Published in final edited form as:

ACM BCB. 2014 ; 2014: 122–127. doi:10.1145/2649387.2649411.

Integrated miRNA and mRNA Analysis of Time Series Microarray Data

Julian Dymacek and

Mary Babb Randolph Cancer Center West Virginia University Morgantown, WV 26506, USA
jdymacek@mix.wvu.edu

Nancy Lan Guo

Mary Babb Randolph Cancer Center West Virginia University Morgantown, WV 26506, USA
lguo@hsc.wvu.edu

Abstract

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.

1. INTRODUCTION

While dynamic temporal regulatory effects of microRNA (miRNA) are not well known, a miRNA helps regulate many mRNA and therefore act on a multitude of proteins [1, 14]. The expression of over half of the genes in the human genome may be regulated by miRNA [7]. 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 [16].

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-

Copyright is held by the owner/author(s). Publication rights licensed to ACM.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

General Terms
Algorithms

or down-regulation may produce more potential target pairs especially when using predicted miRNA/mRNA pairs [3, 5, 11, 21]. 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 [25, 26] 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 [18]. 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.

2. METHODS

The integrated mRNA and miRNA analysis (Figure 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.

2.1 Non-negative Matrix Factorization

Non-negative matrix factorization [12] (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 to find matrices P and C such that $D = C * P$. 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.

2.2 mRNA Analysis

The MEGPath [6] system was used to identify mRNA genes potentially involved with the pathology. The MEG-Path 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 [20] 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 2.

2.3 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 3). A normal distribution was tested for 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 brosis pathology can be seen in Figure 2.

2.4 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 [10], miRecords [23], and TargetScan [13]. 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 [8] 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 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 $i = 1 \dots (n - 1)$:

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

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:

$$\exists i=0 \dots (n - 1) \quad \text{where} \quad R'_i * M'_i \leq 0.$$

An example of target pairs, between let-7c and three mRNA, found through this analysis can be seen in Figure 2. The target pairs have differing second derivatives in at least one time point.

3. RESULTS AND IMPLEMENTATION

Results were obtained from an *in vivo* dose-response time series multi-wall carbon nano-tube (MWCNT) aspiration exposure experiment [17]. 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.

3.1 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(dn), 10, 20, 40, or 80 μ g 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 [17]. The average thickness of the alveolar connection tissue was used for fibrosis scores. These were found from the morpho-metric analysis of Sirius Red staining for connective tissue [15].

3.2 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 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 *ndDelta* 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.

3.3 mRNA Results

Gene sets significantly related to the time series 80 μ g dose fibrosis pathology 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 [9] were used. The 2,996 significant probes were used with the NMF algorithm to identify three underlying patterns from the four time points. Genome-wide coefficients were then found relating each gene to the fibrosis pathology, coefficient error in

¹<http://www.mwcnttranscriptome.org>

fluence was minimized by identifying sets of genes. The brosis pathology coefficients were used to identify significantly represented sets of genes. Significantly related gene sets were found from the MSigDB curated databases with 30 gene sets found from the C2 database and 39 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 [19].

All genes from the gene sets associated with the brosis pathology were screened through Ingenuity Pathway Analysis (IPA), an online curated literature based tool (ingenuity.com). The 89 mRNA genes related to brosis found by IPA were kept for further analysis and are listed in Table 1.

3.4 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.

The NMF algorithm was run with the 80 μ g dose brosis time series constraint. Three patterns were found over the four time points. Since individual miRNA were being identified, the coefficients relating miRNA to the brosis 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 2.

After identifying the significant miRNA, potential mRNA targets were iterated 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 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.

4. 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 brosis pathology pattern and potential regulators of mRNA identified with brosis. In particular the miRNA let-7c may have implications in lung brosis [2] and was shown, with potential mRNA targets, in Figure 2. Likewise, mir-31 has been shown to be involved in lung brosis regulation [24],

suggesting an active role in attempting to suppress MWCNT caused lung brosis. Other identified miRNA with potential lung brosis involvement are mir-326 [4] and mir-375 [22]. The predicted miRNA /mRNA targets could be validated *in vitro*. Although demonstrated on MWCNT toxicity data, this integrated approach could be used in other applications.

ACKNOWLEDGMENTS

Dr. Nancy L. Guo is supported by NIH R01ES021764; R01/ R56LM009500. Julian Dymacek is supported by an NSF training grant on nanotoxicology. We would like to thank Dr. Brandi Synder-Talkington for input and IPA expertise.

REFERENCES

1. Baek D, Vill en J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of micrnas on protein output. *Nature*. Sep; 2008 455(7209):64–71. [PubMed: 18668037]
2. Banerjee S, Xie N, Cui H, Tan Z, Yang S, Icyuz M, Abraham E, Liu G. Microna let-7c regulates macrophage polarization. *J Immunol*. Jun; 2013 190(12):6542–9. [PubMed: 23667114]
3. Cheng C, Li LM. Inferring microna activities by combining gene expression with microna target prediction. *PLoS One*. 2008; 3(4):e1989. [PubMed: 18431476]
4. Das S, Kumar M, Negi V, Pattnaik B, Prakash YS, Agrawal A, Ghosh B. Microna-326 regulates profibrotic functions of transforming growth factor- in pulmonary fibrosis. *Am J Respir Cell Mol Biol*. May; 2014 50(5):882–92. [PubMed: 24279830]
5. Dong J, Jiang G, Asmann YW, Tomaszek S, Jen J, Kislinger T, Wigle DA. Microna networks in mouse lung organogenesis. *PLoS One*. 2010; 5(5):e10854. [PubMed: 20520778]
6. Dymacek, J.; Guo, NL. Proceedings of the 2011 IEEE International Conference on Bioinformatics and Biomedicine, BIBM'11. IEEE Computer Society; Washington, DC, USA: 2011. Systems approach to identifying relevant pathways from phenotype information in dose-dependent time series microarray data.; p. 290-293.
7. Friedman RC, Farh KK-H, Burge CB, Bartel DP. Most mammalian mrnas are conserved targets of micrnas. *Genome Res*. Jan; 2009 19(1):92–105. [PubMed: 18955434]
8. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. mirbase: microna sequences, targets and gene nomenclature. *Nucleic Acids Res*. Jan; 2006 34(Database issue):D140–4. [PubMed: 16381832]
9. Guo NL, Wan Y-W, Denvir J, Porter DW, Pacurari M, Wolfarth MG, Castranova V, Qian Y. 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*. 2012; 75(18):1129–53. [PubMed: 22891886]
10. Hsu S-D, Tseng Y-T, Shrestha S, Lin Y-L, Khaleel A, Chou C-H, Chu C-F, Huang H-Y, Lin C-M, Ho S-Y, Jian T-Y, Lin F-M, Chang T-H, Weng S-L, Liao K-W, Liao I-E, Liu C-C, Huang H-D. mirtarbase update 2014: an information resource for experimentally validated mirna-target interactions. *Nucleic Acids Res*. Jan; 2014 42(Database issue):D78–85. [PubMed: 24304892]
11. Huang GT, Athanassiou C, Benos PV. mirconnx: condition-specific mrna-microna network integrator. *Nucleic Acids Res*. Jul; 2011 39(Web Server issue):W416–23. [PubMed: 21558324]
12. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. 1999; 401(6755):788–791. [PubMed: 10548103]
13. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often anked by adenosines, indicates that thousands of human genes are microna targets. *Cell*. Jan; 2005 120(1):15–20. [PubMed: 15652477]
14. Lim LP, Lau NC, Garrett-Engle P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. Microarray analysis shows that some micrnas downregulate large numbers of target mrnas. *Nature*. Feb; 2005 433(7027):769–73. [PubMed: 15685193]

15. Mercer RR, Hubbs AF, Scabilloni JF, Wang L, Battelli LA, Friend S, Castranova V, Porter DW. Pulmonary fibrotic response to aspiration of multi-walled carbon nanotubes. Part Fibre Toxicol. 2011; 8:21. [PubMed: 21781304]
16. Mo M-H, Chen L, Fu Y, Wang W, Fu SW. Cell-free circulating mirna biomarkers in cancer. J Cancer. 2012; 3:432–48. [PubMed: 23074383]
17. Porter DW, Hubbs AF, Mercer RR, Wu N, Wolfarth MG, Sriram K, Leonard S, Battelli L, Schwegler-Berry D, Friend S, Andrew M, Chen BT, Tsuruoka S, Endo M, Castranova V. Mouse pulmonary dose- and time course-responses induced by exposure to multi-walled carbon nanotubes. Toxicology. Mar; 2010 269(2-3):136–47. [PubMed: 19857541]
18. Schulz MH, Pandit KV, Lino Cardenas CL, Ambalavanan N, Kaminski N, Bar-Joseph Z. Reconstructing dynamic microRNA-regulated interaction networks. Proc Natl Acad Sci U S A. Sep; 2013 110(39):15686–91. [PubMed: 23986498]
19. Snyder-Talkington BN, Dymacek J, Porter DW, Wolfarth MG, Mercer RR, Pacurari M, Denvir J, Castranova V, Qian Y, Guo NL. System-based identification of toxicity pathways associated with multi-walled carbon nanotube-induced pathological responses. Toxicol Appl Pharmacol. Oct; 2013 272(2):476–89. [PubMed: 23845593]
20. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. Oct; 2005 102(43):15545–50. [PubMed: 16199517]
21. Wang X, Wang X. Systematic identification of microRNA functions by combining target prediction and expression profiling. Nucleic Acids Res. 2006; 34(5):1646–52. [PubMed: 16549876]
22. Wang Y, Huang C, Reddy Chintagari N, Bhaskaran M, Weng T, Guo Y, Xiao X, Liu L. mir-375 regulates rat alveolar epithelial cell trans-differentiation by inhibiting wnt/-catenin pathway. Nucleic Acids Res. 2013; 41(6):3833–44. [PubMed: 23396279]
23. Xiao F, Zuo Z, Cai G, Kang S, Gao X, Li T. mirecords: an integrated resource for microRNA-target interactions. Nucleic Acids Res. Jan; 2009 37(Database issue):D105–10. [PubMed: 18996891]
24. Yang S, Xie N, Cui H, Banerjee S, Abraham E, Thannickal VJ, Liu G. mir-31 is a negative regulator of fibrogenesis and pulmonary fibrosis. FASEB J. Sep; 2012 26(9):3790–9. [PubMed: 22661007]
25. Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics. Jul; 2011 27(13):i401–9. [PubMed: 21685098]
26. Zhang S, Liu C-C, Li W, Shen H, Laird PW, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res. Oct; 2012 40(19):9379–91. [PubMed: 22879375]

Figure 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.

Figure 2.

The dose 80 brosis pathology and signi cantly related let-7c are plotted in solid lines. The brosis pathology was used as a constraint in the NMF algorithm for both miRNA and mRNA. The mRNA in dashed lines were found to be related to brosis by both the pathology and IPA. All mRNA and let-7c were identi ed as target pairs, with differing second derivatives in at least one time point.

Figure 3.

Histogram of the coefficients corresponding to the pathology pattern. Each bar shows the percentage of coefficients in the range ending with the label. Error values were subtracted from coefficients making some less than 0. A normal distribution was t to the data.

Table 1

mRNA genes that were found to be significantly related to the time series dose 80 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	NFKB1A	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			

Table 2

miRNA/mRNA pairs that are significant with the fibrosis pathology and verified using the second derivative test.

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