen patterns.

3.3.2 Comparison with other Non-negative Matrix Factorization Algorithms

Since our system utilizes a NMF algorithm we tested our algorithm's ability to identify data reconstruction patterns against other NMF algorithms. Algorithms were used from the NMF package for R [44] that is designed for bioinformatics use and for algorithm comparison. The NMF package includes eight standard algorithms: Brunet, Frobenius, KL, Lee, Offset, SNMF/l, and SNFM/R. The Lee algorithm is the original NMF algorithm. The SAM-Set of significantly changing probes was used to generate 100 sets of 300 randomly selected genes. Patterns and reconstruction error were found for each set of randomly selected genes. The pattern finding utilized both the inflammation (dose 40) pathology constraint as well as no pathology. The algorithms were run four times and the average reconstruction error was calculated. The algorithms were ranked in descending order of reconstruction error with the largest error being ranked zero and the smallest error being ranked nine. The Pattern Finding algorithm, both with ($p < 0.01$) and without ($p < 0.01$) pathology information, had average ranks significantly higher than all other algorithms. As seen

in Figure 3.4, both the constrained and unconstrained algorithms finished with an average rank of more than eight.

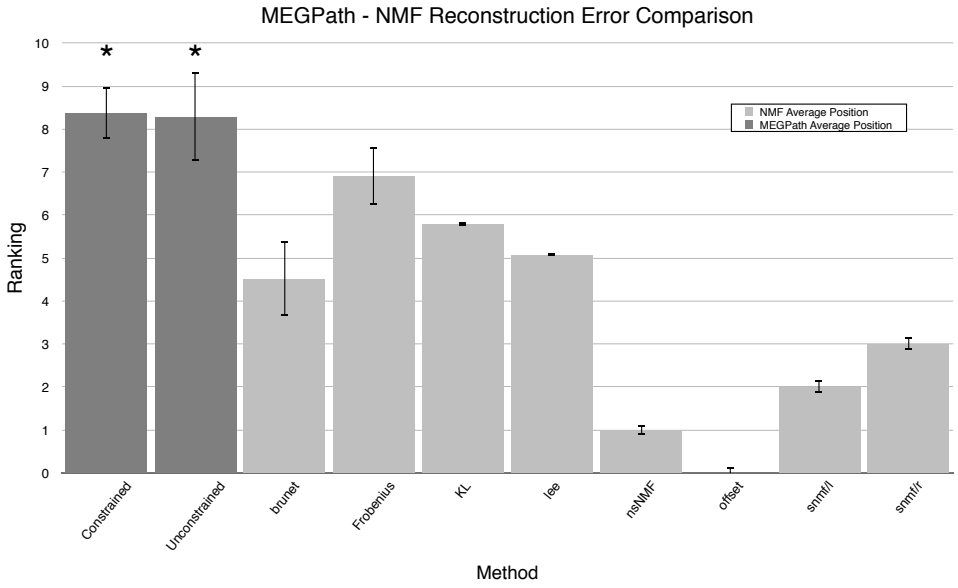

Figure 3.4: A comparison of the Pattern Finding algorithm to traditional NMF algorithms by the average reconstruction error. Gene expression for dose 40 µg across the four days was used to generate 100 sets of 300 genes each. Each algorithm was run four times on each set with the rankings assigned by descending order of error. The Pattern Finding algorithm (utilizing both pathology constraints and no constraints) performed significantly ($p < 0.01$; Wilcoxon Signed-Rank Test and Mann-Whitney) better than all other algorithms.

3.3.3 Functional Process Evaluation

The Functional Process Evaluation algorithm was also evaluated using a mock data set. First, a "genome" of 2,500 genes was created with 100 of the genes being labeled as interesting. Coefficients were randomly generated using a normal distribution with the 100 highest coefficients being assigned to the interesting genes. A set consisting of 50 randomly chosen interesting genes was generated and compared to 99 sets of randomly chosen genes. Only the set consisting of the interesting genes was found to

be significant ($p < 0.01$).

3.4 Results

In many *in vivo* experiments, disease pathologies may be observed. However, the underlying molecular mechanisms may not be understood. As seen in Figure 3.1, the MEGPath system is used in conjunction with dose response, time series mRNA microarray data and quantitative pathology scores. In total, 480 mice were used to generate data in an aspiration experiment with scores being found at 1, 7, 28, and 56 days post exposure and eight mice exposed to 0, 10, 20, 40 and 80 µg of MWCNT at each time point. The mice were divided into three groups: 1) genome-wide mRNA expression profiling was performed on 160 mouse lung tissue, 2) BAL inflammation scores found for 160 mice, and 3) morphometric analysis of Sirius red staining for collagen on 160 mice [83, 96]. These data provided genome-wide expression data as well as matching quantitative pathology for lung inflammation and lung fibrosis. The combination of data allows the system to identify both pathways and genes which may be directly related to the pathology. The MEGPath system consists of four parts: Gene Identification, Pattern Finding, Coefficient Expansion, and Functional Process Evaluation.

The Gene Identification step was run on the 41,059 mRNA probes and found a total of 2,996 probes that were significantly changed. We call this set of significant probes the SAM-Set. The Pattern Finding algorithm found patterns from the SAM-Set. Pattern coefficients were then found for each gene in the genome.

The Functional Process Evaluation step was used to identify leading sets. Gene sets from the MSigDB [111] C2 Canonical Pathways and C5 Gene Ontology databases were used. The C2 and C5 databases consist of a combined 2,334 curated sets of genes

corresponding to metabolic and signaling pathways as well as sets derived from the GO project. These gene sets provide curated information on functional relationships between genes.

3.4.1 Evaluation of the MEGPath system

To check if the MEGPath system was capable of identifying patterns from expression data, we generated 3,000 test expressions. All expressions were generated from a random combination of the same three randomly chosen probes, A_51_P227275, A_51_P454519, and A_52_P342202. The Pattern Finding algorithm was run 16 consecutive times, each time finding three patterns. The identified patterns closely resemble the original probes' patterns as seen in Figure 3.5, demonstrating that the system can identify patterns known to be in the data. In addition, this demonstrates that the normalization of the gene expression does not affect the ability of the system to find relevant patterns.

To compare if the patterns found by the system were better than randomly chosen patterns, three probes were randomly chosen and used for data reconstruction. This randomized process was run 100 consecutive times (Figure 3.6), each time with new probes as patterns. Overall the random patterns produced an average reconstruction error of 1727.318. However, using patterns found while incorporating the Gene Identification, Pattern Finding and Coefficient Expander steps yielded over eight consecutive runs an average reconstruction error of 998.515. The found patterns resulted in a significant ($p < 0.01$) reduction in error over randomly chosen patterns with the found patterns always performing better than the randomly chosen ones.

To justify using the Gene Identification step, the SAM-set patterns were checked to see if they reduced the overall error. Patterns were found using the full genome

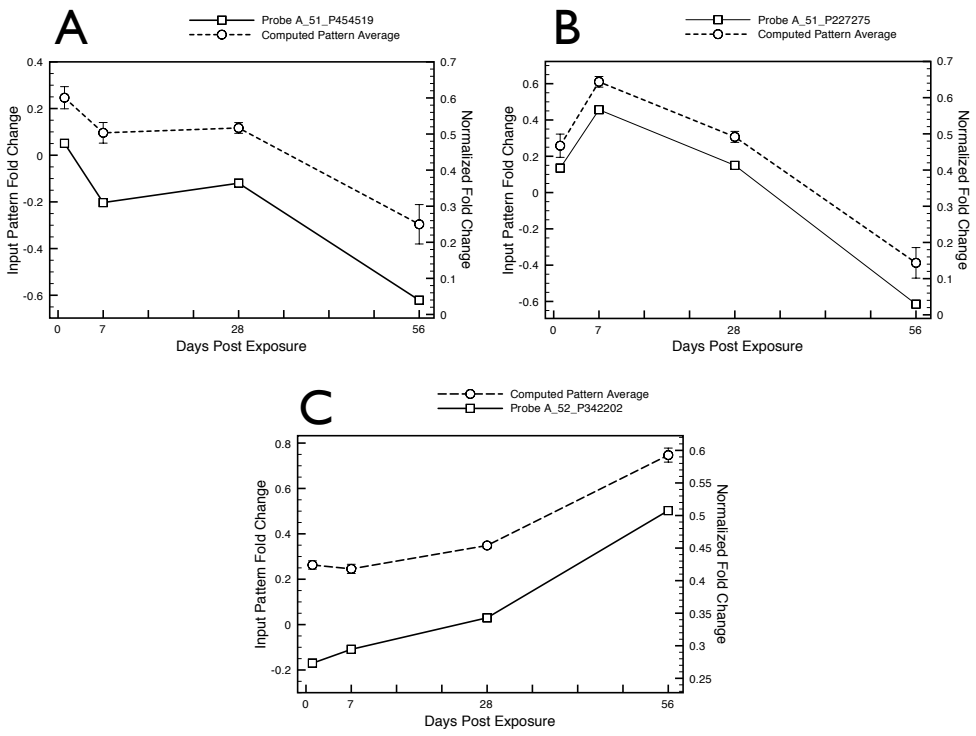

Figure 3.5: Patterns were found in mock data generated from three actual probes. The mock data consisted of 3,000 "genes" which were generated from three randomly selected probes. Gene expression for dose 40 µg across the four days was used. The patterns shown are the average of 16 consecutive runs.

and the total reconstruction error was computed. For comparison, patterns were found using the SAM-Set only and the total error was computed after using the Coefficient Expander step. Eight consecutive trials were performed and, using the Gene Identification step, always led to lower total error as seen in Table 3.1. Using an unpaired t-test, the means of the two groups significantly differed ($p < 0.01$). Since the SAM-Set is always a strict subset of the whole genome this means that using the Gene Identification step reduces pattern finding time while not increasing reconstruction error.

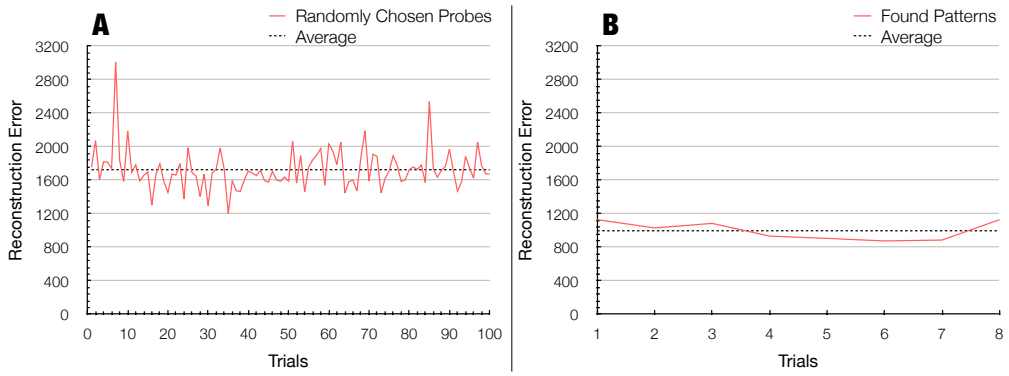

Figure 3.6: A comparison of the reconstruction error of both the Pattern Finding algorithm and randomly chosen probes. Figure A shows the results of 100 consecutive reconstructions each using randomly chosen probes as the patterns. Figure B shows the results of eight consecutive runs with patterns identified from the MEGPath system. The system identified patterns performed significantly ($p < 0.01$) better on average than the randomly chosen probes.

Table 3.1: Comparison of error of reconstruction between patterns found using all probes and the 2996 probes. Using an unpaired t-test, the two groups were found to differ significantly with $p < 0.01$.

All Probes	2996 Probes
1170.233	1131.003
1209.015	1032.582
1269.886	1086.619
1298.348	934.891
1170.796	907.779
1138.710	876.944
1182.705	887.915
1195.762	1130.390

3.4.2 Incorporating histopathological data

The MEGPath system can incorporate quantitative pathological data. Pathological scores of pulmonary inflammation data (BAL polymorphonuclear leukocytes) for the 40 µg dose across the time points [96] were used to identify inflammation related

Table 3.2: Selected Pathways found significant to Pattern 1 in Dose 40 μg, using an constrained search and the C2 database.

Pathway	Adjusted p-value
SIGNALING IN IMMUNE SYSTEM	0.0
REGULATION OF INSULIN SECRETION	0.0
SIGNALING IN IMMUNE SYSTEM	0.0
REGULATION OF INSULIN SECRETION	0.0
LYSOSOME	0.0
T CELL RECEPTOR SIGNALING PATHWAY	0.0
ECM RECEPTOR INTERACTION	0.0
PRIMARY IMMUNODEFICIENCY	0.0
G2 PATHWAY	0.0
CELL CYCLE MITOTIC	0.01318
CHEMOKINE SIGNALING PATHWAY	0.01318
G1 S TRANSITION	0.01318
PLATELET ACTIVATION TRIGGERS	0.01318
FMLP PATHWAY	0.01318
INTESTINAL IMMUNE NETWORK FOR IGA PRODUCTION	0.02292
FORMATION OF PLATELET PLUG	0.02726
IL2RB PATHWAY	0.02726

pathways. The leading set of a pathway is the subset of genes that are most closely related to the pattern. Since genes are allowed in multiple co-expression groups, not all genes in a pathway's leading set have identical looking expression. However, as seen in Figure 3.7, the average expression of the leading set closely resembles the pattern. The pathological pattern was fixed as Pattern 1 with the other two patterns being found automatically by the system. The SAM-Set probes were used in the Pattern Finding component. In total, 111 gene sets were found to be significant with Pattern 1 for 40 µg dose across the time points with a constrained search. Several significant pathways are related to cell proliferation, immune response and chemokine, which are reported to be relevant to inflammation in the literature (Table 3.2).

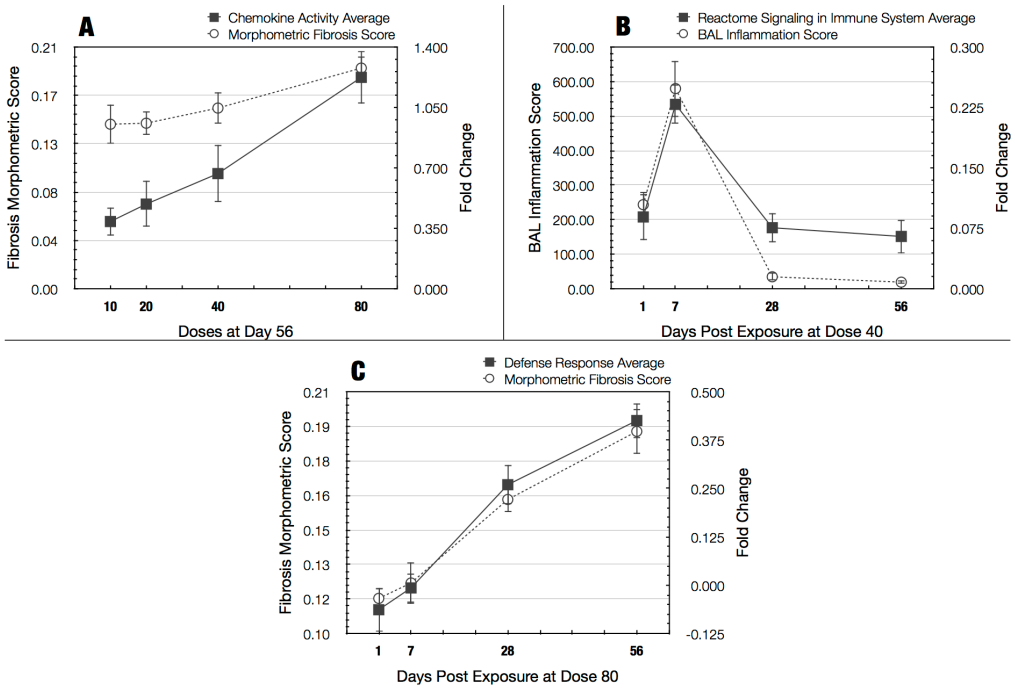

Figure 3.7: Average of the genes in the leading set of Chemokine Activity from the C5 database (A), Reactome Signaling in Immune System (B) from the C2 database, and Defense Response from the C5 database. The constrained histopathological patterns are also shown: day 56 fibrosis (A), dose 40 µg inflammation (BAL polymorphonuclear leukocytes) (B), and dose 80 µg fibrosis (C).

In order to identify mechanisms related to fibrosis, the average thickness of the alveolar connective tissue [83] was used as a pathological pattern. Patterns were found for the 80 µg dose across the time points and for day 56 across the doses. As with the inflammation data, Pattern 1 was constrained to the pathological pattern and was normalized to the range of 0 to 1 and the SAM-Set probes were used to find patterns. In total, 69 gene sets were found to be significant with Pattern 1 for 80 µg dose across the time points with a constrained search. Likewise, 85 gene sets were found to be significant with Pattern 1 for day 56 across the doses using a constrained search.

Experimental validation was achieved by first using Ingenuity Pathway Analysis

to identify, from the leading sets identified by our system, the genes with known involvement in inflammation or fibrosis. In total, IPA found 67 genes that were involved with inflammation with *ccl2* occurring most frequently in leading sets. *Ccl2* was selected for *in vitro* analysis. After treatment with MWCNT, protein levels were significantly up-regulated, suggesting that *ccl2* responded similarly *in vitro* as *in vivo*. These results suggest that *in vitro* study may be used for future study of *ccl2* and MWCNT [37].

3.4.3 Without histopathological data

The MEGPath system can find pathways, functions and genes when using both pathology data and expression data. Pathogenesis progress is important for toxicity as affected organs may not be known. We further explored to see if we could identify potential pathologies using only mRNA expression data. Patterns were found using the different doses across days. As demonstrated in Figure 3.8, the system finds gene sets whose expression may vary at points but as a group strongly resembles the pattern, genes may be influenced by multiple functions and not have identical patterns of expression. All patterns were found using the SAM-Set probes with coefficients expanded to all probes using the Coefficient Expander algorithm.

Significant pathways, Table 3.3, were found to match Pattern 1 for the 40 µg dose across the time points. The average expression of the leading sets resembles the lung inflammation pattern (BAL polymorphonuclear leukocytes) reported by Porter et al.[96] in the same animal studies. The results show when no pathological data is provided, our system finds potential pathological patterns and related pathways from time series gene expression data.

The MEGPath system can also be used to identify patterns using multiple days

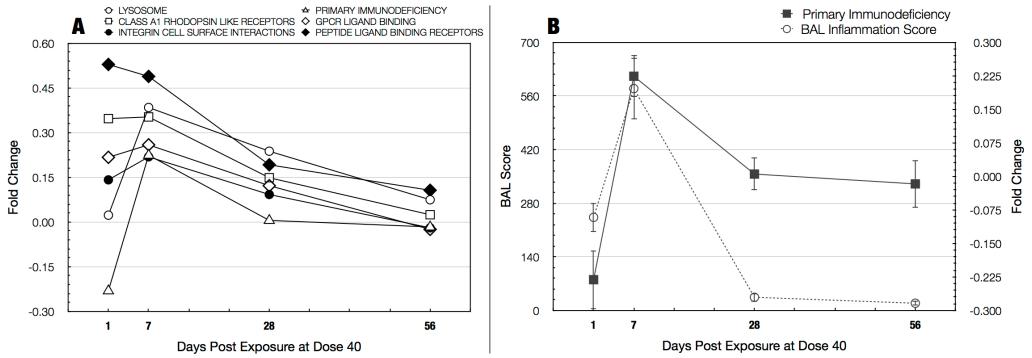

Figure 3.8: (A) The average expression for the leading sets which matched Pattern 1 using an unconstrained search. (B) The average expression of the leading set for KEGG Primary Immunodeficiency with the inflammation histopathology (BAL polymorphonuclear leukocytes) included for comparison.

Table 3.3: Pathways found significant to Pattern 1 in dose 40 µg, using an unconstrained search and the C2 database.

Pathway	Adjusted p-value
KEGG LYSOSOME	0.0
KEGG PRIMARY IMMUNODEFICIENCY	0.04393
REACTOME CLASS A1 RHODOPSIN LIKE RECEPTORS	0.04393
REACTOME GPCR LIGAND BINDING	0.0
REACTOME INTEGRIN CELL SURFACE INTERACTIONS	0.0
REACTOME PEPTIDE LIGAND BINDING RECEPTORS	0.0

and multiple doses. Seven patterns were found from the SAM-Set of probes across eight conditions, using the 40 µg dose and the 80 µg dose across the time points. Many of the patterns exhibit strong similarity between the 40 µg dose sections and the 80 µg dose section (Figure 3.9). The results suggest similar dose response in both the 40 µg and 80 µg doses and the potential for future experiments not needing the higher dose.

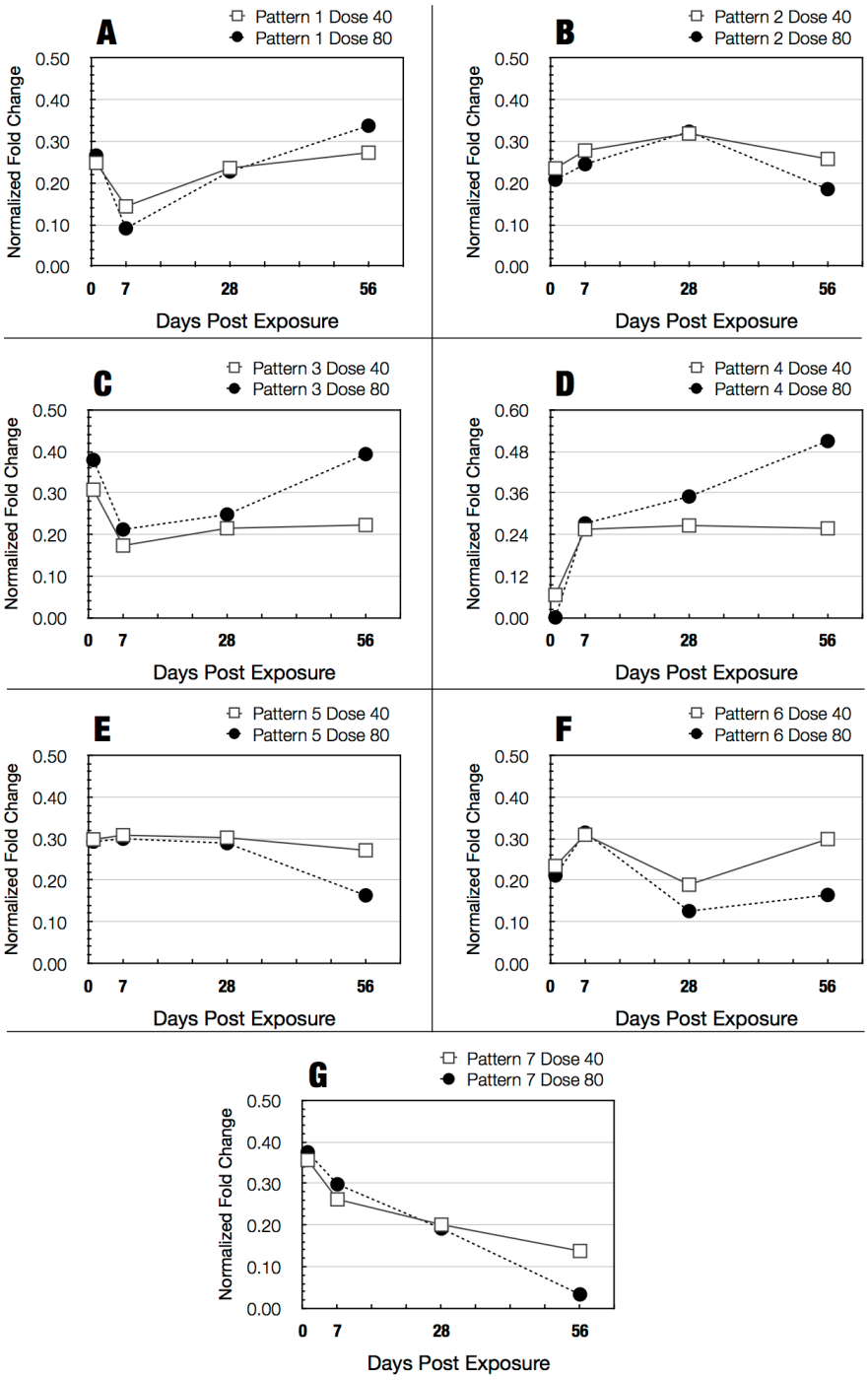

Figure 3.9: The seven patterns found using an unconstrained search and eight conditions. The SAM-set was used but included the conditions for dose 40 µg and dose 80 µg, both across the four time points. The patterns are split into the dose 40 µg segment and dose 80 µg segment to show similarity.

3.4.4 ClueGO Visualization

A visualization of the gene relationships in a leading set can be created with the ClueGo [12] tool. For these gene lists, a functional analysis of only a single cluster was conducted. The gene cluster list was populated using the model species as Homo Sapiens and allowing ClueGO to automatically select correct identifiers. ClueGO allows the user to specify the level of detail they wish to acquire from the analysis, ranging from highly detailed results to a broader more general global setting. In this analysis, the slider was placed upon the medium setting showing no favor towards the global or detailed option. The GO biological process was selected as the ontology of choice using all available codes. The leading sets for the Immune Response pathway (Figure 3.10) and Response To External Stimulus pathway (Figure 3.11) were visualized.

3.5 Discussion

Dose response and time series experiments are an important tool for identifying toxicity and disease progression. Previous studies have used *in vivo* and *in vitro* genome-wide mRNA expression data to infer toxicity. In addition to microarray data, benchmark dose techniques have been combined with Gene Ontology annotations, to identify biological processes related to toxicity. However, even if quantitative pathology is observed, the underlying molecular mechanisms may be difficult to reveal.

Our system, Figure 3.1, uses both mRNA expression data and quantitative pathology scores to identify significantly changing pathways and interesting genes. We were able to identify results that are able to be experimentally validated [37].

The MEGPath system, involving a combination of Gene Identification, Pattern Finding, Coefficient Expansion and Functional Process Evaluation methods, is a com-

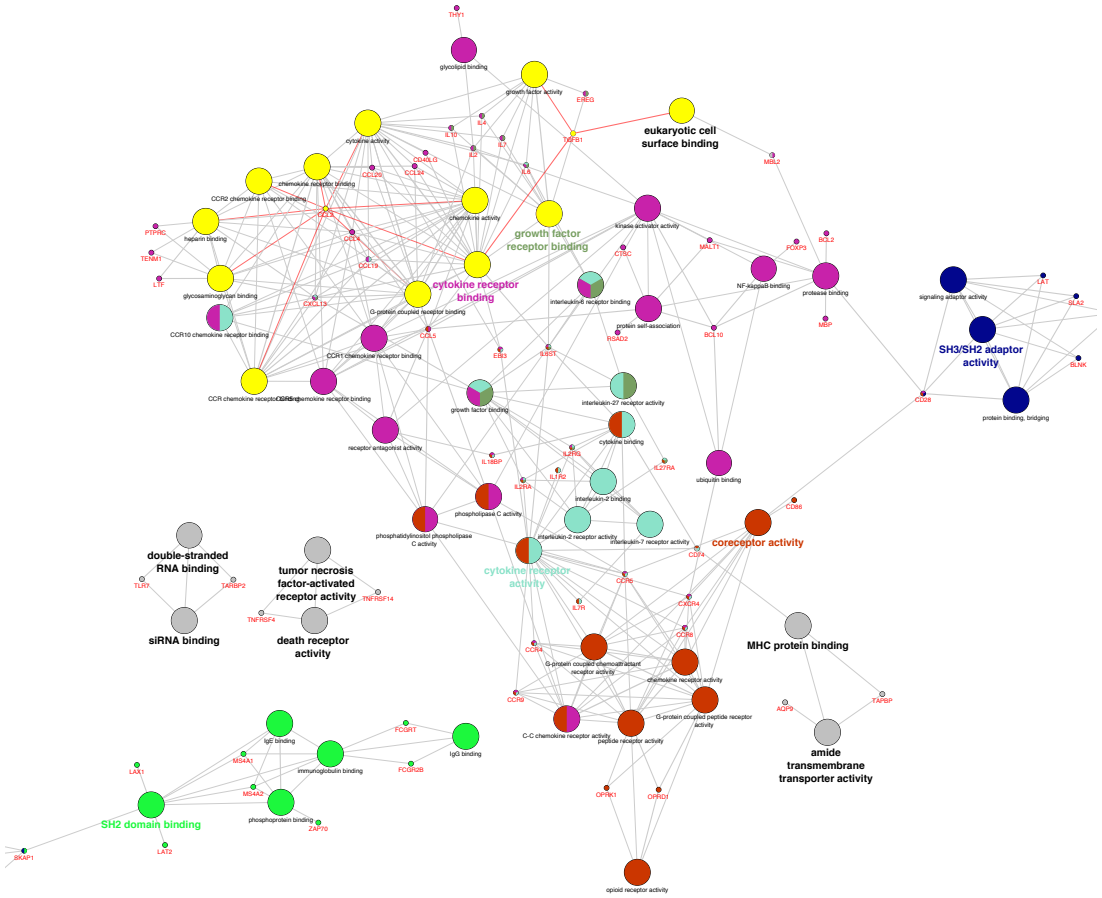

Figure 3.10: The Immune Response leading set as visualized by ClueGO.

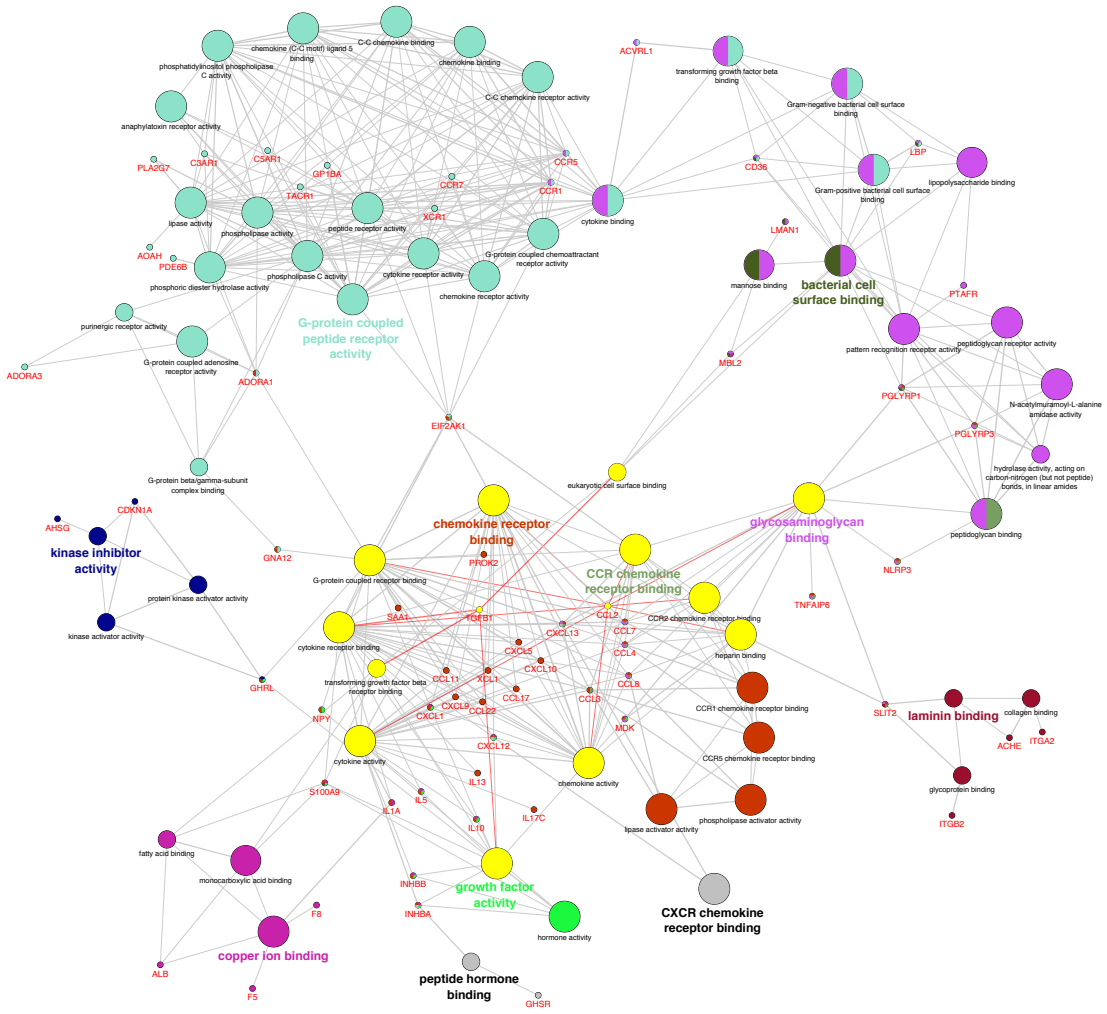

Figure 3.11: The Response to External Stimulus leading set as visualized by ClueGO.

putationally efficient technique to model dose dependent time series microarray data on the genome-scale. Our system identifies non-parametric patterns capable of reconstructing both dose dependent and time series microarray data, while also incorporating quantitative pathological data and known biological interactions. Genes are allowed in multiple co-expression groups and so can be significantly related with multiple functions and patterns.

However, the MEGPath system is not suited for working with data consisting of less than three treatment conditions. Also, unlike parametric systems, no information can be implied from unobserved experimental conditions. While our system suggests that two doses may behave the same, we can make no claim about unobserved middle doses. The system also depends on curated gene sets, which depending on granularity, may be too general.

When no pathological data is available, this system is able to identify potential pathological phenomena and related pathways. Future experiments can be developed for testing potential pathologies by observing trends in the identified significant pathways. The MEGPath system has potential applications in areas outside of toxicology such as chemo response, chemo sensitivity, and pharmacogenomics. For instance, using mRNA expression data from blood could reveal the effects of unobserved diseases or organs with potential diseases. Observing patterns across multiple days or doses could help reduce the number of conditions needed in future experiments. Repetitive higher doses could be removed, or long durations might be shortened.

3.6 Publications

Some the work described in this chapter has been published in proceedings of the IEEE International Conference on Bioinformatics and Biomedicine. In addition a journal

manuscript has been prepared for submission to PLOS Computational Biology.

J. Dymacek and N. L. Guo. Systems approach to identifying relevant pathways from phenotype information in dose-dependent time series microarray data. In *Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine, BIBM '11*, pages 290293, Atlanta, GA, USA, 2011. IEEE Computer Society.

Chapter 4

Integrated miRNA/mRNA Analysis

In this chapter we will explore the integration of miRNA with mRNA. The integrated mRNA and miRNA analysis (Figure 4.1) of time series microarray data will involve separate processing of both the expression data sets described in Chapter 3. Central to the analysis is the identification of genes (miRNA and mRNA) that correspond with the given quantitative pathology. After identifying genes, only genes that are either predicted or known target pairs will be kept. In addition, for each target pair there must be gene expression evidence of regulation. In Chapter 5, the results will be analyzed with Ingenuity Pathway Analysis and pathology involvement will be validated from the literature.

The dynamic temporal regulatory effects of microRNA are not well known. We introduce a technique for integrating miRNA and mRNA time series microarray data with known disease pathology. The integrated analysis includes identifying both mRNA and miRNA that are significantly similar to the quantitative pathology. Potential regulatory miRNA/mRNA target pairs are identified through databases of

both predicted and validated pairs. Finally, potential target pairs are filtered by examining the second derivatives of the fold changes over time. Our system was used on genome-wide microarray expression data of mouse lungs ($n = 160$) following aspiration of multi-walled carbon nanotubes. This system shows promise of readily identifying miRNA for further study as potential biomarker use.

The remainder of this chapter is organized as follows. Section 4.1 describes the methods used to analyze both the mRNA and miRNA data. Results are described in Section 4.2.

4.1 Methods

The integrated mRNA and miRNA analysis (Figure 4.1) of time series microarray data involves separate processing of both the expression data sets. Central to the analysis is the identification of genes (miRNA and mRNA) that correspond with the given quantitative pathology. After identifying genes, only genes that are either predicted or known target pairs are kept. In addition, for each target pair there must be gene expression evidence of regulation.

4.1.1 Non-negative Matrix Factorization

Non-negative matrix factorization [71] (NMF) allows for the identification of underlying patterns in multi-dimensional microarray data. These patterns can be thought of as biological functions responding to a disease or exposure. By fixing, or constraining, one pattern to a known pathology, we can identify genes that are strongly influenced by a function resembling the pathology.

Let D be the original fold change data (either mRNA or miRNA expression data), containing values at multiple conditions for each probe. The NMF algorithm tries

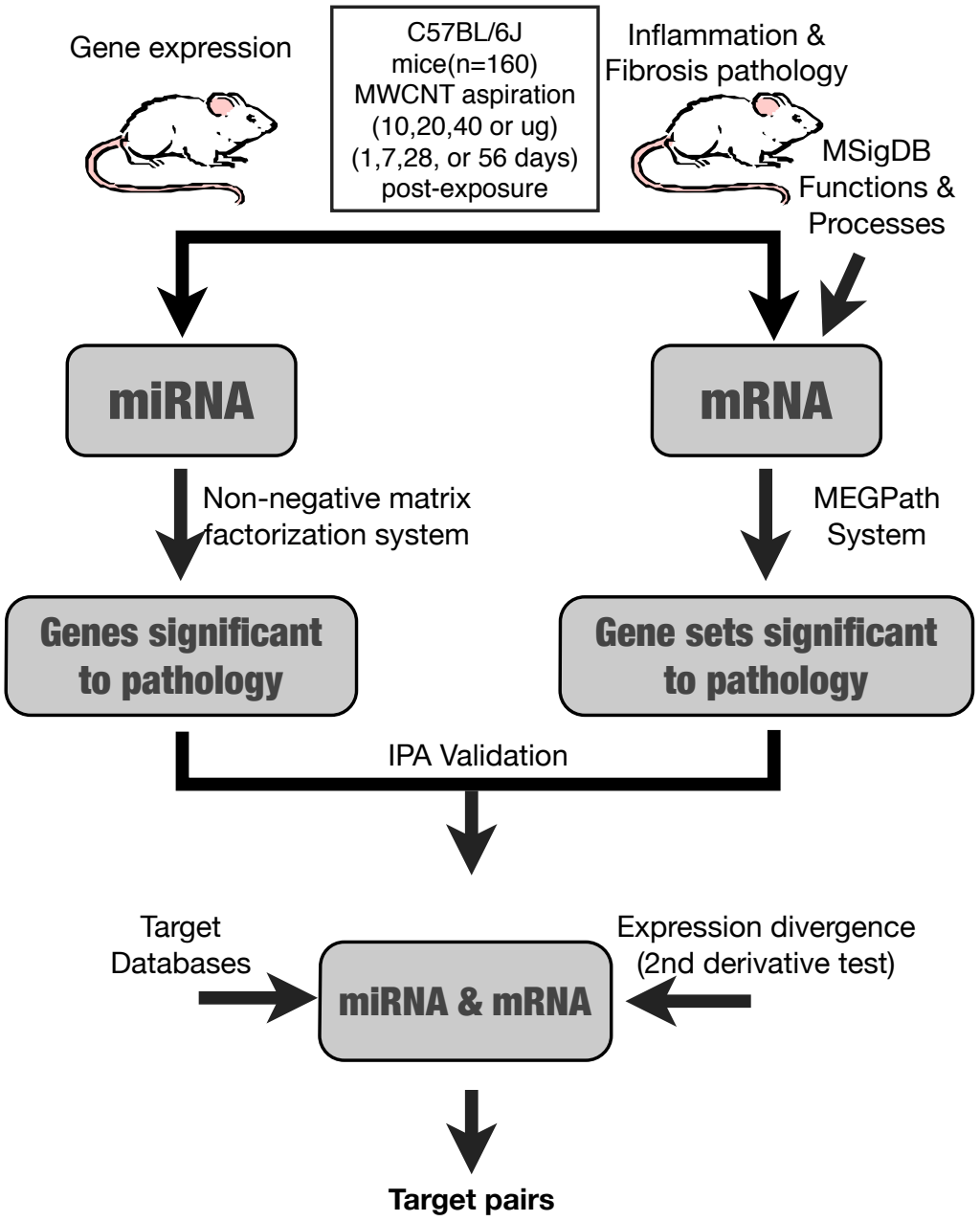

Figure 4.1: Flow diagram of an integrated miRNA/mRNA analysis. Both microarray data and pathology data are used in the miRNA and mRNA analysis. Genes significant with the pathology are identified and then potential target pairs analyzed.

to find matrices P and C such $D = C * P$. The pattern matrix, P, consists of underlying biological functions that can be used to reconstruct the expression data. The coefficient matrix, C, relates each probe to the each pattern. Due to noise, it is unlikely that an exact solution can be found. The algorithm tries to minimize the difference between the original fold change matrix (D) and the reconstructed fold change ($C * P$).

Our NMF algorithm works as a Monte-Carlo Markov Chain, where a probability density function is associated with each entry in the P and C matrices. To satisfy the non-negative constraint the fold change data is normalized to the $[0 - 1)$ domain. The density functions are updated at each algorithm step and final entry values are found using a simulated annealing process.

Unlike traditional NMF algorithms, our algorithm allows constraints to be added. Constraints are implemented through manipulation of the density functions. A density function can be shared across multiple entries or constrained to always return a specific value. Pathology patterns are encoded as a single density function with a time series pathology constraint encoding each constraint entry as a relative offset from the previous time point.

4.1.2 mRNA Analysis

The MEGPath [34] system was used to identify mRNA genes potentially involved with the pathology. The MEGPath system was designed to identify sets of genes that, as a group, are significantly related to a known pathology pattern. The first step was to identify a subset of genes which were significantly changing. Since most of the genes change very little over time, the patterns should be found from genes with more noticeable changes. Genes with fold changes that were changing significantly

were identified at each dose and time condition and considered significant.

The MEGPath system uses the significant genes and constraint pathology along with the described NMF algorithm to identify underlying biological patterns. The identified patterns are then used to find genome wide coefficients. These coefficients (C) relate each gene to each of the patterns (P) and are used to identify sets of genes that are significantly related to the pattern. The gene sets used are from the curated MSigDB [111] database allowing for annotations for each gene's function. Genes may be contained in multiple gene sets. The gene sets are functionally related and are not required to be co-expressed. An example of a pathology pattern and some of the identified mRNA genes can be seen in Figure 4.2.

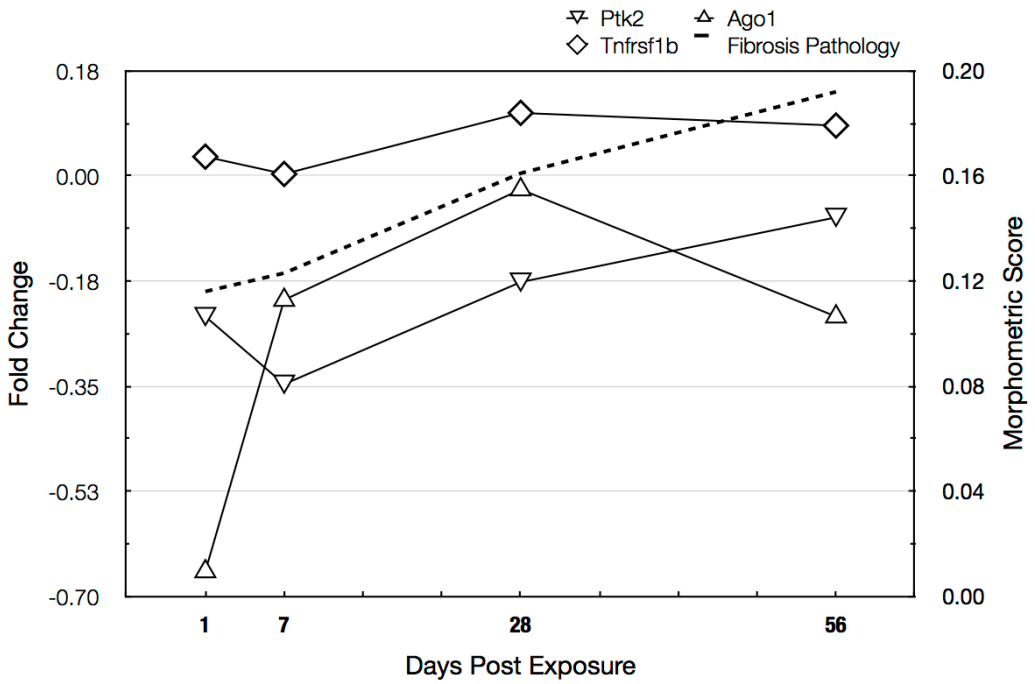

Figure 4.2: The dose 80 µg fibrosis pathology and significantly related mRNA. The fibrosis pathology was used as a constraint in the NMF algorithm for both miRNA and mRNA. The mRNA were found to be related to fibrosis by both the pathology and IPA.

4.1.3 miRNA Analysis

The MEGPath system identifies mRNA that are functionally involved with a constraint pathology. Unfortunately, the MEGPath system relies on functional annotations from the MSigDB for each mRNA gene and there are currently no similar functional annotation databases for miRNA. Our methodology identifies miRNA that are functionally involved with the pathology by utilizing known potential miRNA targets and both mRNA and miRNA expression data to select pairs that show signs of being regulated.

The described constrained NMF algorithm was used to relate the constraint pathology to the miRNA. Patterns (P) and each probe's corresponding coefficients (C) were computed from the time series miRNA microarray data. The first pattern was constrained to match the pathology. A probe's error was calculated as the absolute difference between the normalized reconstructed probe expression and the normalized original probe expression.

Coefficients corresponding to the constrained pathology pattern were considered for significance. The probe's error values were subtracted from the probe's coefficients to eliminate probes with high reconstruction error. This step reduces the opportunity for false positives generated from probes with noisy data. These modified genome-wide coefficients were then plotted to visibly check for a normal distribution (Figure 4.4). A normal distribution was fitted to the modified coefficients and probes with coefficients with a probability less than 5% were kept. Only the coefficients for the pathology pattern were used. An example of a miRNA, let-7c, and related lung fibrosis pathology can be seen in Figure 4.3.

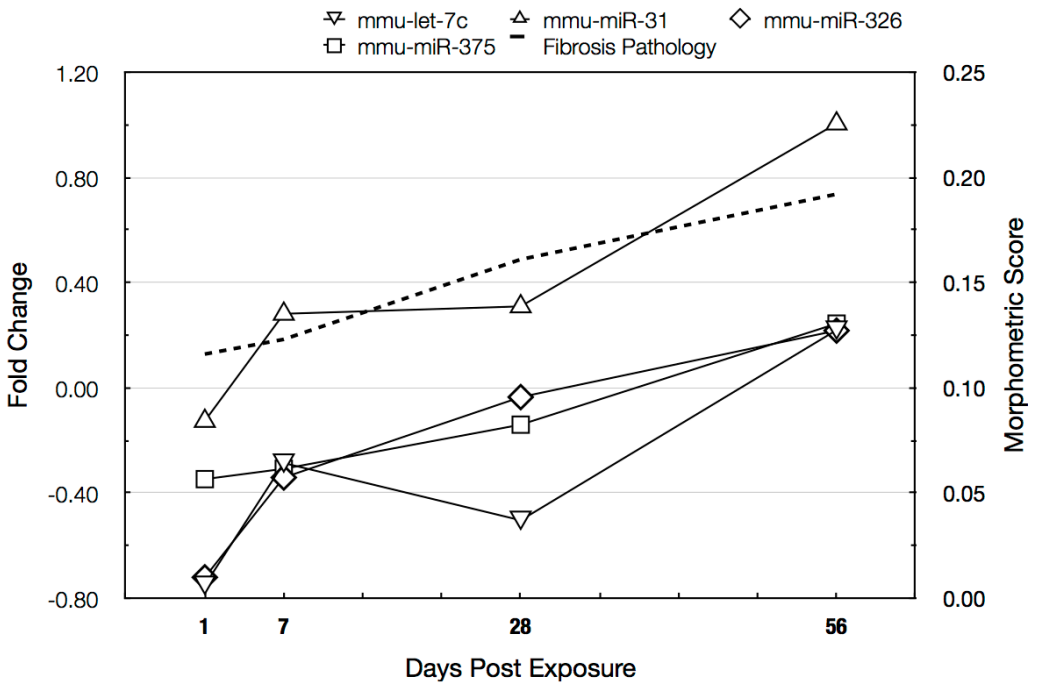

Figure 4.3: The dose 80 µg fibrosis pathology and significantly related miRNA. The fibrosis pathology was used as a constraint in the NMF algorithm for both miRNA and mRNA.

4.1.4 Integrated Analysis

After identifying both mRNA and miRNA for further study, an integrated analysis was applied. The integrated analysis was performed in two steps.

Potential target pairs were identified from databases. Three databases were used: miRTarBase [56], miRecords [131], and TargetScan [73]. The TargetScan database provides predicted regulatory miRNA/mRNA pairs. Both miRTarBase and miRecords provide a mix of published validated pairs as well as predicted pairs. Both validated human and mouse pairs were kept. The miRBase [48] website was used to translate a probe's gene name into the most recent form.

The potential target pairs were then filtered according to the gene expression data. Traditionally, a negative correlation analysis is performed as the miRNA and mRNA

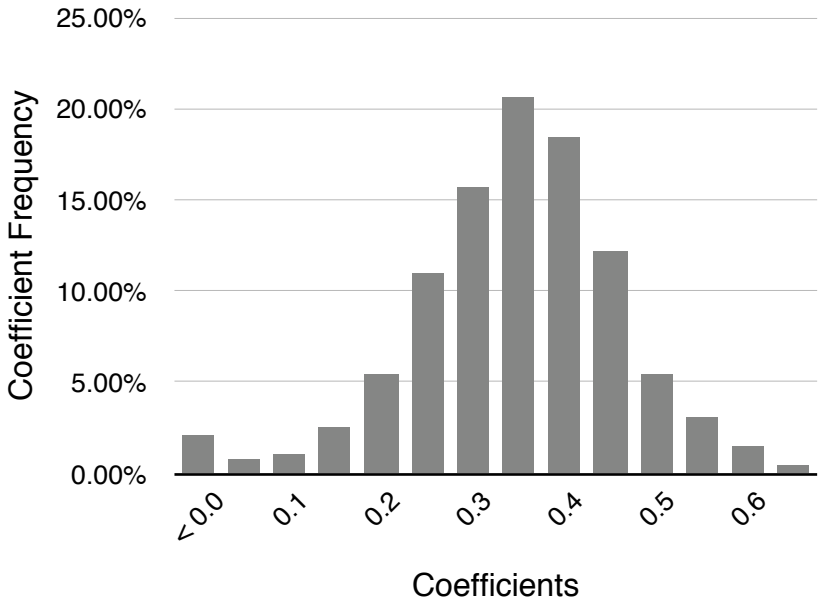

Figure 4.4: Histogram of the coefficients corresponding to the fibrosis pathology pattern. Each bar shows the number of coefficients in the range ending with the label. Error values were subtracted from coefficients making some be less than 0. A normal distribution was fitted to the data.

expressions should be in opposite directions to signify regulation. Each miRNA may target many mRNA; hence, over the course of time a miRNA's expression may be changing to help regulate multiple mRNA and not be "opposite" of a targeted mRNA. In addition, the discrete time points of time series data may miss critical moments where a miRNA's expression changes. These issues were addressed by using the second derivatives of the fold change. The second derivative is the "change of the change", or given a gene G's fold change G_i at times $i = 1 \ldots (n-1)$:

$$G'_i = (G_{i+1} - G_i) - (G_i - G_{i-1}).$$

The second derivative is not defined for the first time point so the first fold change is

duplicated:

$$G'_0 = (G_1 - G_0) - (G_0 - G_0) = (G_1 - G_0).$$

A miRNA/mRNA pair are considered targeted pairs if the second derivatives at the same time point are opposite signs; hence, for miRNA R and mRNA M to be a targeted pair:

$$\exists\, i = 0 \ldots (n-1)\ where\ R'_i * M'_i \leq 0.$$

An example of target pairs, between let-7c and three mRNA, found through this analysis can be seen in Figure 4.5. The target pairs have differing second derivatives in at least one time point.

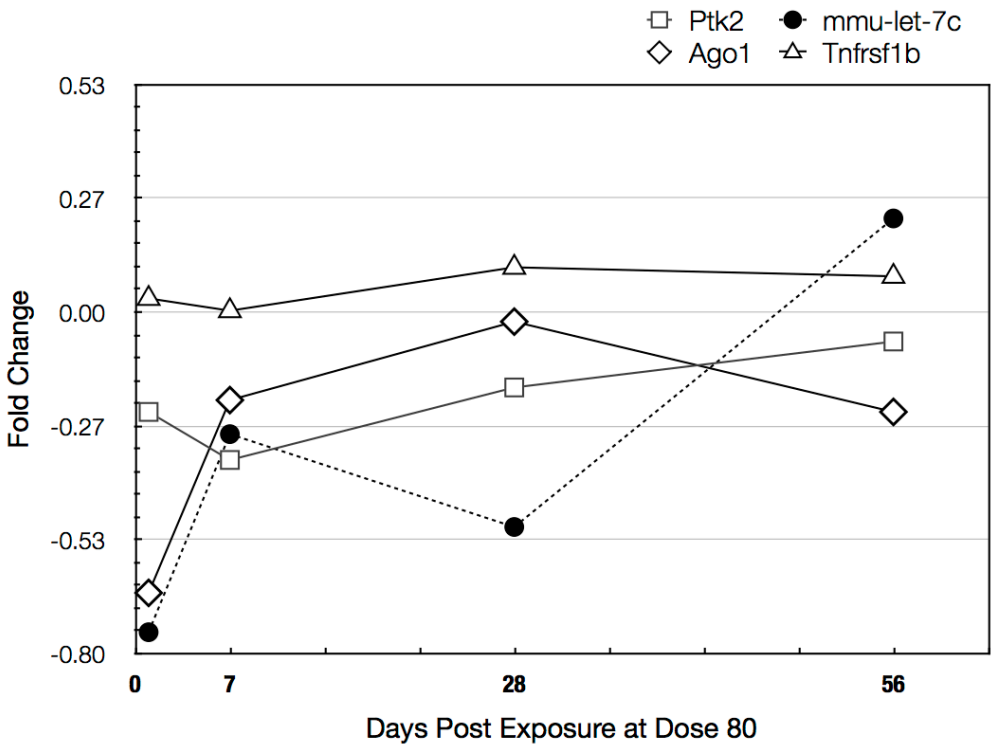

Figure 4.5: The miRNA let-7c was found to be significantly related to the fibrosis pathology. The mRNA were found to be related to fibrosis by both the pathology and IPA. All mRNA and let-7c were identified as target pairs, with differing second derivatives in at least one time point.

Table 4.1: mRNA genes that were found to be significantly related to the time series dose 80 µg fibrosis pathology. Genes were filtered by IPA to be involved in fibrosis.

ACTC1	EGFR	LGALS3	S100A4
ADORA1	EGR1	LGMN	SELE
ADORA3	F11	MMP8	SELP
ADORA2B	FAS	MMP9	SERPINE1
AGO1	FCGR2B	MMP12	SLC4A1
ARG1	FN1	MMP13	SLC8A1
ARID4A	GCLC	MMP14	SMAD4
ATF3	GSK3B	MYD88	SMURF2
BDKRB2	HBEGF	NFKBIA	SOAT1
BMPR2	HIF1A	OSM	SOCS1
C3	HMGCS1	PDGFRA	SOCS3
CCL2	HPX	PLA2G10	SOD2
CCL17	IGF1	PLAT	STAB1
CCL24	IL5	PLAUR	TIMP1
CCR1	IL6	PROC	TLR2
CD74	IL11	PTGIR	TNF
CEBPB	IL12B	PTGS2	TNFAIP3
CSF3	IL1B	PTK2	TNFRSF1B
CTSB	IL1R1	PTX3	TNNC1
CTSK	IL1RN	RASSF1	VEGFA
CX3CL1	INHBA	RELB	VIM
DAG1	KCNN4	RGS16	WRN
EDNRB			

4.2 Results and Implementation

Results were obtained from an *in vivo* dose-response time series multi-wall carbon nano-tube (MWCNT) aspiration exposure experiment [96]. The experimental results indicated lung damage, inflammation, and fibrosis.

All code was written in Java and the analyses were run on a standard laptop computer.

Table 4.2: mRNA genes that were found to be significantly related to the time series dose 40 µg inflammation pathology. Genes were filtered by IPA to only those involved in inflammation.

ABCC1	ADAM8	ADORA1	ADORA3	AGT
AGTR2	AHSG	AIF1	ANGPTL2	ANXA1
AOC3	BLNK	C3AR1	CAPG	CCDC88A
CCL2	CCL3	CCL4	CCL5	CCL7
CCL17	CCL20	CCL24	CCR1	CD9
CD14	CD44	CD63	CD69	CD74
CDKN1A	CEBPB	CHRNA7	CLEC5A	CLEC7A
CSF3	CX3CL1	CXADR	CXCL1	CXCL2
CXCL3	CXCL5	CXCL10	CXCL12	CYBA
CYSLTR1	DPP4	EGR1	EPAS1	FCER1G
FCGR2B	FN1	FOS	FOXP3	GHRL
GIT2	GNAI2	HYAL1	IGF1	IKBKG
IL5	IL6	IL16	IL24	IL12B
IL1B	IL1R1	IL1RN	IL23A	IL2RA
INHBA	ISG15	ITGB2	ITGB6	ITGB7
LGALS3	LITAF	LUM	LY75	MCL1
MDK	MEFV	MMP8	MMP9	MMP14
MSR1	MYD88	NAD+	NCKAP1L	NCL
NFIL3	NPY	NR1D1	NR2C2	OLR1
OSM	PIK3R5	PLA2G7	PLA2G10	PLAUR
PPBP	PRG2	PRKCD	PROC	PROCR
PTGER3	PTGS1	PTGS2	PTN	PTPN2
PTX3	RGS1	RIPK3	S100A8	SDC4
SELE	SELP	SERPINE1	SFRP1	SLC11A1
SOCS1	SOCS3	SPACA3	SPHK1	SPP1
STEAP2	TNF	TNFAIP3	TNFRSF4	TNFRSF9
VAV1	VEGFA	WNT5A	XCR1	

4.2.1 Data

The data set consisted of dose-dependent time series mRNA and miRNA microarray expression data. Microarray data came from 160 MWCNT exposed mice (C57BL/6J). The doses were 0(dm), 10, 20, 40, or 80 µg of MWCNT. Total RNA was extracted from the mouse lungs at 1, 7, 28, and 56 days post-exposure for each dose condition. Agilent Mouse Whole Genome Arrays were used for mRNA expression profiling. In total the mRNA genome contained 41,059 probes and the miRNA genome consisted of 484 probes. Our mRNA data has been deposited to the NCBI Gene Expression Omnibus (GEO) repository with accession number GSE29042, miRNA data is in the process of being deposited. We also maintain a website for browsing both data sets on the web.[1] The microarray data were log-transformed for analysis. In addition to mRNA and miRNA data, 160 mice were used for quantitative inflammation scores and 160 mice were used for quantitative fibrosis scores. Inflammation scores were derived from the analysis of BAL fluid [96]. The average thickness of the alveolar connection tissue was used for fibrosis scores. These were found from the morphometric analysis of Sirius Red staining for connective tissue [83].

4.2.2 Significantly Changing Probes

Significant mRNA and miRNA probes were found using the same procedure. Missing data were imputed using the K-means nearest neighbor algorithm as implemented by the *impute.knn* function in the impute R package from Bioconductor (Seattle, WA). Using the Bioconductor package, a set of differentially expressed genes for each dose and time point were identified by performing a two-class unpaired Significance Analysis of Microarrays (SAM) between the treated samples and the dose zero samples

[1] http://www.mwcnttranscriptome.org

from the corresponding time point. A threshold delta value was chosen to produce a false discovery rate of 1%(mRNA) and 5%(miRNA) using the findDelta function from the same package. The list of significant probes was filtered by only keeping probes that were at least 1.5 fold up- or down-regulated. Fold changes were computed from the data before imputation of missing values.

4.2.3 mRNA Results

Gene sets significantly related to the time series 80 µg dose fibrosis and the time series 40 µg dose inflammation pathologies were obtained from the mRNA microarray data. The SAM analysis was run on all conditions to obtain significantly changing mRNA at either a dose or time condition. In addition, genes significant with linear models [49] were used. The 2,996 significant probes were used with the NMF algorithm to identify three underlying patterns from the four time points for both pathologies. Genome-wide coefficients were then found relating each gene to the pathologies, coefficient error influence was minimized by identifying sets of genes. All gene sets were from the MSigDB curated databases.

The fibrosis pathology coefficients were used to identify significantly represented sets of genes, identifying 30 gene sets found from the C2 database and 39 sets from the C5 database. The inflammation pathology coefficients were used to identify significantly represented sets of inflammation genes, with 50 gene sets found from the C2 database and 61 sets from the C5 database. Many genes were found in multiple gene sets, two such genes CCL2 and VEGFA were validated *in vitro* as changing expression when exposed to MWCNT [37].

All genes from the gene sets associated with the fibrosis pathology were screened through Ingenuity Pathway Analysis (IPA), an online curated literature based tool

Table 4.3: miRNAs associated with the fibrotic pathology and their associated mRNA binding partners. Each pair passes the second derivative test. Significantly changed miRNAs are highlighted in bold.

miRNA	**miRTarBase, miRecords Experimentally Confirmed**	**TargetScan Predicted**
let-7c-5p	AGO1, PTK2	IL6, RGS16, TNFAIP3, TNFRSF1B
miR-205-5p	VEGFA	IL1R1, PTX3
miR-23b-3p	SMAD4	EDNRB, FAS, GSK3B, IL11
miR-31-5p	HIF1A, SELE	HBEGF
miR-326-3p	AGO1	RASSF1
miR-328-3p	AGO1	
miR-330-3p	AGO1, VEGFA	RASSF1
miR-34c-3p		PDGFRA, SERPINE1, SMAD4
miR-375-3p	KCNN4	BMPR2, EDNRB, RGS16
miR-455-3p	AGO1	
miR-652-3p	AGO1	
miR-92b-3p		

(ingenuity.com). The 89 mRNA genes related to fibrosis found by IPA were kept for further analysis and are listed in Table 4.1. Likewise, genes from the inflammation associated gene sets were screened through IPA. The 125 mRNA genes related to inflammation are listed in Table 4.2.

Table 4.4: miRNAs associated with the inflammation pathology and their associated mRNA binding partners. Each pair passes the second derivative test. Significantly changed miRNAs are highlighted in bold.

miRNA	**miRTarBase, miRecords Experimentally Confirmed**	**TargetScan Predicted**
miR-1224-5p		
miR-147-3p	VEGFA	
miR-188-5p		
miR-290a-5p		
miR-327		
miR-3474		
miR-380-3p		
miR-449a-5p		GNAI2, SERPINE1
miR-494-3p		TNFRSF9
miR-551b-3p		
miR-667-3p		
miR-696		
miR-703		
miR-877-5p		
miR-881-5p		
miR-92a-2-5p		MCL1, WNT5A

Table 4.5: All miRNAs significantly changed from controls identified from mice exposed to 10, 20, 40, or 80 µg MWCNT at 1, 7, 28, and 56 days post-exposure. There were no significant miRNAs identified at Dose 10.

miR-103	miR-1188	miR-125b-3p	miR-125b-5p
miR-126-3p	miR-129-5p	miR-1306	miR-130b
miR-132	miR-142-5p	miR-146a	miR-146b
miR-147	miR-149	miR-15a	miR-15a*
miR-15b	miR-16	miR-16*	miR-188-5p
miR-1892	miR-1897-5p	miR-18b	miR-1902
miR-1903	miR-1904	miR-1906	miR-1932
miR-1935	miR-1937c	miR-195	miR-1951
miR-196a	miR-196b	miR-1982*	miR-199a-3p
miR-199a-5p	miR-200a	miR-200b	miR-21
miR-2132	miR-2133	miR-2140	miR-22
miR-221	miR-222	miR-223	miR-26b
miR-296-3p	miR-297a	miR-297c	miR-29b*
miR-30c	miR-30c-1*	miR-30e	miR-31
miR-322	miR-323-5p	miR-327	miR-328
miR-330	miR-341	miR-342-3p	miR-3470b
miR-3473	miR-34a	miR-34b-3p	miR-34c
miR-370	miR-382	miR-429	miR-434-3p
miR-449a	miR-449b	miR-449c	miR-450a-3p
miR-466b-5p	miR-466h	miR-467e	miR-467h
miR-471	miR-486	miR-669a	miR-669e
miR-673-3p	miR-679	miR-696	miR-711
miR-714	miR-720	miR-744	miR-92a*

4.2.4 miRNA and Integrated Results

The SAM analysis was performed on the miRNA microarray data and identified 92 probes which were significantly changed in at least one dose time condition (Table 4.5).

The NMF algorithm was run with the 80 µg dose fibrosis time series constraint. Three patterns were found over the four time points. Since individual miRNA were being identified, the coefficients relating miRNA to the fibrosis pattern were modified by subtracting the reconstruction error to penalize noisy genes.

A normal distribution was fitted to the modified coefficients with a mean of

0.318 and a standard deviation of 0.128. Coefficients with scores greater than 0.5285 ($p < 0.05$) were considered to be related to the pathology and are listed in Table 4.3.

Similarly, the NMF algorithm was run with the 40 µg dose inflammation time series constraint. Three patterns were found over the four time points. The coefficients relating miRNA to the fibrosis pattern were modified by subtracting the reconstruction error to penalize noisy genes. A normal distribution was fitted to the modified coefficients with a mean of 0.271 and a standard deviation of 0.0849. Coefficients with scores greater than 0.4108 ($p < 0.05$) were considered to be related to the pathology and are listed in Table 4.4.

After identifying the significant miRNA, potential mRNA targets were filtered by using three databases. Additional databases could be used. Finally, potential pairs needed to have a differing second derivative in at least one time point. Only one of the miRNA significant with the fibrosis pathology did not end with a target, miR-92b. Two miRNA were both significantly changed in the SAM analysis and related to the pathology: miR-31 and miR-328. Only four of the miRNA significant with the inflammation pathology had targets: miR-147-3p, miR-449a-5p, miR-494-3p, and miR-92a-2-5p. Five miRNA changed significantly in the SAM analysis and were related to the inflammation pathology: miR-147-3p, miR-188-5p, miR-327, miR-449a-5p, and miR-696.

4.3 Discussion

This study presents a methodology for integrating both miRNA and mRNA time series data along with quantitative pathology information for identifying important miRNA regulated biological processes underlying pathogenesis. The use of a constrained NMF algorithm allows for the identification of miRNA significantly related

to the pathology while still allowing gene expression to be influenced by multiple functions. The integration of mRNA provides additional functional annotation information from IPA and MSigDB. Potential miRNA regulated mRNAs can be identified by the second derivative test, encompassing both negative correlation aspects and temporal responses.

Our system has been able to identify pairs of miRNA and potentially regulated mRNA. All of the identified miRNA were related to the quantitative pathology patterns and potential regulators of mRNA identified with fibrosis or inflammation. In particular the miRNA let-7c may have implications in lung fibrosis [9] and was shown, with potential mRNA targets, in Figure 4.5. Likewise, mir-31 has been shown to be a involved in lung fibrosis regulation [133], suggesting an active role in attempting to suppress MWCNT caused lung fibrosis. Other identified miRNA with potential lung fibrosis involvement are mir-326 [26] and mir-375 [129]. In addition, miR-449a and miR-92a have been shown to change expression with titanium dioxide nanoparticle exposure and have potential involvement with lung inflammation [51]. The predicted miRNA /mRNA targets could be validated *in vitro* or biological connections explored using IPA [36]. Although demonstrated on MWCNT toxicity data, this integrated approach could be used in other applications.

4.4 Publications

Some the work described in this chapter has been published in proceedings of the ACM International Conference on Bioinformatics and Computational Biology. In addition, a journal manuscript has been prepared for submission to Bioinformatics.

J. Dymacek and N. L. Guo. Integrated mirna and mrna analysis of time series microarray data. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pages 122127, Newport Beach,

CA, USA, 2014. ACM.

Chapter 5

Results Analysis

5.1 Biological Validation of the MEGPath System

Integrating *in vivo*, *in vitro* studies, and *in silico* analysis is a recent endeavor in toxicological sciences. Novel methods for the analysis of current *in vivo* data are needed to develop predictive *in vitro* models so as to determine the toxicity profile of multiple material variants, such as various types of carbon nano-tubes. By identifying the leading gene sets of the significant functions and pathways, our system can extract genes that are strongly associated with the inflammation and fibrosis pathologies and that have potential involvement in inflammation and collagen production. The use of Ingenuity Pathway Analysis [1] allows for global analysis of our leading sets throughout the body of accepted scientific literature so as to target our results to those genes known to be involved in inflammation and fibrosis.

[1] (Ingenuity®Systems, www.ingenuity.com)

5.1.1 Ingenuity Pathway Analysis

Data were analyzed through the use of Ingenuity Pathway Analysis (IPA). A network/My Pathway is a graphical representation of the molecular relationships between molecules. Molecules are represented as nodes and the biological relationship between two nodes is represented as an edge (line). All edges are supported by at least one reference from the literature, from a textbook, or from canonical information stored in the Ingenuity® Knowledge Base. Human, mouse, and rat orthologs of a gene are stored as separate objects in the Ingenuity® Knowledge Base but are represented as a single node in the network. Nodes are displayed using various shapes that represent the functional class of the gene product.

A total of 773 significant inflammation genes identified in the computational system were subjected to an Inflammatory Response Inflammation overlay to determine which genes in the significant inflammation leading set were directly involved in inflammation according to IPA (Table 5.2). A total of 890 significant fibrosis genes were subjected to an Organismal Injury and Abnormalities Fibrosis overlay to determine which genes in the significant fibrosis leading set were directly involved in fibrosis according to IPA (Table 5.1). To determine the interactions between genes which have only been experimentally observed in the lung, the Build-Trim tool of IPA was used. Direct and indirect interactions were trimmed to a Confidence Level of Experimentally Observed, and Tissue & Cell Lines were trimmed to Organ Systems of Lung and Cell Line as Lung Cancer Cell Line.

5.1.2 Cell Culture

Small airway epithelial cells (SAEC) were cultured in SABM media (Lonza) supplemented with a SingleQuot Kit (Lonza). Cells were maintained at 37 °C with 5%

Table 5.1: mRNA genes that were found to be significantly related to the time series dose 80 µg fibrosis pathology. Genes were filtered by IPA to be involved in fibrosis.

ACTC1	ADORA1	ADORA2B	ADORA3	ADRA2A
ARG1	BDKRB2	BMPR2	C3	CCL17
CCL2	CCL8	CCR1	CEBPB	CXCL10
CXCL12	EDNRB	EIF2C1	EPHA2	F11
FAS	FCGR2B	FLT3	FN1	GSK3B
HIF1A	HMGCR	HMGCS1	HPX	IGF1
IL11	IL12B	IL1B	IL1R1	IL1RN
IL2RA	IL5	IL6	IRF7	LGALS3
LYVE1	MDK	MMP12	MMP13	MMP14
MX1	MYD88	OAS2	OSM	PDE3A
PDPN	PLA2G10	PROC	PTGIR	PTGS2
PTK2	S100A4	SELE	SELP	SMAD4
SMURF2	SOCS1	SSTR4	THBS1	TIMP1
TNF	TNFAIP3	TNFRSF1B	VEGFA	

Table 5.2: mRNA genes that were found to be significantly related to the time series dose 40 µg inflammation pathology. Genes were filtered by IPA to only those involved in inflammation.

ABCC1	ADORA2B	ADORA3	AGT	APP
C3AR1	CARD11	CCL2	CCL4	CCL5
CD14	CD44	CD48	CD86	CEBPB
CORT	CTSD	CTSS	CXCL12	CYBA
EGFR	FCER1G	FCGR2B	FN1	GHRL
GJA1	ICOS	IGF1	IKBKG	IL12B
IL1B	IL1R1	IL1RN	IL21R	IL23A
IL24	IL2RA	IL6	ITGB2	JUNB
MC2R	MMP9	MYD88	NFKBIA	OLR1
OSM	PBK	PGF	PLA2G10	PLA2G7
PTGER3	PTGS1	PTGS2	RIPK3	SELP
SLC11A1	SOCS1	SOD2	SPHK1	SPP1
THBS1	TNF	TNFAIP3	TNFRSF4	TNFRSF9
TNFSF10	VCAM1			

CO2.

Enzyme-Linked Immunosorbent Assay (ELISA)

SAEC were plated at 60,000 cells per well in a 24-well dish and grown at 37 °C for 48 hours. Cells were serum starved overnight followed by exposure to 1 µg/ml or 2.5 µg/ml MWCNT for 24 hours. Conditioned media were collected and assayed for vascular endothelial growth factor A (*vegfa*) and C-C motif chemokine 2 (*ccl2*) protein expression levels using DuoSet ELISA Development Systems from R & D Systems (Minneapolis, MN) according to manufacturer's protocol. Statistical analysis was done using a two-sample t-test assuming unequal variances.

Cellular RNA isolation

RNA was isolated from SAEC using RNAprotect Cell Reagent and an RNeasy Mini Kit from Qiagen according to the manufacturer's protocol (Qiagen, Valencia, CA). RNA concentrations were determined using a NanoDrop 1000 Spectrophotometer (NanoDrop Technologies, Wilmington, DE) and RNA quality was assessed using an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA).

Real-Time Polymerase Chain Reaction

Total RNA (1 µg) was converted into complementary DNA (cDNA) using a High Capacity cDNA Reverse Transcription Kit from Applied Biosystems (Life Technologies, Carlsbad, CA). All quantitative real-time PCR (qRT-PCR) reactions were performed on a 7500 Real-Time PCR system from Applied Biosystems. Each treatment group consisted of three biological replicates. qRT-PCR analysis for each biological replicate was performed in triplicate, and the Ct values obtained were normalized to the *18s* housekeeping gene. Validated gene expression assays from Applied Biosystems

were employed to carry out the mRNA expression profiling. The following gene expression assays were used: *vegfa* (`Hs00900055_m1`); *ccl2* (`Hs00234140_m1`); and *18s* (`Hs99999901_s1`). Thermal cycling conditions were as follows: 50 °C for 2 minutes, 95 °C for 10 minutes, followed by 40 cycles of 95 °C for 15 seconds and 60 °C for 10 minutes.

5.1.3 Results

Identification of biological processes with expression patterns resembling MWCNT-induced inflammation or fibrosis pathology

The MEGPath system was used to identify genes and biological processes with transcriptional activities, which matched the observed pathological patterns of lung inflammation or fibrotic collagen in the alveolar wall in the MWCNT-exposed mice. The Gene Identification step found 2,996 unique probes which were significantly up-regulated or down-regulated using Significance Analysis of Microarrays (SAM) or a linear model showing significant dose-response or dose and time interactions. Using this set of 2,996 genes, quantitative BAL and pathological data of MWCNT-induced inflammation or quantitative morphometric analysis of fibrosis were used as input patterns to find gene coefficients for reconstruction of the gene expression. Specifically, results for three sets of data were found, two sets relating to fibrosis (morphometrically determined changes in collagen within the alveolar wall) [83] and one relating to inflammation (BAL) [96]. Pathology data for fibrosis at dose 80 µg across the four time points was fitted as an input pattern. The computational system found 69 total significant leading sets, the subset of genes that was used to compute the Functional Process Evaluation (FPE) score, representing the level of correlation with the fibrosis morphometric data for each biological process in the databases. Morphometric data

for fibrosis occurring on day 56 across four doses was fit in the computational system with 85 significant leading sets found. Lastly, inflammation BAL scores at dose 40 µg across four time points was used, and 111 leading sets were found to be significantly correlated with the inflammation pattern.

Example results for each of the pathology data are shown in Figure 5.1. The average of the mRNA expression of genes in the leading set closely resembled the pathology data, indicating that in general, the transcriptional activities of the leading set genes correlated with changes in the pathology. The leading set Reactome Hemostsasis (Figure 5.1E) was found in the C2 Canonical Pathways database and consisted of 147 genes. The leading sets of Immune System Process (Figure 5.1C) and Response To External Stimulus (Figure 5.1F) were found in the C5 database and consisted of 163 genes and 103 genes. *Ccl2* (Figure 5.1A) was contained in the leading set of Immune System Process. Although the *ccl2* expression does not exactly follow the pattern, the average of all gene expression in the leading set does. The same can be seen for *vegfa* (Figure 5.1B). Importantly, our computational system does not constrain genes to being in only one leading set, allowing for genes to be involved in multiple processes. For instance, *ccl2* was found to be involved in both MWNCT-induced fibrosis (Figure 5.1F) and inflammation (Figure 5.1D).

Determination of genes functionally involved in inflammation and fibrosis

To determine which genes were significantly altered in response to MWCNT exposure, leading set genes which attained a fold change of 1.5-fold or greater were input into IPA to determine if they were functionally involved in inflammation or fibrosis according to currently accepted literature.

The inflammation and fibrosis biological processes consisted of 773 and 890 unique genes, respectively, identified to be significantly altered ($\geq$ 1.5 fold change) after

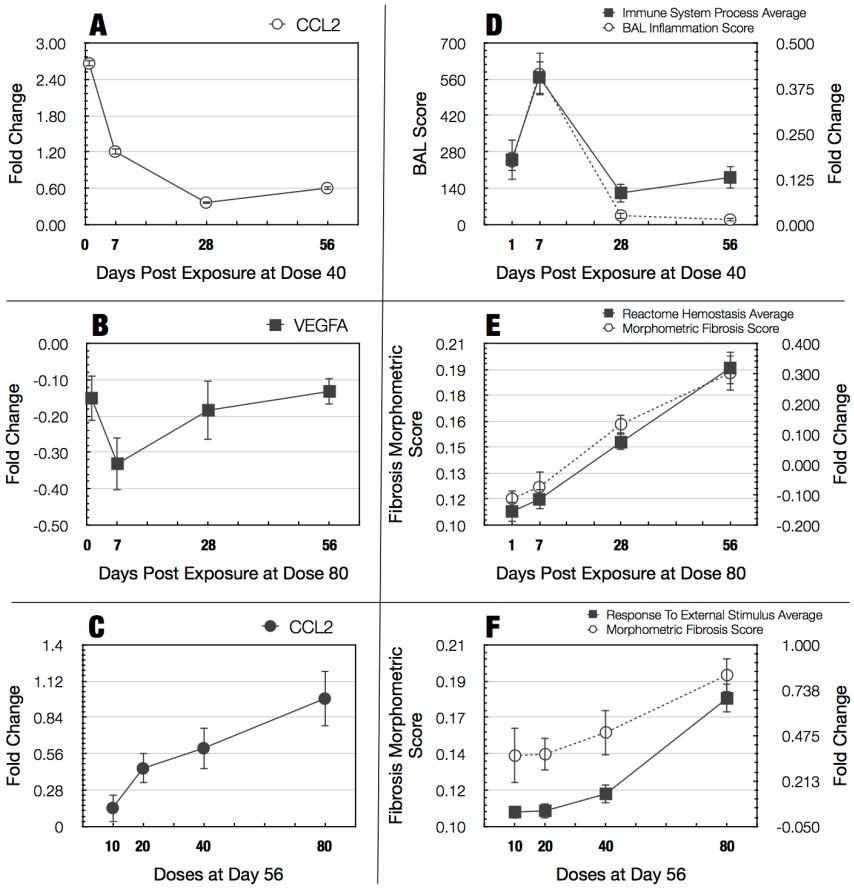

Figure 5.1: Three leading sets found to be significant in a search of the C5 and C2 Canonical Pathway databases using pathological data. Computations were based on the observed experimental data points only; lines have been added to emphasize the patterns used in the computational system. For each pathway, (D) Immune System Process , (E) Reactome Hemostasis, and (F) Response to External Stimulus, the average of all the genes in the leading set shows strong similarity to the pathology data. Expression fold change values are shown for *ccl2*, which was found in the leading sets in (D) and (F), at (C) day 56 and (A) dose 40. *Vegfa*, found in the leading set from (E), fold change is shown in (B).

MWCNT exposure (significant inflammation) with a false discovery rate (FDR) of 1% in SAM analysis. Of the 773 significant inflammation genes, 67 were determined to be directly involved in inflammation by IPA (Table 5.2). Of the 890 significant fibrosis genes, 69 were determined to be directly involved in fibrosis by IPA (Table 5.1).

A heat map of gene expression for the 67 significant inflammation genes (Figure 5.2) suggested the up- and down-regulation of multiple genes in response to MWCNT exposure. Expression of *c3ar1*, *fcgr2b*, *pbk*, *pla2g10*, *il2ra*, *il1rn*, *ptgs1*, *cd14*, *igf1*, *ccl2*, *ccl4*, *il1b*, *pla2g7*, *tnfrsf4*, *ghrl*, *slc11a1*, *tnfaip3*, *cd44*, *adora2b*, *gja1*, *tnf*, *ptgs2*, *junb*, *cd86*, *cyba*, *fcer1g*, *ripk3*, and *socs1* was up-regulated on all days at almost all doses. Expression of *itgb2*, *icos*, *il12b*, *ctss*, *ctsd*, *cd48*, and *il21r* was down-regulated at day 1 but increased in expression at almost all doses on days 7 and 28 and all doses on day 56. Expression of *fn1*, *osm*, *selp*, *thbs1*, *pgf*, *tnfsf9*, *adora3*, *il23a*, *myd88*, *il1r1*, *sod2*, *cebpb*, and *nfkbia* was up-regulated at all doses on day 1 with a decrease in expression over time and down-regulation at most doses on day 56. *Spp1* was highly up-regulated on all days, particularly at doses 40 and 80 µg, while *il6* was highly up-regulated on Day 1 and had a sustained increase in expression over time. Expression of *ptger3*, *ikbkg*, *cxcl12*, *ccl5*, *tnfsf10*, *card11*, *il24*, *mc2r*, *cort*, *mmp9*, *vcam1*, *agt*, *sphk1*, *app*, *egfr*, and *abcc1* was down-regulated across all days at most doses.

Of the 69 significant fibrosis genes (Figure 5.3), *il1rn*, *lgals3*, *pla2g10*, *ccl17*, *adra2a*, *cxcl12*, *fcgr2b*, *s100a4*, *igf1*, *mx1*, *ccl8*, *arg1*, *mmp13*, *il1b*, *sele*, *hpx*, *timp1*, *ccl2*, *adora2b*, *hmgcr*, *hmgcs1*, *tnfaip3*, *tnfrsf1b*, *adora3*, *c3*, *tnf*, *tpgs2*, and *hif1a* were up-regulated on all days at almost all doses. Expression of *il12b*, *flt3*, *mdk*, *adora1*, and *il2ra* was decreased on day 1 but increased over time, while expression of *pdpn*, *myd88*, *il1r1*, *cebpb*, *mmp14*, *fn1*, *socs1*, *irf7*, *selp*, *osm*, *thbs1*, *oas2*, *ptgir*, and *sstr4* was increased on day 1 and decreased over time. *Il6*, *cxcl10*, *ccr1*, and *mmp12* were

highly expressed on day 1 and remained up-regulated over time, while *fas*, *smad4*, *vegfa*, *eif2c1*, *epha2*, *ptk2*, *gsk3b*, *proc*, *f11*, *lyve1*, *pde3a*, *ednrb*, *bdkrb2*, *actc1*, *bmpr2*, and *smurf2* were down-regulated across all days at almost all doses.

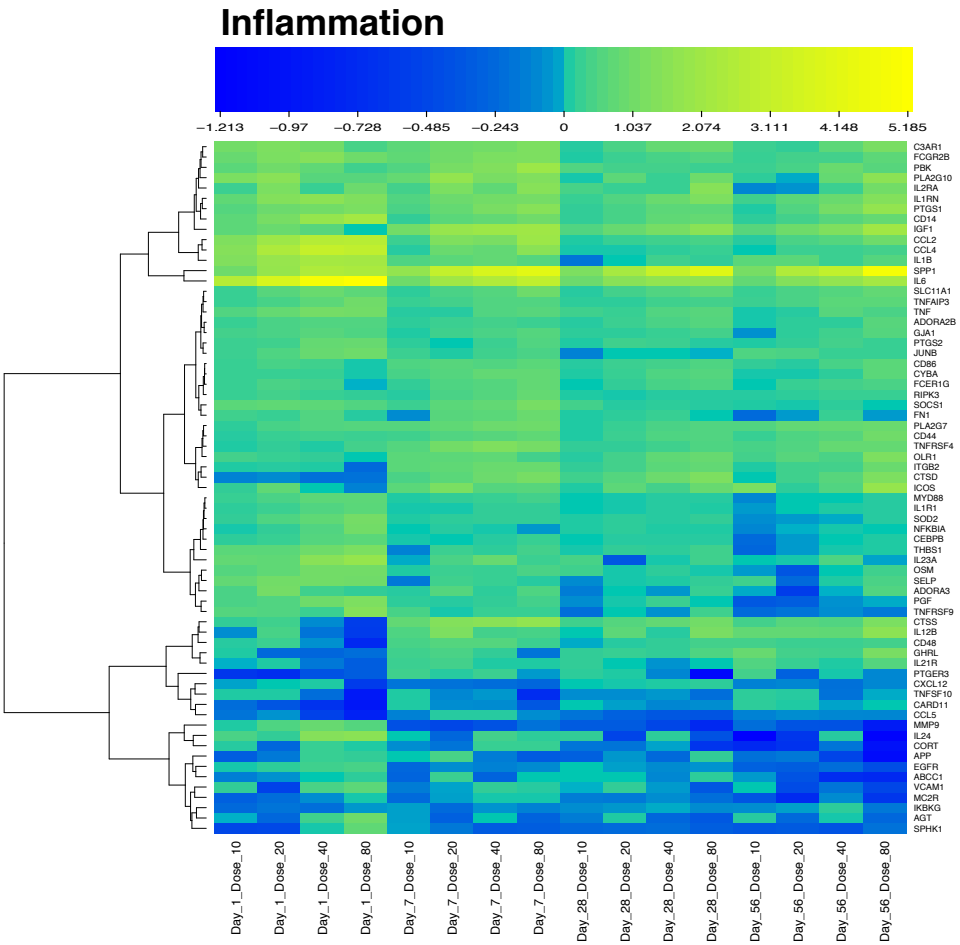

Figure 5.2: Heatmap representation of genes significantly altered above 1.5-fold with an FDR of 1% in SAM analysis in inflammation. *In vivo* gene expression of 67 significant inflammation genes across days 1, 7, 28 and 56 at doses 10, 20, 40, and 80 µg.

Using IPA and these 67 inflammation genes and 69 fibrosis genes, we determined those genes which were significantly involved in IPA Function and Disease Annotations associated with MWCNT-induced fibrosis. A recent report by Mishra et al. [84]

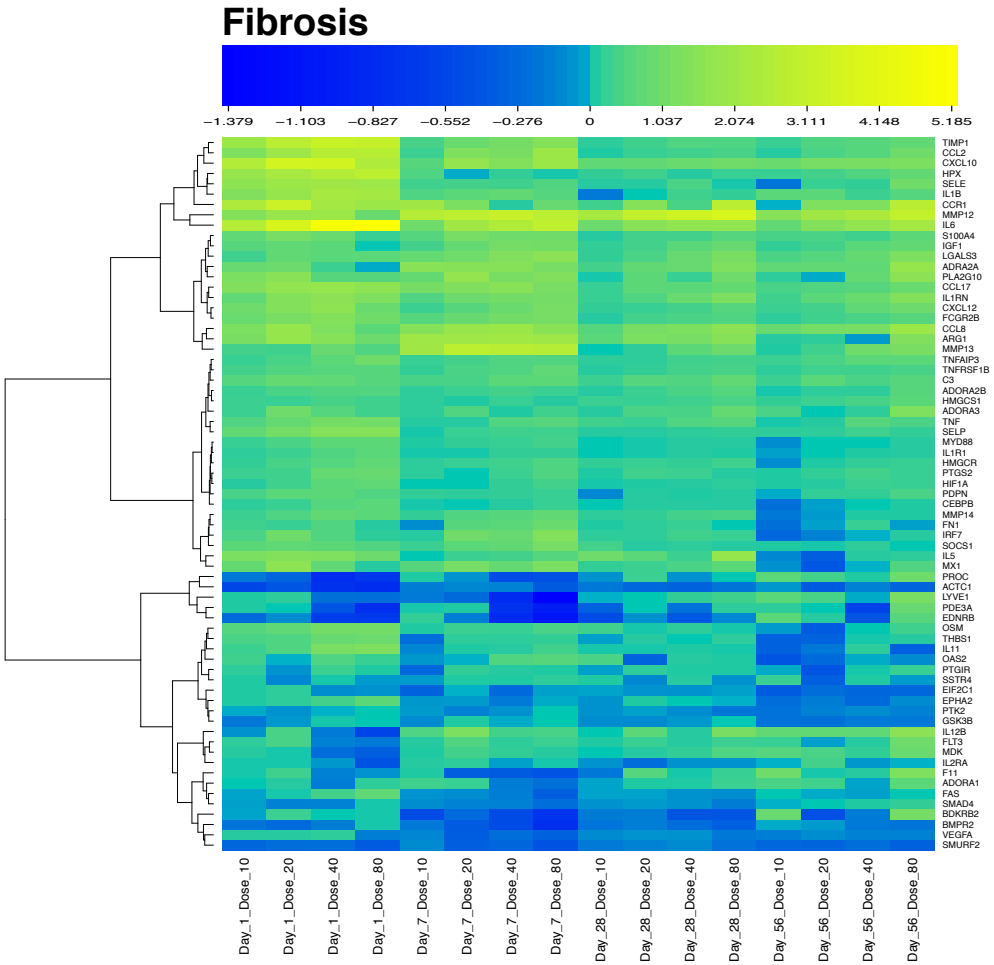

Figure 5.3: Heatmap representation of genes significantly altered above 1.5-fold with an FDR of 1% in SAM analysis in fibrosis. *In vivo* gene expression of 69 significant fibrosis genes across days 1, 7, 28 and 56 at doses 10, 20, 40, and 80 µg

determined that low, physiologically relevant doses of MWCNT equivalent to those in our mouse study could significantly elevate the levels of transforming growth factor β (*tgf-β*) and matrix metalloproteinase-9 (*mmp-9*) in lung epithelial cells, as well as increase mechanisms of collagen production and cellular activation. Therefore, we used IPA to determine which genes in our significant inflammation and fibrosis gene sets were involved in these processes (Tables 5.1 and 5.2). Many inflammation genes were involved in general cell activation by functional association with the IPA function and disease annotations, including Cell Movement, Proliferation of Cells, and Morphology of Cells. Genes found in the significant inflammation set were also involved in the function and disease annotations, including Injury of Lung (*ccl2*, *cd14*, *il6*, *il1r1*, *olr1*, *ptgs1*, *ptgs2*, *selp*, *sphk1*, and *tnf*), Degradation of Connective Tissue (*fn1*, *il6*, *il1b*, *il1rn*, *osm*, *ptgs1*, *ptgs2*, and *tnf*), as well as the signaling pathway *vegf* Signaling (*pgf*). No significant inflammatory genes were found in the *tgf-β* signaling pathway according to IPA. Many fibrosis genes were also involved in the general cell activation function and disease annotations, such as Cell Movement, Proliferation of cells, and Morphology of Cells. Several genes in the significant fibrosis set were involved in the function and disease annotations, including Injury of Lung (*adra2a*, *c3*, *ccl2*, *hif1a*, *il5*, *il6*, *il1r1*, *mmp12*, *ptgs2*, *selp*, *tnf*, and *vegfa*), Degradation of Connective Tissue (*fcgr2b*,*fn1*,*il6*, *il1b*, *il1rn*, *mmp13*, *osm*, *ptgir*, *tnf*, and *tnfrsf1b*), as well as the signaling pathway *vegf* Signaling (*actc1*, *hif1a*, *ptk2*, and *vegfa*). Interestingly, three genes in the significant fibrosis set, *bmpr2*, *smad4*, and *smurf2*, were involved in the IPA *tgf-β* Signaling pathway, again suggesting that *tgf-β* signaling may play an important role in the progression of fibrosis and that the computational system was efficient in determining those biological processes which were functionally related to MWCNT-induced inflammation and fibrosis. An additional analysis of the significant inflammation (Figure 5.4) and fibrosis (Figure 5.5) genes by IPA determined those

genes that have been experimentally shown to have an interaction specifically in the lung.

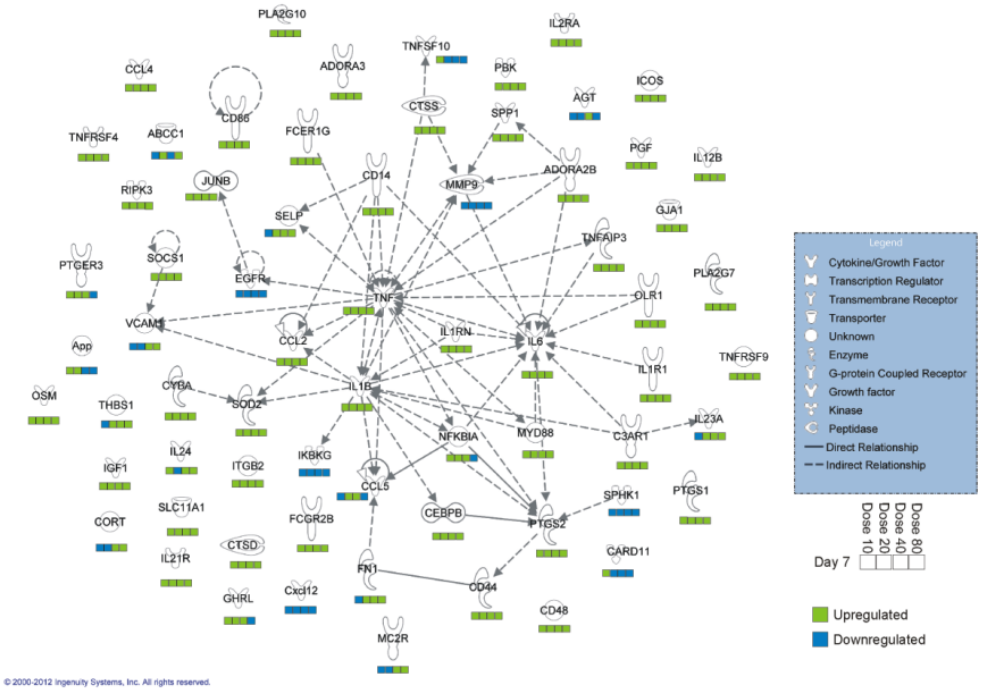

Figure 5.4: IPA analysis of the 67 significant inflammation genes to determine those interactions, which specifically occur in the lung.

5.1.4 *Vegfa* and *ccl2 in vivo* and *in vitro* RNA expression

The 67 inflammation genes and 69 fibrosis genes were ranked by their frequency of inclusion in the biological processes significantly correlated with the pathological data (Figure 5.6). Two genes, *ccl2* and *vegfa*, were selected for *in vitro* validation. *Ccl2* was the top ranked gene that was involved in the most biological processes correlated with the inflammation and among the top 20 genes involved in the most biological processes correlated with the fibrosis. Consistently, in the IPA lung interaction net-

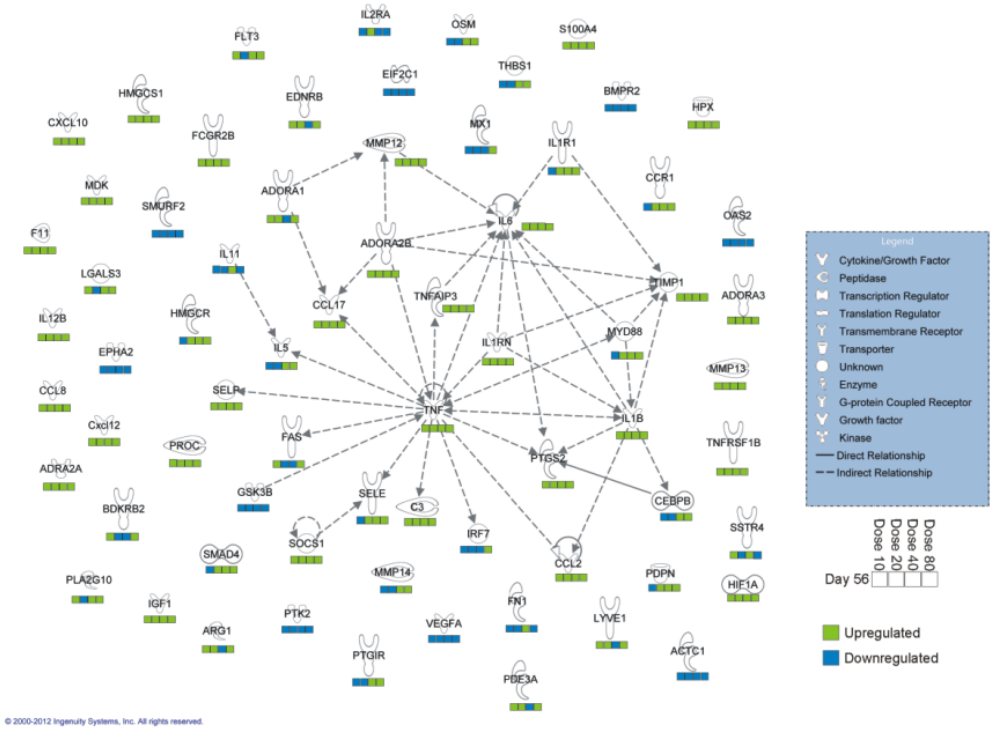

Figure 5.5: IPA analysis of the 69 significant fibrosis genes to determine those interactions, which specifically occur in the lung.

works (Figure 5.4 and 5.5), *ccl2* is in a hub that interacts with both *tnf* and *il1β* hubs in the inflammation and fibrosis networks. *Vegfa* was found to be functionally associated with the fibrosis leading set and is integral in the formation of new blood vessels [38].

Neovascularization is necessary for the formation of fibrotic tissue, and *vegf* has been suggested as a serum biomarker for ranking the severity of idiopathic pulmonary fibrosis [108, 118, 3]. In a separate study, angiogenesis was observed after MWCNT exposure in human endothelial cells and in a coculture of both human epithelial and endothelial cells following epithelial exposure [107]. Based on these results, *ccl2* and *vegfa* were analyzed for their *in vitro* mRNA and protein expression levels following MWCNT exposure to validate the *in vivo* analysis.

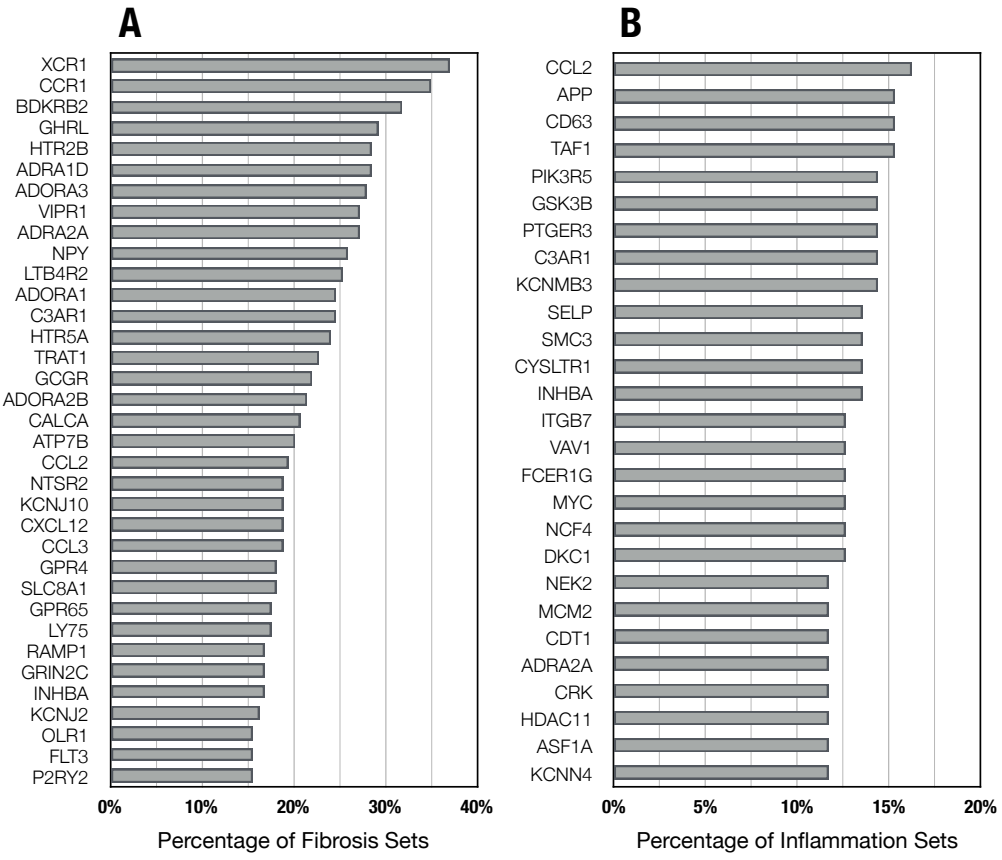

Figure 5.6: (A) Ranking of significant fibrosis genes by their frequency of appearance in biological processes significantly correlated with histopathological data. (B) Ranking of significant inflammation genes by their frequency of appearance in biological processes significantly correlated with histopathological data.

In vivo mRNA levels of *vegfa* showed stable expression levels across all days and doses with a significant decrease in expression on day 56 at dose 40 µg (Figure 5.1B) and closely resembled the time-course of the morphometric collagen score data and leading set average of the biological process Reactome Hemostasis (Figure 5.1E). *Ccl2* showed a consistent dose-dependent increase in mRNA expression on all days with significant increases at all doses on day 1, doses 20, 40, and 80 µg on day 7 and doses 40 and 80 µg on day 56 (Figure 5.1C). *Ccl2 in vivo* mRNA expression data closely resembled the fibrosis day 56 dose-response morphometric analysis and leading set average of biological process Response to External Stimulus (Figure 5.1F) and was similar to the inflammation BAL pattern and leading set average for Immune System Process (Figure 5.1D).

To assess the ability of MWCNT to induce similar RNA expression changes in vitro, SAEC were exposed to MWCNT at either 1 µg/ml (approximately equivalent to the *in vivo* dose of 20-40 µg [96]) or 2.5 µg/ml (approximately equivalent to the *in vivo* dose of 80 µg [96]) for 24 hours, and their mRNA expression levels analyzed. MWCNT exposure at both 1 µg/ml and 2.5 µg/ml exposure levels induced modest but significant increases in *vegfa* mRNA expression *in vitro* in a dose-dependent manner. MWCNT exposure at both 1 µg/ml and 2.5 µg/ml levels induced an increase in *ccl2* mRNA expression with a significant increase at 1 µg/ml.

5.1.5 *Vegfa* and *ccl2 in vitro* protein expression

To determine if the change in *in vitro* mRNA expression levels after exposure to MWCNT resulted in an increase in protein expression, conditioned media from cells exposed to either 1 µg/ml or 2.5 µg/ml MWCNT for 24 h was collected and analyzed by ELISA for *vegfa* and *ccl2* protein expression. *Vegfa* showed significant increases

in protein expression levels over control after 24 h of MWCNT exposure. *Ccl2* also showed significant increases in protein expression levels after 24 h of exposure. This demonstrated that the increase in mRNA expression levels of *vegfa* and *ccl2* after MWCNT exposure *in vitro* resulted in a concordant increase in protein expression and indicated that a similar increase may occur after *in vivo* exposure.

5.1.6 Discussion

Using a novel computational system, the correlation of global mRNA expression profiles to the changes in BAL score and morphometric analysis was analyzed. This identified transcription-related biological processes with expression patterns resembling the pathological patterns of inflammation and fibrosis in MWCNT-exposed mice, allowing for the identification of critical toxicity pathways and potential mechanisms for intervention. The results showed that this systematic analysis could identify relevant genes and pathways in MWCNT-induced lung injury from *in vivo* studies, which were further validated in *in vitro* experiments.

The use of IPA to determine if genes significantly altered in the leading sets were involved in inflammation or fibrosis allowed for an in depth analysis based upon data derived from relationships between genes and disease states taken from the currently accepted literature knowledge base. These analyses were rooted in and verified by experimental results collated from numerous sources. A total of 67 significantly altered genes were determined by IPA to be directly involved in the inflammatory process while 69 significantly altered genes were determine by IPA to be directly involved in fibrosis. Of the significantly altered genes, two genes, *ccl2* and *vegfa*, were chosen to determine their *in vivo* and *in vitro* expression levels due to their roles in the cell during the development of inflammation and fibrosis as well as their rankings during

gene profiling.

The dose-dependent increase in *ccl2* mRNA expression at all days and doses *in vivo* suggests its role in the initial inflammatory process. Although the *in vivo* mRNA levels of *vegfa* remained relatively constant across all days and doses, the *in vivo* protein levels are unknown and may enhance collagen production. *In vitro* levels of *ccl2* and *vegfa* mRNA also increased with increasing dose, reflecting what is seen in the *in vivo* analysis. *In vitro* analysis of the protein levels of *ccl2* and *vegfa* suggests that even modest increases in mRNA levels were able to significantly up-regulate protein expression, and a similar increase in protein expression may occur *in vivo*. The analogous changes to *vegfa* mRNA levels *in vitro*, with subsequent increases in protein levels, suggest that MWCNT may have a similar effect *in vitro* to that seen *in vivo*. This may allow for potentially significant cellular processes to be identified by computational means and for the analysis of the mechanisms and signaling cascades behind MWCNT-induced effects to be validated in an *in vitro* manner.

5.2 Integrated mRNA and miRNA Analysis

mRNA significantly associated with MWCNT-induced inflammatory and fibrotic pathological patterns and functionally involved in lung inflammation and fibrosis in IPA analysis were used to identify miRNA targets and mRNA/miRNA regulatory networks. There were two sets of miRNA used in the analysis. One set consists of miRNA that were significantly changed after MWCNT exposure and found by IPA to be functionally involved in inflammation and fibrosis. The second set of miRNA were associated with the pathological patterns [35]. Both sets were used in the integrated miRNA/mRNA analysis. Potential mRNA targets of significant miRNA were identified using the miRTarBase [56], miRecords [131], and TargetScan [73]

databases. To be considered as a potential mRNA/miRNA target pair, the expression profiles of each mRNA/miRNA pair needed to have a differing second derivative in at least one time point, indicating a divergence in expression during miRNA regulated post-transcriptional activities. The integrated pathways of identified miRNAs and mRNAs were then analyzed and visualized with IPA. The results are provided in Tables 5.3-5.5, listing miRNA/mRNA regulations identified through our analysis that were either experimentally confirmed (Column 3 in Tables 5.3-5.5) or predicted as highly conserved target pairs (Column 4 in Tables 5.3-5.5). In comparison, the miRNA-mediated regulations retrieved from the IPA database are listed in the last column in Tables 5.3-5.5. It is worth noting that the regulations stored in the IPA database may or may not overlap with the results identified with our algorithms.

5.2.1 Inflammation pathology

In total, 16 miRNA were associated with the inflammatory pathology, five of which (*mir-147-3p*, *mir-188-5p*, *mir-327*, *mir-449a-5p*, and *mir-696*) were significantly up- or down-regulated after MWCNT exposure. Four miRNAs, *mir-147-3p*, *mir-449a-5p*, *mir-494-3p*, and *mir-92a-2-5p*, had predicted or experimentally confirmed mRNA targets that were also associated with the inflammatory pathology (Table 5.4). The integrated inflammation pathway based upon miRNAs and mRNAs in Table 5.4 is shown in Figure 5.7. The identified relationships highlighted with a red solid/dash line were not available in the IPA database. Specifically, a regulatory relationship between *mir-147-3p* and *vegfa* identified in our analysis was experimentally confirmed [135] (designated as a solid red line in Figure 5.7). Our analysis also predicted highly conserved target pairs between *mir-92a-2-5p* and *mcl1* and *wnt5a* (dashed red lines in Figure 5.7). The gene expression direction (up- or down-regulation relative to control)

Table 5.3: miRNAs associated with the fibrotic pathology and their experimentally confirmed and predicted mRNA binding partners. Significantly changed miRNAs are highlighted in bold.

miRNA	**Experimentally Confirmed (miRTarBase, miRecords)**	**TargetScan Predicted (highly conserved)**	**IPA Relationships**
let-7c-5p	AGO1 [53], PTK2 [53]	IL6, RGS16, TNFAIP3, TNFRSF1B	FAS, HBEGF, IL6 [112], RGS16, TNFRSF1B
miR-205-5p	VEGFA [135]	IL1R1, PTX3	IL1R1, SMAD4
miR-23b-3p	SMAD4 [99]	EDNRB, FAS, GSK3B, IL11	FAS, IL11, GSK3B, TNFAIP3
miR-31-5p	HIF1A [104], SELE [109]	HBEGF	EDNRB, IL1R1 HBEGF
miR-326-3p	AGO1 [53]	RASSF1	RASSF1, TNFAIP3
miR-328-3p	AGO1 [53]		TNFRSF1B
miR-330-3p	AGO1 [53], VEGFA [135]	RASSF1	BMPR2
miR-34c-3p		PDGFRA, SERPINE1, SMAD4	
miR-375-3p	KCNN4 [124]	BMPR2, EDNRB, RGS16	BMPR2, RGS16
miR-455-3p	AGO1 [53]		
miR-652-3p	AGO1 [53]		
miR-92b-3p			BMPR2 [15], EDNRB

Table 5.4: miRNA associated with the inflammatory pathology and their experimentally confirmed and predicted mRNA binding partners. Significantly changed miRNA are highlighted in bold.

miRNA	**Experimentally Confirmed (miRTarBase, miRecords)**	**TargetScan Predicted (highly conserved)**	**IPA Relationships**
miR-1224-5p			
miR-147-3p	VEGFA [135]		
miR-188-5p			
miR-290a-5p			
miR-327			
miR-3474			
miR-380-3p			
miR-449a-5p		GNAI2, SERPINE1	
miR-494-3p		TNFRSF9	SERPINE1, WNT5A
miR-551b-3p			
miR-667-3p			
miR-696			
miR-703			
miR-877-5p			
miR-881-5p			
miR-92a-2-5p		MCL1, WNT5A	

at doses 10, 20, 40 or 80 µg MWCNT on post-exposure Day 7 are shown for each mRNA and miRNA. This time point was chosen because quantitative bronchoalveolar lavage scores from post-exposure Day 7 [83, 96] were the peak of the inflammation pathology.

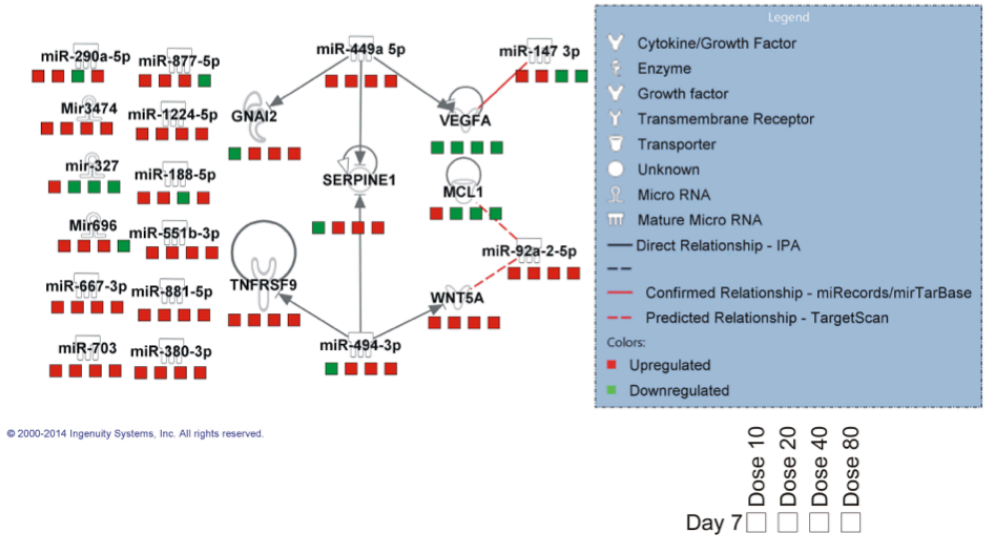

Figure 5.7: Regulatory network of mRNAs and miRNAs transcriptionally related to the dose 40 post-exposure bronchoalveolar lavage inflammatory pathological pattern.

5.2.2 Fibrosis pathology

Among the 12 miRNAs associated with the fibrotic patterns, two miRNAs (*mir-31-5p* and *mir-328-3p*) were significantly (FDR < 5 %; SAM analysis) up- or down-regulated after MWCNT exposure (Table 5.3). All 12 miRNAs had at least one mRNA target as based upon the miRTarBase, miRecords, TargetScan, or IPA databases (Table 5.3). The integrated fibrotic pathway based upon miRNAs and mRNAs in Table 5.3 is shown in Figure 5.8. In addition to the functional relationships found by

IPA, our system identified experimentally confirmed regulations between *ago1* and *let-7c-5p*, *mir-455-3p*, *mir-652-3p*, *mir-326-3p*, *mir-328-3p*, and *mir-330-3p* [53] and between *vegfa* and *mir-330-3p* [135]. Our system predicted highly conserved target pairs between *ednrb* and *mir-375-3p* and *mir-23b-3p*; *rassf1* and *mir-330-3p*; and *mir-34c-3p* and *pdgfr*, *smad4*, and *serpine1*. The gene expression direction (up- or down-regulation relative to control) at doses 10, 20, 40 or 80 µg MWCNT at post-exposure Day 56 are shown for each mRNA and miRNA. This time point was chosen because quantitative morphometric analysis of Sirius Red staining for collagen at Day 56 [83] were the peak of the fibrosis pathology.

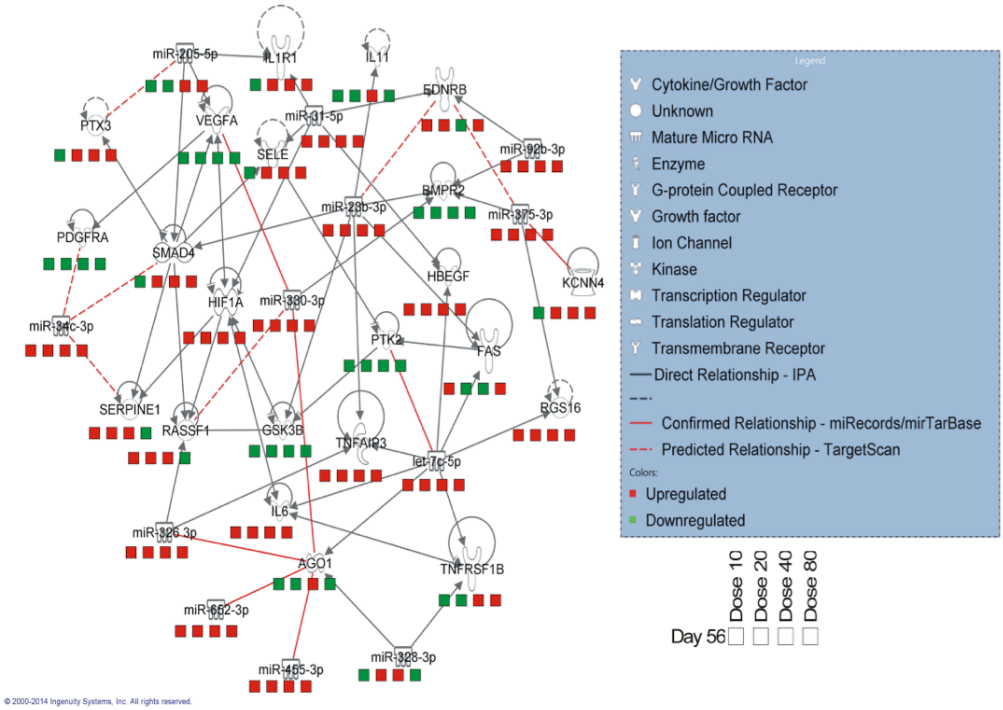

Figure 5.8: Regulatory network of mRNAs and miRNAs transcriptionally related to the dose 80 Sirius Red staining fibrotic pathological pattern.

5.2.3 Ingenuity Pathway Analysis

Next, we used the set of mRNAs associated with the MWCNT-induced fibrotic pathological patterns and functionally involved with fibrosis in IPA analysis to identify their miRNA regulators. A total of 10 miRNAs were identified as potentially involved in MWCNT-induced fibrosis (Table 5.5). Among them, seven miRNAs (*mir-125b-5p*, *mir-126a-3p*, *mir-16-5p*, *mir-199a-5p*, *mir-21-5p*, *mir-30c-5p*, and *mir-322*) had a significant (FDR < 5 %; SAM analysis) expression change after MWNCT exposure. All 10 miRNAs had at least one mRNA target based upon the miRTarBase, miRecords, TargetScan, or IPA databases. The integrated fibrotic pathway for the miRNAs and mRNAs listed in Table 5.3 is shown in Figure 5.9. In addition to the functional relationships retrieved with IPA, our analysis identified numerous post-transcriptional regulations that have been previously experimentally confirmed, including regulatory relationships between *mir-125-5p* and *il1rn* [54], *tnf* [122], *tnfaip3* [68], and *mmp13* [132]; *mir-126a-3p* and *vegfa* [144]; *mir-18a-3p* and *hif1a* [50], *smad4* [29], and *hmgcs1* [50]; *mir-26a-5p* and *egr1* [21], *smad4* [30], *gsk3b* [87], *il6* [134], *ago1* [53]; *mir-21a-5p* and *fas* [100], *bmpr2* [94], *egfr* [143], *plat*, *ptx3*, *tnfaip3*, *ccr1* [117], *vegfa* [79], *mmp9* [89], *ptk2* , *arid4a* [43], among many others listed in Table 5.5 (highlighted with solid red lines in Figure 5.9). Our analysis predicted highly conserved target pairs between *mir-18a-3p* and *igf1* and *tnfaip3*; *mir-322-5p* and *ptgs2*, *vegfa*, *kcnn4*, *cxcl1*, and *smurf2*; *mir-30c-5p* and *ednrb*; and *mir-26a-5p* and *ptx3* (designated by dashed red lines in Figure 5.9). The gene expression direction (up- or down-regulation relative to control) at doses 10, 20, 40, or 80 µg MWCNT at post-exposure Day 56 are shown for mRNAs and miRNAs in Figure 5.9.

These results indicate that our algorithms could identify miRNA-mediated post-transcriptional regulations in MWCNT-treated mice, either experimentally confirmed

Table 5.5: miRNAs associated with fibrosis based upon IPA analysis and their experimentally confirmed and predicted mRNA binding partners. Significantly changed miRNAs are highlighted in bold.

miRNA	**Experimentally Confirmed (miRTarBase, miRecords)**	**TargetScan Predicted (highly conserved)**	**IPA Relationships**
miR-125-5p	IL1RN [54], MMP13 [132], TNF [122], TNFAIP3 [68]	BMPR2, SMAD4, TNFRSF1B, VEGFA ,	BMPR2, SMAD4, TNFRSF1B TNFRSF1B
miR-126a-3p	VEGFA [144]		
miR-141-3p		CSF3	EGFR, HMGCS1, TNFAIP3
miR-16-5p	EGFR [103], KNCC4 [103], SMURF2 [14], VIM [103], VEGFA	CXCL1	AGO1, IGF1, PTGS2 [103]
miR-18a-5p	HIF1A [50], HMGCS1 [50], SMAD4 [29]	IGF1, TNFAIP3	
miR-199a-5p	HIF1A [98], SMAD4 [142]	GSK3B, VEGFA, SERPINE1	AGO1, EDNRB, GSK3B, VEGFA, SERPINE1
miR-21-5p	ARID4A [43],CCR1 [117], BMPR2 [94], EGFR [143], MMP9 [89], PLAT [117], PTK2 [43], PTX3 [117], FAS [100], VEGFA [79], TNFAIP3 [117]		SMURF2, TNF [140]
miR-26a-5p	AGO1 [53], EGR1 [21] GSK3B [87],IL6 [134] SMAD4 [30]	MMP14, PTGS2, PTX3	IGF1
miR-30c-5p	AGO1 [53], SOCS1 [141], VIM [14]	ARID4A, ACTC1, IGF1 EDNRB, IGF1	ARID4A ACTC1, IGF1, SERPINE1
mir-322			CX3CL1, KCNN4, PTGS2, VEGFA, SMURF2

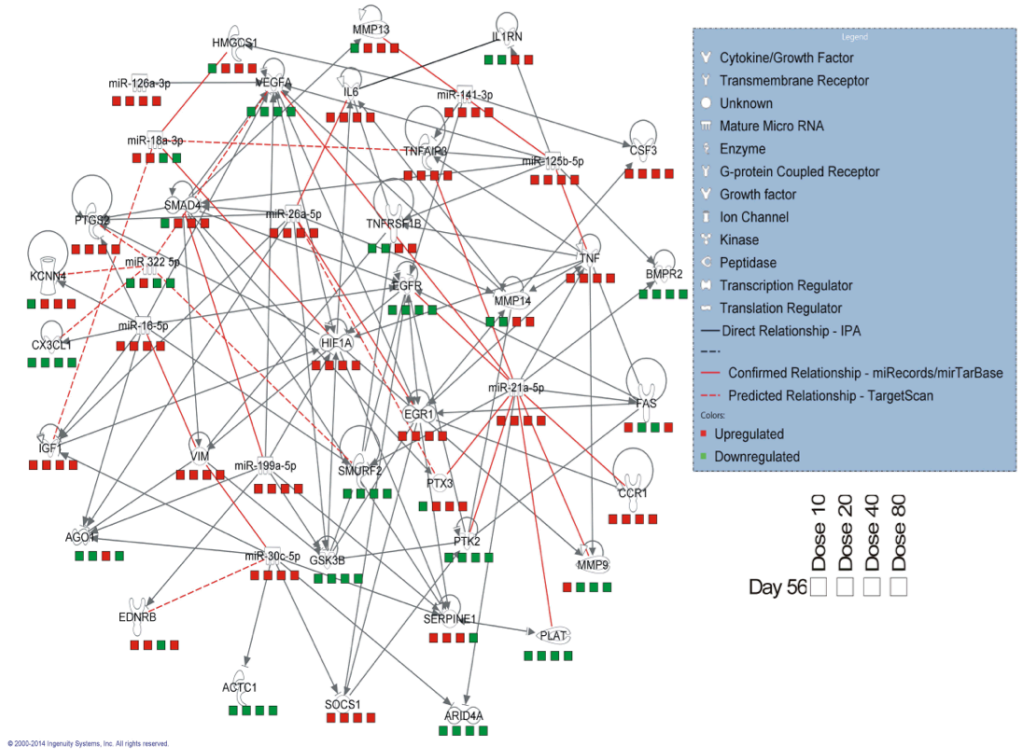

Figure 5.9: Regulatory network of mRNAs transcriptionally related to the dose 80 Sirius Red staining fibrotic pathological pattern and miRNAs experimentally validated to be involved in fibrosis according to the Ingenuity®Knowledge Base.

interactions or highly conserved target pair predictions, many of which were not available in the Ingenuity®Knowledge Base. The identified miRNAs and mRNAs were significantly associated with MWCNT-induced pathological patterns and/or significantly up- or down-regulated in the mouse lung following MWCNT exposure, indicating their potential involvement in pathogenesis and utility as biomarkers for disease. The integrated pathway analysis of miRNA and mRNA and the revealing of their regulatory interactions further elucidated their functional roles in molecular disease mechanisms.

5.2.4 Potential Signaling Pathways in MWCNT-induced Lung Inflammation and Fibrosis

To determine signaling pathways potentially involved in the inflammatory and fibrotic pathological responses to MWCNT exposure, all functionally related mRNAs and miRNAs were analyzed by IPA using a Core Analysis. The top five canonical pathways significant to the mRNAs and miRNAs found in the integrated inflammation analysis (Table 5.4) were Axonal Guidance Signaling, IL-6 Signaling, Corticotropin Releasing Hormone Signaling, Ovarian Cancer Signaling, and Hepatic Fibrosis/Hepatic Stellate Cell Activation. The top five canonical pathways significant to the mRNAs and miRNA found in the integrated fibrosis analysis using the NMF algorithm (Table 5.3) were Hepatic Fibrosis/Hepatic Stellate Cell Activation; Role of Osteoblasts, Osteoclasts, and Chondrocytes in Rheumatoid Arthritis; NF-κB Signaling; Role of Macrophages, Fibroblasts, and Endothelial Cells in Rheumatoid Arthritis; and HMGB1 Signaling. The top five canonical pathways significant to the mRNAs and miRNAs found in the integrated fibrosis (IPA) analysis (Table 5.5) were Hepatic Fibrosis/Hepatic Stellate Cell Activation; Role of Osteoblasts, Osteoclasts, and Chondrocytes in Rheumatoid Arthritis; Colorectal Cancer Metastasis Signaling; ILK Signaling; and Granulocyte Adhesion and Diapedesis. These results suggest that the miRNA and mRNA regulatory networks identified in our analysis are reflective of the inflammatory response and tissue remodeling biological processes during the onset and progression of fibrosis following MWCNT exposure.

5.2.5 Discussion

After we used the second derivative analysis of expression profiles to identify potential regulatory miRNA/mRNA pairs, mRNA and miRNA functional relationships

were assessed through the Ingenuity®Knowledge Base, as well as by the miRTarBase [56], miRecords [131], and TargetScan [73] databases. The Ingenuity®Knowledge Base is based upon the collated findings of patient phenotypes and disease, cellular, molecular, and sequence mechanisms, and all connections between miRNAs and mRNAs are supported by at least one reference in the scientific literature. The TargetScan database provided predicted regulatory mRNA/miRNA pairs, whereas both miRTarBase and miRecords provided a mix of published, experimentally validated, and predicted mRNA and miRNA pairings. In our analysis, only those binding relationships considered to be highly conserved by TargetScan were considered to be predicted targets in our analysis (Tables 5.45.5). IPA considers all relationships, both highly and poorly conserved; therefore, additional predicated relationships were found through the TargetScan database using IPA that were not identified through our second derivative analysis. It is worth noting that *vegfa* was experimentally confirmed in the fibrosis analysis to be an experimentally confirmed target of the tumor suppressor miRNA *mir-205-5p* [130] and the oncogenic miRNA *mir-330-3p* [135]. In our previous study, *vegfa* was predicted to be involved in both MWCNT-induced lung inflammation and fibrosis using both *in vivo* and *in vitro* mRNA and protein assays [37]. The interactions with these potential miRNA regulators could provide new insights into post-transcriptional regulatory mechanisms involved in MWCNT-induced lung angiogenesis, inflammation, and fibrosis. In addition, *mir-23b-3p* was experimentally confirmed as regulator of *smad4* [99], and *mir-34c-3p* was identified as a potential regulator of *smad4* in MWCNT-induced fibrosis (Table 5.3), revealing potential involvement of the *tgf-β* signaling pathway.

The regulatory networks of mRNA and miRNA determined by both IPA analysis and predicted binding due to sequence similarity in this study give a detailed view of potential regulatory networks and signaling pathways involved in MWCNT-induced

inflammation and fibrosis. Core analysis of the significant integrated inflammatory mRNA and miRNA detected pathways involved cytoskeletal remodeling, acute-phase and stress responses, cell adhesion and growth, and the accumulation of ECM proteins. Core analysis of the integrated fibrotic mRNA and miRNA (both NFM and IPA) detected pathways involved in the accumulation of ECM protein, chronic inflammation, innate and acquired immunity, cellular transcription and metastasis, integrin/ECM signaling, and leukocyte migration, suggesting that the miRNA and mRNA regulatory networks determined by our analysis are reflective of the inflammatory response and tissue remodeling that takes place during MWCNT exposure and the onset of fibrosis.

5.3 Publications

The work described in this chapter has been published in two journal manuscripts both as joint first author with Dr. Brandi Synder-Talkington.

J. Dymacek, B. N. Snyder-Talkington, D. W. Porter, M. G. Wolfarth, R. R. Mercer, M. Pacurari, J. Denvir, V. Castranova, Y. Qian, and N. L. Guo. System based identification of toxicity pathways associated with multi-walled carbon nanotube-induced pathological responses. *Toxicol Appl Pharmacol*, 272(2):476 89, Oct 2013.

J. Dymacek, B. N. Snyder-Talkington, D. W. Porter, R. R. Mercer, M. G. Wolfarth, V. Castranova, Y. Qian, and N. L. Guo. mrna and mirna regulatory networks reflective of multi-walled carbon nanotube-induced lung inflammatory and fibrotic pathologies in mice. *Toxicol Sci*, Dec 2014.

Chapter 6

Summary and Future Work

6.1 Summary

This research focuses on developing algorithms to identify genes that are functionally related to our bodies' response to disease. As genomic medical research evolves, we continue to explore whether genes can predict disease or prescribe a cure. By identifying gene response to certain stimuli, it may be possible to develop a standard method for early prognosis of a disease and personalized treatment. Potential applications include early detection of lung disease in factory workers, early detection of lung cancer, or early prognosis of chemotherapy drug treatment.

This thesis outlines a computational system for finding novel hypotheses about involvement of diseases, processes, and functions from a combination of pathological data, gene annotations, miRNA/mRNA regulatory information, and time series dose response microarray data. The use of matrix factorization, optimization, and randomized algorithms allows the computational system to reduce calculations from days to hours while still running on a standard laptop. This research is a small step towards successful personalized medicine.

Analyzing time series microarray data is a difficult task; however, temporal information is useful for discovering functional mechanisms and causal relationships. My system utilizes a non-negative matrix factorization (NMF) algorithm for extracting underlying basis patterns in the gene expression data. These patterns represent underlying functions guiding gene response. The patterns are then related to prior biological knowledge in the form of known annotated pathways and relationships between genes. The NMF algorithm incorporates pathological data and discovers related sets of genes which are significantly correlated to the pathology.

This algorithm finds biologically relevant pathways and genes with and without pathological information. It has been used on genome-wide expression profiles of mouse lungs following aspiration of well dispersed multi-walled carbon nanotubes (MWCNT), and has detected MWCNT-induced lung inflammation, lung fibrosis, and related pathways. The identified significant pathways and genes are supported by evidence in the literature and by biological validation (both *in vitro* and *in vivo*).

MWCNT are an important class of engineered nano-materials with broad applications in many industries [91]. Concerns over potential MWCNT-induced toxicity have emerged, particularly due to the structural similarity between asbestos and MWCNT [31]. Previous studies have shown that MWCNT induce lung damage, including inflammatory granulomas and substantial interstitial lung fibrosis [83, 96]. Pulmonary fibrosis has a poor clinical outcome [13, 40] and may be a potential precursor to lung cancer [137]. However, there are no clinically applicable biomarkers for early detection and no effective treatment for pulmonary fibrosis due to its late diagnosis and poorly understood molecular mechanisms for initiation [55, 127].

Recently, there is emerging interest in exploring miRNAs as potential therapeutic targets and biomarkers for diagnosis and prognosis. Advantages of miRNA biomarkers include their presence in various bodily fluids [18, 85] and greater stability in

prepared tissue samples, including formalin fixation, relative to mRNA [62]. The use of miRNA markers in the selection of appropriate treatments has the possibility for improving patient outcomes by determining the best application of existing drugs. Additionally, identifying miRNA biomarkers may aid in the development of novel treatments through the elucidation of new pathways [7, 61]. Our computational system could aid in the process of identifying miRNA biomarkers and in practice has helped identify several miRNA for further studies.

Results from this system were used as preliminary findings in a successful NIH R01 grant application. ("Systematic assessment of multi-walled carbon nanotubes in pulmonary disease" 1R01ES021764-01 Guo (PI) NIEHS/NHLBI $1,665,000)

Articles

J. Dymacek and N. L. Guo. Systems approach to identifying relevant pathways from phenotype information in dose-dependent time series microarray data. In *Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine, BIBM '11*, pages 290293, Atlanta, GA, USA, 2011. IEEE Computer Society.

J. Dymacek and N. L. Guo. Integrated mirna and mrna analysis of time series microarray data. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pages 122127, Newport Beach, CA, USA, 2014. ACM.

J. Dymacek, B. N. Snyder-Talkington, D. W. Porter, M. G. Wolfarth, R. R. Mercer, M. Pacurari, J. Denvir, V. Castranova, Y. Qian, and N. L. Guo. System based identification of toxicity pathways associated with multi-walled carbon nanotube-induced pathological responses. *Toxicol Appl Pharmacol*, 272(2):476 89, Oct 2013.

J. Dymacek, B. N. Snyder-Talkington, D. W. Porter, R. R. Mercer, M. G. Wolfarth, V. Castranova, Y. Qian, and N. L. Guo. mrna and mirna regulatory networks reflective of multi-walled carbon nanotube-induced lung inflammatory and fibrotic pathologies in mice. *Toxicol Sci*, Dec 2014.

Posters

Dymacek, Julian et al. Identifying Significant Biological Processes from Pathological Information and mRNA Microarray Data. Poster at 27th Annual Meeting of the Allegheny-Erie Society of Toxicology. 2013.

Support

NSF, "IGERT: Research and Education in Nanotoxicity" training grant, 2012-2014

WVNano, "Cancer, Energy and Security Nanotechnology STEM Graduate Education" training grant, 2011-2012

6.2 Limitations

The MEGPath system is not suited for modeling experimental data from experiments with less than three conditions. The system runs in $O(n * m^3)$ time for n genes and m conditions meaning it is not ideal for experiments with high numbers of conditions. Also, unlike parametric systems, no information can be implied from unobserved experimental conditions. While the system suggests that two doses may behave the same, it can make no claim about unobserved middle doses. The system also depends on curated gene sets, which depending on granularity, may be too general. The use of second derivatives in the miRNA analysis detects rudimentary regulatory relationships but implies a linear relationship over time or doses.

6.3 Future Work

Time Series Data Across Experiments

My collaborators at the National Institute of Occupational Safety and Health have recently completed a year long MWCNT lung exposure experiment. This current ex-

periment is a logical successor to the 56 day experiment I have worked on. Combining microarray data from both experiments will provide a new, longer term data set useful for the community at large. Unfortunately, changes in microarray technology and slight experimental differences make a straight forward combination difficult.

Gold Standard Data Set

A new mock "gold" standard data set would be useful for future exploration and comparison of similar systems. The new data set would need to have underlying functions influencing genes and underlying networks connecting genes. In addition the networks would need to be described as gene sets.

Better miRNA Integration

miRNA are a regulatory mechanism of mRNA. An integrated analysis of miRNA and the mRNA targets provides a deeper understanding of the biological mechanisms of disease response. There is the potential to relate miRNA back to the mRNA annotated gene sets.

Distributed Implementation

A distributed implementation of the MEGPath matrix factorization algorithm can reduce computation time from hours to minutes. There are many opportunities for exploring distributed computing, parallel algorithms, and cloud computing. Obtaining the probability density functions are trivially parallel.

Alternate Distance Measures

Currently, the Frobenius norm is used for distance calculations and for updating the probability density functions in the MEGPath system. Possible improvements

include the potential for limited sampling to find covariance. There is potential to both increase accuracy and speed with direct tie-ins to a distributed implementation.

Bibliography

[1] C. A. Afshari, H. K. Hamadeh, and P. R. Bushel. The evolution of bioinformatics in toxicology: Advancing toxicogenomics. *Toxicological Sciences*, 120(suppl 1):S225–S237, 2011.

[2] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, 97(18):10101–6, Aug 2000.

[3] M. Ando, E. Miyazaki, T. Ito, S. Hiroshige, S.-i. Nureki, T. Ueno, R. Takenaka, T. Fukami, and T. Kumamoto. Significance of serum vascular endothelial growth factor level in patients with idiopathic pulmonary fibrosis. *Lung*, 188(3):247–252, 2010.

[4] B. Andreopoulos, A. An, X. Wang, and M. Schroeder. A roadmap of clustering algorithms: finding a match for a biomedical application. *Briefings in Bioinformatics*, 10(3):297–314, 2009.

[5] I. Androulakis, E. Yang, and R. Almon. Analysis of time-series gene expression data: Methods, challenges, and opportunities. *Annual Review of Biomedical Engineering*, 9(1):205–228, 2007.

[6] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, 25(1):25–29, May 2000.

[7] R. Avraham and Y. Yarden. Regulation of signalling by microrna. *Biochem Soc Trans*, 40(1):26–30, Feb 2012.

[8] D. Baek, J. Villén, C. Shin, F. D. Camargo, S. P. Gygi, and D. P. Bartel. The impact of microrna on protein output. *Nature*, 455(7209):64–71, Sep 2008.

[9] S. Banerjee, N. Xie, H. Cui, Z. Tan, S. Yang, M. Icyuz, E. Abraham, and G. Liu. Microrna let-7c regulates macrophage polarization. *J Immunol*, 190(12):6542–9, Jun 2013.

[10] Z. Bar-Joseph. Analyzing time series gene expression data. *Bioinformatics*, 20(16):2493–2503, 2004.

[11] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300, 1995.

[12] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, and J. Galon. Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, 25(8):1091–3, Apr 2009.

[13] J. A. Bjoraker, J. H. Ryu, M. K. Edwin, J. L. Myers, H. D. Tazelaar, D. R. Schroeder, and K. P. Offord. Prognostic significance of histopathologic subsets in idiopathic pulmonary fibrosis. *Am J Respir Crit Care Med*, 157(1):199–203, Jan 1998.

[14] J. Bockhorn, K. Yee, Y.-F. Chang, A. Prat, D. Huo, C. Nwachukwu, R. Dalton, S. Huang, K. E. Swanson, C. M. Perou, O. I. Olopade, M. F. Clarke, G. L. Greene, and H. Liu. Microrna-30c targets cytoskeleton genes involved in breast cancer cell invasion. *Breast Cancer Res Treat*, 137(2):373–82, Jan 2013.

[15] M. Brock, M. Trenkmann, R. E. Gay, B. A. Michel, S. Gay, M. Fischler, S. Ulrich, R. Speich, and L. C. Huber. Interleukin-6 modulates the expression of the bone morphogenic protein receptor type ii through a novel stat3-microrna cluster 17/92 pathway. *Circ Res*, 104(10):1184–91, May 2009.

[16] L. D. Burgoon, Q. Ding, A. N'jai, E. Dere, A. R. Burg, J. C. Rowlands, R. A. Budinsky, K. E. Stebbins, and T. R. Zacharewski. Automated dose-response analysis of the relative hepatic gene expression potency of tcdf in c57bl/6 mice. *Toxicol Sci*, 112(1):221–8, Nov 2009.

[17] R. Chavez-Alvarez, A. Chavoya, and A. Mendez-Vazquez. Discovery of possible gene relationships through the application of self-organizing maps to dna microarray databases. *PLoS One*, 9(4):e93233, 2014.

[18] X. Chen, Z. Hu, W. Wang, Y. Ba, L. Ma, C. Zhang, C. Wang, Z. Ren, Y. Zhao, S. Wu, R. Zhuang, Y. Zhang, H. Hu, C. Liu, L. Xu, J. Wang, H. Shen, J. Zhang, K. Zen, and C.-Y. Zhang. Identification of ten serum microrna from a genome-wide serum microrna expression profile as novel noninvasive biomarkers for non-small cell lung cancer diagnosis. *Int J Cancer*, 130(7):1620–8, Apr 2012.

[19] C. Cheng and L. M. Li. Inferring microrna activities by combining gene expression with microrna target prediction. *PLoS One*, 3(4):e1989, 2008.

[20] Y. Cheng and G. M. Church. Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol*, 8:93–103, 2000.

[21] S. W. Chi, J. B. Zang, A. Mele, and R. B. Darnell. Argonaute hits-clip decodes microrna-mrna interaction maps. *Nature*, 460(7254):479–86, Jul 2009.

[22] J. Chou and P. Bushel. Discernment of possible mechanisms of hepatotoxicity via biological processes over-represented by co-expressed genes. *BMC Genomics*, 10(1):272, 2009.

[23] J. W. Chou, T. Zhou, W. K. Kaufmann, R. S. Paules, and P. R. Bushel. Extracting gene expression patterns and identifying co-expressed genes from microarray data reveals biologically responsive processes. *BMC Bioinformatics*, 8:427, 2007.

[24] T. H. Cormen, C. Stein, R. L. Rivest, and C. E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2nd edition, 2001.

[25] D. Croft, A. F. Mundo, R. Haw, M. Milacic, J. Weiser, G. Wu, M. Caudy, P. Garapati, M. Gillespie, M. R. Kamdar, B. Jassal, S. Jupe, L. Matthews, B. May, S. Palatnik, K. Rothfels, V. Shamovsky, H. Song, M. Williams, E. Birney, H. Hermjakob, L. Stein, and P. D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Res*, 42(Database issue):D472–7, Jan 2014.

[26] S. Das, M. Kumar, V. Negi, B. Pattnaik, Y. S. Prakash, A. Agrawal, and B. Ghosh. Microrna-326 regulates profibrotic functions of transforming growth factor- in pulmonary fibrosis. *Am J Respir Cell Mol Biol*, 50(5):882–92, May 2014.

[27] D. Dembélé and P. Kastner. Fuzzy c-means method for clustering microarray data. *Bioinformatics*, 19(8):973–80, May 2003.

[28] K. Devarajan. Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS Comput Biol*, 4(7):e1000029, 07 2008.

[29] M. Dews, J. L. Fox, S. Hultine, P. Sundaram, W. Wang, Y. Y. Liu, E. Furth, G. H. Enders, W. El-Deiry, J. M. Schelter, M. A. Cleary, and A. Thomas-Tikhonenko. The myc-mir-17 92 axis blunts tgfbeta signaling and production of multiple tgfbeta-dependent antiangiogenic factors. *Cancer Res*, 70(20):8233–46, Oct 2010.

[30] B. K. Dey, J. Gagan, Z. Yan, and A. Dutta. mir-26a is required for skeletal muscle differentiation and regeneration in mice. *Genes Dev*, 26(19):2180–91, Oct 2012.

[31] K. Donaldson, R. Aitken, L. Tran, V. Stone, R. Duffin, G. Forrest, and A. Alexander. Carbon nanotubes: a review of their properties in relation to pulmonary toxicology and workplace safety. *Toxicol Sci*, 92(1):5–22, Jul 2006.

[32] J. Dong, G. Jiang, Y. W. Asmann, S. Tomaszek, J. Jen, T. Kislinger, and D. A. Wigle. Microrna networks in mouse lung organogenesis. *PLoS One*, 5(5):e10854, 2010.

[33] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley, New York, 2. edition, 2001.

[34] J. Dymacek and N. L. Guo. Systems approach to identifying relevant pathways from phenotype information in dose-dependent time series microarray data. In *Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine*, BIBM '11, pages 290–293, Washington, DC, USA, 2011. IEEE Computer Society.

[35] J. Dymacek and N. L. Guo. Integrated mirna and mrna analysis of time series microarray data. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '14, pages 122–127, New York, NY, USA, 2014. ACM.

[36] J. Dymacek, B. N. Snyder-Talkington, D. W. Porter, R. R. Mercer, M. G. Wolfarth, V. Castranova, Y. Qian, and N. L. Guo. mrna and mirna regulatory networks reflective of multi-walled carbon nanotube-induced lung inflammatory and fibrotic pathologies in mice. *Toxicol Sci*, Dec 2014.

[37] J. Dymacek, B. N. Snyder-Talkington, D. W. Porter, M. G. Wolfarth, R. R. Mercer, M. Pacurari, J. Denvir, V. Castranova, Y. Qian, and N. L. Guo. System-based identification of toxicity pathways associated with multi-walled carbon nanotube-induced pathological responses. *Toxicol Appl Pharmacol*, 272(2):476–89, Oct 2013.

[38] N. Ferrara and T. Davis-Smyth. The biology of vascular endothelial growth factor. *Endocrine reviews*, 18(1):4–25, 1997.

[39] A. F. Filipsson, S. Sand, J. Nilsson, and K. Victorin. The benchmark dose method–review of available models, and recommendations for application in health risk assessment. *Crit Rev Toxicol*, 33(5):505–42, 2003.

[40] K. R. Flaherty, W. D. Travis, T. V. Colby, G. B. Toews, E. A. Kazerooni, B. H. Gross, A. Jain, R. L. Strawderman, A. Flint, J. P. Lynch, and F. J. Martinez. Histopathologic variability in usual and nonspecific interstitial pneumonias. *Am J Respir Crit Care Med*, 164(9):1722–7, Nov 2001.

[41] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel. Most mammalian mrnas are conserved targets of microrna. *Genome Res*, 19(1):92–105, Jan 2009.

[42] A. Frigyesi, S. Veerla, D. Lindgren, and M. Hoglund. Independent component analysis reveals new and biologically significant structures in micro array data. *BMC Bioinformatics*, 7(1):290, 2006.

[43] G. Gabriely, T. Wurdinger, S. Kesari, C. C. Esau, J. Burchard, P. S. Linsley, and A. M. Krichevsky. Microrna 21 promotes glioma invasion by targeting matrix metalloproteinase regulators. *Mol Cell Biol*, 28(17):5369–80, Sep 2008.

[44] R. Gaujoux and C. Seoighe. A flexible r package for nonnegative matrix factorization. *BMC Bioinformatics*, 11:367, 2010.

[45] R. Gaujoux and C. Seoighe. Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infect Genet Evol*, 12(5):913–21, Jul 2012.

[46] D. Ghosh and A. M. Chinnaiyan. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18(2):275–286, 2002.

[47] D. Greenbaum, C. Colangelo, K. Williams, and M. Gerstein. Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol.*, 4(9):117, 2003.

[48] S. Griffiths-Jones, R. J. Grocock, S. van Dongen, A. Bateman, and A. J. Enright. mirbase: microrna sequences, targets and gene nomenclature. *Nucleic Acids Res*, 34(Database issue):D140–4, Jan 2006.

[49] N. L. Guo, Y.-W. Wan, J. Denvir, D. W. Porter, M. Pacurari, M. G. Wolfarth, V. Castranova, and Y. Qian. Multiwalled carbon nanotube-induced gene signatures in the mouse lung: potential predictive value for human lung cancer risk and prognosis. *J Toxicol Environ Health A*, 75(18):1129–53, 2012.

[50] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, Jr, A.-C. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl. Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip. *Cell*, 141(1):129–41, Apr 2010.

[51] S. Halappanavar, P. Jackson, A. Williams, K. A. Jensen, K. S. Hougaard, U. Vogel, C. L. Yauk, and H. Wallin. Pulmonary response to surface-coated nanotitanium dioxide particles includes induction of acute phase response genes, inflammatory cascades, and changes in microrna: a toxicogenomic study. *Environ Mol Mutagen*, 52(6):425–39, Jul 2011.

[52] H. K. Hamadeh, M. Todd, L. Healy, J. T. Meyer, A. M. Kwok, M. Higgins, and C. A. Afshari. Application of genomics for identification of systemic toxicity triggers associated with vegf-r inhibitors. *Chemical Research in Toxicology*, 23(6):1025–1033, 2010.

[53] A. Helwak, G. Kudla, T. Dudnakova, and D. Tollervey. Mapping the human mirna interactome by clash reveals frequent noncanonical binding. *Cell*, 153(3):654–65, Apr 2013.

[54] M. H. Hofmann, J. Heinrich, G. Radziwill, G. Radziwil, and K. Moelling. A short hairpin dna analogous to mir-125b inhibits c-raf expression, proliferation, and survival of breast cancer cells. *Mol Cancer Res*, 7(10):1635–44, Oct 2009.

[55] R. J. Homer, J. A. Elias, C. G. Lee, and E. Herzog. Modern concepts on the role of inflammation in pulmonary fibrosis. *Arch Pathol Lab Med*, 135(6):780–8, Jun 2011.

[56] S.-D. Hsu, Y.-T. Tseng, S. Shrestha, Y.-L. Lin, A. Khaleel, C.-H. Chou, C.-F. Chu, H.-Y. Huang, C.-M. Lin, S.-Y. Ho, T.-Y. Jian, F.-M. Lin, T.-H. Chang, S.-L. Weng, K.-W. Liao, I.-E. Liao, C.-C. Liu, and H.-D. Huang. mirtarbase update 2014: an information resource for experimentally validated mirna-target interactions. *Nucleic Acids Res*, 42(Database issue):D78–85, Jan 2014.

[57] D. Huang and W. Pan. Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics*, 22(10):1259–1268, 2006.

[58] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res*, 37(1):1–13, Jan 2009.

[59] D. W. Huang, B. T. Sherman, Q. Tan, J. R. Collins, W. G. Alvord, J. Roayaei, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. The david gene functional classification tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*, 8(9):R183, 2007.

[60] G. T. Huang, C. Athanassiou, and P. V. Benos. mirconnx: condition-specific mrna-microrna network integrator. *Nucleic Acids Res*, 39(Web Server issue):W416–23, Jul 2011.

[61] M. V. Iorio and C. M. Croce. microrna involvement in human cancer. *Carcinogenesis*, 33(6):1126–33, Jun 2012.

[62] M. Jung, A. Schaefer, I. Steiner, C. Kempkensteffen, C. Stephan, A. Erbersdobler, and K. Jung. Robust microrna stability in degraded rna preparations from human tissue and cell samples. *Clin Chem*, 56(6):998–1006, Jun 2010.

[63] C. Jutten and J. Herault. Blind separation of sources, part 1: An adaptive algorithm based on neuromimetic architecture. *Signal Process.*, 24(1):1–10, Aug. 1991.

[64] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, Jan 2000.

[65] B.-Y. Kang, S. Ko, and D.-W. Kim. Sicago: Semi-supervised cluster analysis using semantic distance between gene pairs in gene ontology. *Bioinformatics*, 26(10):1384–1385, 2010.

[66] P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*, 8(2):e1002375, 2012.

[67] H. Kim and H. Park. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, 23(12):1495–1502, 2007.

[68] S.-W. Kim, K. Ramasamy, H. Bouamar, A.-P. Lin, D. Jiang, and R. C. T. Aguiar. Micrornas mir-125a and mir-125b constitutively activate the nf-b pathway by targeting the tumor necrosis factor alpha-induced protein 3 (tnfaip3, a20). *Proc Natl Acad Sci U S A*, 109(20):7865–70, May 2012.

[69] W. Kong, C. R. Vanderburg, H. Gunshin, J. T. Rogers, and X. Huang. A review of independent component analysis application to microarray gene expression data. *Biotechniques*, 45(5):501–20, Nov 2008.

[70] A. K. Kopec, L. D. Burgoon, D. Ibrahim-Aibo, A. R. Burg, A. W. Lee, C. Tashiro, D. Potter, B. Sharratt, J. R. Harkema, J. C. Rowlands, R. A. Budinsky, and T. R. Zacharewski. Automated dose-response analysis and comparative toxicogenomic evaluation of the hepatic effects elicited by tcdd, tcdf, and pcb126 in c57bl/6 mice. *Toxicol Sci*, 118(1):286–97, Nov 2010.

[71] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[72] S.-I. Lee and S. Batzoglou. Application of independent component analysis to microarrays. *Genome Biol*, 4(11):R76, 2003.

[73] B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microrna targets. *Cell*, 120(1):15–20, Jan 2005.

[74] M. Li, B. Wang, Z. Momeni, and F. Valafar. Pattern recognition techniques in microarray data analysis. In *Proc Int Conf on Mathematics and Engineering Techniques in Medicine and Biological Sciences 2002 (METMBS'02), Las Vegas*, pages 610–616, 2002.

[75] J. C. Liao, R. Boscolo, Y.-L. Yang, L. M. Tran, C. Sabatti, and V. P. Roychowdhury. Network component analysis: reconstruction of regulatory signals in biological systems. *Proc Natl Acad Sci U S A*, 100(26):15522–7, Dec 2003.

[76] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–40, Jun 2011.

[77] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, Jan 2002.

[78] L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some micrornas downregulate large numbers of target mrnas. *Nature*, 433(7027):769–73, Feb 2005.

[79] L.-Z. Liu, C. Li, Q. Chen, Y. Jing, R. Carpenter, Y. Jiang, H.-F. Kung, L. Lai, and B.-H. Jiang. Mir-21 induced angiogenesis through akt and erk activation and hif-1 expression. *PLoS One*, 6(4):e19139, 2011.

[80] S. P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[81] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans Comput Biol Bioinform*, 1(1):24–45, 2004.

[82] S. C. Madeira, M. C. Teixeira, I. Sá-Correia, and A. L. Oliveira. Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *IEEE/ACM Trans Comput Biol Bioinform*, 7(1):153–65, 2010.

[83] R. R. Mercer, A. F. Hubbs, J. F. Scabilloni, L. Wang, L. A. Battelli, S. Friend, V. Castranova, and D. W. Porter. Pulmonary fibrotic response to aspiration of multi-walled carbon nanotubes. *Part Fibre Toxicol*, 8:21, 2011.

[84] A. Mishra, Y. Rojanasakul, B. T. Chen, V. Castranova, R. R. Mercer, and L. Wang. Assessment of pulmonary fibrogenic potential of multiwalled carbon nanotubes in human lung cells. *Journal of Nanomaterials*, 2012:4, 2012.

[85] P. S. Mitchell, R. K. Parkin, E. M. Kroh, B. R. Fritz, S. K. Wyman, E. L. Pogosova-Agadjanyan, A. Peterson, J. Noteboom, K. C. O'Briant, A. Allen, D. W. Lin, N. Urban, C. W. Drescher, B. S. Knudsen, D. L. Stirewalt, R. Gentleman, R. L. Vessella, P. S. Nelson, D. B. Martin, and M. Tewari. Circulating micrornas as stable blood-based markers for cancer detection. *Proc Natl Acad Sci U S A*, 105(30):10513–8, Jul 2008.

[86] M.-H. Mo, L. Chen, Y. Fu, W. Wang, and S. W. Fu. Cell-free circulating mirna biomarkers in cancer. *J Cancer*, 3:432–48, 2012.

[87] J. S. Mohamed, M. A. Lopez, and A. M. Boriek. Mechanical stretch up-regulates microrna-26a and induces human airway smooth muscle hypertrophy by suppressing glycogen synthase kinase-3. *J Biol Chem*, 285(38):29336–47, Sep 2010.

[88] T. D. Moloshok, R. R. Klevecz, J. D. Grant, F. J. Manion, W. F. Speier, and M. F. Ochs. Application of bayesian decomposition for analysing microarray data. *Bioinformatics*, 18(4):566–575, 2002.

[89] T. Moriyama, K. Ohuchida, K. Mizumoto, J. Yu, N. Sato, T. Nabae, S. Takahata, H. Toma, E. Nagai, and M. Tanaka. Microrna-21 modulates biological functions of pancreatic cancer cells including their proliferation, invasion, and chemoresistance. *Mol Cancer Ther*, 8(5):1067–74, May 2009.

[90] M. F. Ochs, L. Rink, C. Tarn, S. Mburu, T. Taguchi, B. Eisenberg, and A. K. Godwin. Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. *Cancer Res.*, 69(23):9125–9132, Dec 2009.

[91] M. Pacurari, V. Castranova, and V. Vallyathan. Single- and multi-wall carbon nanotubes versus asbestos: are the carbon nanotubes a new health risk to humans? *J Toxicol Environ Health A*, 73(5):378–95, 2010.

[92] M. Pacurari, Y. Qian, D. W. Porter, M. Wolfarth, Y. Wan, D. Luo, M. Ding, V. Castranova, and N. L. Guo. Multi-walled carbon nanotube-induced gene expression in the mouse lung: association with lung pathology. *Toxicol Appl Pharmacol*, 255(1):18–31, Aug 2011.

[93] W. Pan, J. Lin, and C. Le. Model-based cluster analysis of microarray gene-expression data. *Genome Biology*, 3(2):research0009.1–research0009.8, 2002.

[94] T. Papagiannakopoulos, A. Shapiro, and K. S. Kosik. Microrna-21 targets a network of key tumor-suppressive pathways in glioblastoma cells. *Cancer Res*, 68(19):8164–72, Oct 2008.

[95] S. D. Peddada, E. K. Lobenhofer, L. Li, C. A. Afshari, C. R. Weinberg, and D. M. Umbach. Gene selection and clustering for time-course and dose-response microarray experiments using order-restricted inference. *Bioinformatics*, 19(7):834–41, May 2003.

[96] D. W. Porter, A. F. Hubbs, R. R. Mercer, N. Wu, M. G. Wolfarth, K. Sriram, S. Leonard, L. Battelli, D. Schwegler-Berry, S. Friend, M. Andrew, B. T. Chen, S. Tsuruoka, M. Endo, and V. Castranova. Mouse pulmonary dose- and time course-responses induced by exposure to multi-walled carbon nanotubes. *Toxicology*, 269(2-3):136–47, Mar 2010.

[97] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical recipes in C (2nd ed.): the art of scientific computing*. Cambridge University Press, New York, NY, USA, 1992.

[98] S. Rane, M. He, D. Sayed, H. Vashistha, A. Malhotra, J. Sadoshima, D. E. Vatner, S. F. Vatner, and M. Abdellatif. Downregulation of mir-199a derepresses hypoxia-inducible factor-1alpha and sirtuin 1 and recapitulates hypoxia preconditioning in cardiac myocytes. *Circ Res*, 104(7):879–86, Apr 2009.

[99] C. E. Rogler, L. Levoci, T. Ader, A. Massimi, T. Tchaikovskaya, R. Norel, and L. E. Rogler. Microrna-23b cluster micrornas regulate transforming growth factor-beta/bone morphogenetic protein signaling and liver stem cell differentiation by targeting smads. *Hepatology*, 50(2):575–84, Aug 2009.

[100] D. Sayed, M. He, C. Hong, S. Gao, S. Rane, Z. Yang, and M. Abdellatif. Microrna-21 is a downstream effector of akt that mediates its antiapoptotic effects via suppression of fas ligand. *J Biol Chem*, 285(26):20281–90, Jun 2010.

[101] A. Schliep, A. Schonhuth, and C. Steinhoff. Using hidden markov models to analyze gene expression time course data. *Bioinformatics*, 19(suppl 1):i255–i263, 2003.

[102] M. H. Schulz, K. V. Pandit, C. L. Lino Cardenas, N. Ambalavanan, N. Kaminski, and Z. Bar-Joseph. Reconstructing dynamic microrna-regulated interaction networks. *Proc Natl Acad Sci U S A*, 110(39):15686–91, Sep 2013.

[103] M. Selbach, B. Schwanhäusser, N. Thierfelder, Z. Fang, R. Khanin, and N. Rajewsky. Widespread changes in protein synthesis induced by micrornas. *Nature*, 455(7209):58–63, Sep 2008.

[104] J. Shen, X. Yang, B. Xie, Y. Chen, M. Swaim, S. F. Hackett, and P. A. Campochiaro. Microrna regulate ocular neovascularization. *Mol Ther*, 16(7):1208–16, Jul 2008.

[105] B. T. Sherman, D. W. Huang, Q. Tan, Y. Guo, S. Bour, D. Liu, R. Stephens, M. W. Baseler, H. C. Lane, and R. A. Lempicki. David knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics*, 8:426, 2007.

[106] J. Shlens. A tutorial on principal component analysis. *CoRR*, abs/1404.1100, 2014.

[107] B. N. Snyder-Talkington, D. Schwegler-Berry, V. Castranova, Y. Qian, and N. L. Guo. Multi-walled carbon nanotubes induce human microvascular endothelial cellular effects in an alveolar-capillary co-culture with small airway epithelial cells. *Part Fibre Toxicol*, 10:35, 2013.

[108] R. M. Strieter. What differentiates normal lung repair and fibrosis? inflammation, resolution of repair, and fibrosis. *Proc Am Thorac Soc*, 5(3):305–10, Apr 2008.

[109] Y. Suárez, C. Wang, T. D. Manes, and J. S. Pober. Cutting edge: Tnf-induced micrornas regulate tnf-induced expression of e-selectin and intercellular adhesion molecule-1 on human endothelial cells: feedback control of inflammation. *J Immunol*, 184(1):21–5, Jan 2010.

[110] N. Subhani, L. Rueda, A. Ngom, and C. J. Burden. Multiple gene expression profile alignment for microarray time-series data clustering. *Bioinformatics*, 26(18):2281–2288, 2010.

[111] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43):15545–50, Oct 2005.

[112] K. Sugimura, H. Miyata, K. Tanaka, R. Hamano, T. Takahashi, Y. Kurokawa, M. Yamasaki, K. Nakajima, S. Takiguchi, M. Mori, and Y. Doki. Let-7 expression is a significant determinant of response to chemotherapy through the regulation of il-6/stat3 pathway in esophageal squamous cell carcinoma. *Clin Cancer Res*, 18(18):5144–53, Sep 2012.

[113] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. S. Lander, and T. R. Golub. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. U.S.A.*, 96(6):2907–2912, Mar 1999.

[114] L. Tari, C. Baral, and S. Kim. Fuzzy c-means clustering with prior biological knowledge. *Journal of Biomedical Informatics*, 42(1):74 – 81, 2009.

[115] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nat. Genet.*, 22(3):281–285, Jul 1999.

[116] A. B. Tchagang, K. V. Bui, T. McGinnis, and P. V. Benos. Extracting biologically significant patterns from short time series gene expression data. *BMC Bioinformatics*, 10:255, 2009.

[117] M. Terao, M. Fratelli, M. Kurosaki, A. Zanetti, V. Guarnaccia, G. Paroni, A. Tsykin, M. Lupi, M. Gianni, G. J. Goodall, and E. Garattini. Induction of mir-21 by retinoic acid in estrogen receptor-positive breast carcinoma cells: biological correlates and molecular targets. *J Biol Chem*, 286(5):4027–42, Feb 2011.

[118] V. J. Thannickal, G. B. Toews, E. S. White, J. P. Lynch Iii, and F. J. Martinez. Mechanisms of pulmonary fibrosis. *Annu. Rev. Med.*, 55:395–417, 2004.

[119] R. S. Thomas, B. C. Allen, A. Nong, L. Yang, E. Bermudez, H. J. Clewell, 3rd, and M. E. Andersen. A method to integrate benchmark dose estimates with genomic data to assess the functional effects of chemical exposure. *Toxicol Sci*, 98(1):240–8, Jul 2007.

[120] R. S. Thomas, H. J. Clewell, 3rd, B. C. Allen, S. C. Wesselkamper, N. C. Y. Wang, J. C. Lambert, J. K. Hess-Wilson, Q. J. Zhao, and M. E. Andersen. Application of transcriptional benchmark dose values in quantitative cancer and noncancer risk assessment. *Toxicol Sci*, 120(1):194–205, Mar 2011.

[121] R. S. Thomas, H. J. Clewell, 3rd, B. C. Allen, L. Yang, E. Healy, and M. E. Andersen. Integrating pathway-based transcriptomic data into quantitative chemical risk assessment: a five chemical case study. *Mutat Res*, 746(2):135–43, Aug 2012.

[122] E. Tili, J.-J. Michaille, A. Cimino, S. Costinean, C. D. Dumitru, B. Adair, M. Fabbri, H. Alder, C. G. Liu, G. A. Calin, and C. M. Croce. Modulation of mir-155 and mir-125b levels following lipopolysaccharide/tnf-alpha stimulation and their possible roles in regulating the response to endotoxin shock. *J Immunol*, 179(8):5082–9, Oct 2007.

[123] L. M. Tran, M. P. Brynildsen, K. C. Kao, J. K. Suen, and J. C. Liao. gnca: a framework for determining transcription factor activity based on transcriptome: identifiability and numerical implementation. *Metab Eng*, 7(2):128–41, Mar 2005.

[124] Y. Tsukamoto, C. Nakada, T. Noguchi, M. Tanigawa, L. T. Nguyen, T. Uchida, N. Hijiya, K. Matsuura, T. Fujioka, M. Seto, and M. Moriyama. Microrna-375

is downregulated in gastric carcinomas and regulates cell survival by targeting pdk1 and 14-3-3zeta. *Cancer Res*, 70(6):2339–49, Mar 2010.

[125] V. G. Tusher, R. Tibshirani, and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

[126] R. Vigário, J. Särelä, V. Jousmäki, M. Hämäläinen, and E. Oja. Independent component approach to the analysis of eeg and meg recordings. *IEEE Trans Biomed Eng*, 47(5):589–93, May 2000.

[127] N. Walter, H. R. Collard, and T. E. King, Jr. Current perspectives on the treatment of idiopathic pulmonary fibrosis. *Proc Am Thorac Soc*, 3(4):330–8, Jun 2006.

[128] X. Wang and X. Wang. Systematic identification of microrna functions by combining target prediction and expression profiling. *Nucleic Acids Res*, 34(5):1646–52, 2006.

[129] Y. Wang, C. Huang, N. Reddy Chintagari, M. Bhaskaran, T. Weng, Y. Guo, X. Xiao, and L. Liu. mir-375 regulates rat alveolar epithelial cell transdifferentiation by inhibiting wnt/-catenin pathway. *Nucleic Acids Res*, 41(6):3833–44, Apr 2013.

[130] H. Wu, S. Zhu, and Y.-Y. Mo. Suppression of cell growth and invasion by mir-205 in breast cancer. *Cell Res*, 19(4):439–48, Apr 2009.

[131] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li. mirecords: an integrated resource for microrna-target interactions. *Nucleic Acids Res*, 37(Database issue):D105–10, Jan 2009.

[132] N. Xu, L. Zhang, F. Meisgen, M. Harada, J. Heilborn, B. Homey, D. Grandér, M. Ståhle, E. Sonkoly, and A. Pivarcsi. Microrna-125b down-regulates matrix metallopeptidase 13 and inhibits cutaneous squamous cell carcinoma cell proliferation, migration, and invasion. *J Biol Chem*, 287(35):29899–908, Aug 2012.

[133] S. Yang, N. Xie, H. Cui, S. Banerjee, E. Abraham, V. J. Thannickal, and G. Liu. mir-31 is a negative regulator of fibrogenesis and pulmonary fibrosis. *FASEB J*, 26(9):3790–9, Sep 2012.

[134] X. Yang, L. Liang, X.-F. Zhang, H.-L. Jia, Y. Qin, X.-C. Zhu, X.-M. Gao, P. Qiao, Y. Zheng, Y.-Y. Sheng, J.-W. Wei, H.-J. Zhou, N. Ren, Q.-H. Ye, Q.-Z. Dong, and L.-X. Qin. Microrna-26a suppresses tumor growth and metastasis of human hepatocellular carcinoma by targeting interleukin-6-stat3 pathway. *Hepatology*, 58(1):158–70, Jul 2013.

[135] W. Ye, Q. Lv, C.-K. A. Wong, S. Hu, C. Fu, Z. Hua, G. Cai, G. Li, B. B. Yang, and Y. Zhang. The effect of central loops in mirna:mre duplexes on the efficiency of mirna-mediated gene regulation. *PLoS One*, 3(3):e1719, 2008.

[136] K. Y. Yeung and W. L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774, 2001.

[137] Y.-Y. Yu, P. F. Pinsky, N. E. Caporaso, N. Chatterjee, M. Baumgarten, P. Langenberg, J. P. Furuno, Q. Lan, and E. A. Engels. Lung cancer risk following detection of pulmonary scarring by chest radiography in the prostate, lung, colorectal, and ovarian cancer screening trial. *Arch Intern Med*, 168(21):2326–32; discussion 2332, Nov 2008.

[138] B. R. Zeeberg, W. Feng, G. Wang, M. D. Wang, A. T. Fojo, M. Sunshine, S. Narasimhan, D. W. Kane, W. C. Reinhold, S. Lababidi, K. J. Bussey, J. Riss, J. C. Barrett, and J. N. Weinstein. Gominer: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*, 4(4):R28, 2003.

[139] S. Zhang, Q. Li, J. Liu, and X. J. Zhou. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules. *Bioinformatics*, 27(13):i401–9, Jul 2011.

[140] S. Zhang, C.-C. Liu, W. Li, H. Shen, P. W. Laird, and X. J. Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res*, 40(19):9379–91, Oct 2012.

[141] X. Zhang, M. Daucher, D. Armistead, R. Russell, and S. Kottilil. Microrna expression profiling in hcv-infected human hepatoma cells identifies potential anti-viral targets induced by interferon-. *PLoS One*, 8(2):e55733, 2013.

[142] X. Zhang, W.-L. Ng, P. Wang, L. Tian, E. Werner, H. Wang, P. Doetsch, and Y. Wang. Microrna-21 modulates the levels of reactive oxygen species by targeting sod3 and tnf. *Cancer Res*, 72(18):4707–13, Sep 2012.

[143] X. Zhou, Y. Ren, L. Moore, M. Mei, Y. You, P. Xu, B. Wang, G. Wang, Z. Jia, P. Pu, W. Zhang, and C. Kang. Downregulation of mir-21 inhibits egfr pathway and suppresses the growth of human glioblastoma cells independent of pten status. *Lab Invest*, 90(2):144–55, Feb 2010.

[144] N. Zhu, D. Zhang, H. Xie, Z. Zhou, H. Chen, T. Hu, Y. Bai, Y. Shen, W. Yuan, Q. Jing, and Y. Qin. Endothelial-specific intron-derived mir-126 is down-regulated in human breast cancer and targets both vegfa and pik3r2. *Mol Cell Biochem*, 351(1-2):157–64, May 2011.