Using NMFAS to Identify Key Biological Pathways Associated With Human Diseases

Hao Guo, Yunping Zhu, Dong Li, Fuchu He §
State Key Laboratory of Proteomics, Beijing Proteome Research Centre
Beijing Institute of radiation Medicine
Beijing, P. R. China
haosmail@gmail.com
§ Corresponding authors: hefc@nic.bmi.ac.cn

Qijun Liu
National Laboratory for Parallel & Distributed Processing
National University of Deference and Technology
Changsha, P. R. China
ivanliuqj@gmail.com

***Abstract*—Gene expression microarray enables us to measure the gene expression levels for thousands of genes at the same time. Here, we constructed the non-negative matrix factorization analysis strategy (NMFAS) to dig the underlying biological pathways related with various diseases by factorizing the pathway expression matrix, which was extracted from microarray matrix using pathway membership information, into the product of row and column vectors. We defined row vector as the pathway activity and column vector as the gene contribution weight. Via comparing the pathway activity of two different sample groups, we can identify significantly expressed pathways. We applied this strategy on two different cases: smoking and type 2 diabetes (DM2). We found 152 differentially expressed pathways by the comparison of pathway activity between smoker and never smoker, including pathways that have been validated in literature, such as "O-Glycans biosynthesis" and "Glutathione metabolism". We also found important genes related to smoking phenotype, such as NQO, HSPA1A, ALDH3A1. As for DM2 analysis, our results suggested 9 pathways were significantly expressed, including typical pathways like "Oxidative phosphorylation" and "mTOR signaling pathway", and found genes like CAPNS1, APP, COX7A1, COX7B, which might play important roles in the cellular regulations of DM2. In conclusion, Our strategy can be efficiently used to integrate gene expression profiles and biological pathway information to identify the key processes underlying human disease and can identify gene pathways missed by alternative approaches.**

Keywords-non-negative matrix factorization; pathways; microarray; smoking; type 2 diabetes

I. INTRODUCTION

The development of high-throughput technologies including microarray experiments have triggered an explosion of large amounts of genome-wide expression profiles. Currently, a common challenge is to search for the expression patterns and uncover the underlying biological meaning of the massive data. Particularly, the identification of differentially expressed genes and pathways associated with the phenotypes of special diseases or therapy responses has attracted extensive attention [1, 2].

Various methods have been developed to identify genes and pathways for this purpose. Most of these methods could be divided into two groups: single gene analysis and gene set analysis. The former analysis strategies include clustering algorithm [3, 4] and some statistical analysis for differentially expressed genes [6, 7]. The common goal of these single gene analysis methods is to find some genes with coordinated expression patterns or a small group of genes to predict the response outcome. They did not consider the inherent relationships of genes. The second group combined the microarray dataset and pathway databases or GO annotations. By using dimension reducing method, e.g. singular value decomposition (SVD) [5], or statistic test, e.g. gene set enrichment analysis (GSEA) [10], researchers obtained a rank of pathways that were probably related to disease phenotypes, and a sorted list of genes in each pathway that have the biggest impact on the whole pathway. Using predefined canonical pathways, these methods identified easily-interpretable biological meanings of the genes and pathways. Based on the combination of expression profiles and information on pathways, we noted that SVD is a kind of dimension reducing methods firstly used in the face recognition technology, which could extract the predominant component of the full structure. One of the drawbacks of using such methods is that the factorized matrix will have negative components, which is not suited to the interpretation of the textual representation or biological meanings behind microarray data.

To address this issue, a novel dimension reducing method named non-negative matrix factorization (NMF) was presented by Lee and Seung [11] in 1999. It could generate non-negative parts-based representation as the low rank approximation of original data matrix and keep data locality in dimensional reduction. It has been reported that NMF performed better than SVD and PCA [11, 12] on face recognition and latent semantic analysis. As the original data matrix of microarray is non-negative, NMF has been proved a natural method to cluster genes and samples for its nonnegative constraint during the factorizing process, which can provide a more intuitive partial view of the whole data matrix. Brunet et al. [13] has adopted NMF to elucidate tumor subtypes and data substructure of the cancer microarray data. Similar work to identify molecular patterns of microarray data has been done by Gao and Church [14]. NMF could also be regarded as a tool to cluster genes and predict functional cellular relationships in gene expression data [15]. More improvements and applications of NMF in computational

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

biology have been introduced by Devarajan [16].

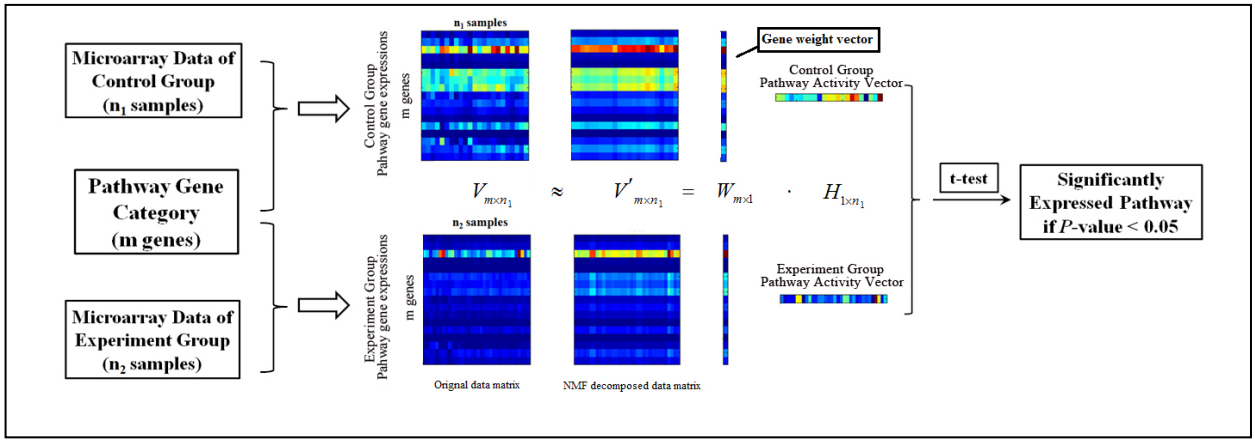

Figure 1. Work flow of NMFAS

In this work, we firstly adopted NMF algorithm to process the combined datasets of expression profiles and pathways information, trying to explore the ability of NMF algorithm to discover disease related pathways.

II. RESULTS

A. *Using NMFAS to identify the pathway underlying phenotype*

For microarray data, we averaged the expression levels of probes for the same gene and then divided the microarray data into two groups (Figure 1): control group and experiment (case) group. For pathways we downloaded, member genes of pathway that are not represented on microarrays are not included for further analysis. As a result, we obtained 453 biological pathways. For each pathway, we generated two data matrix, control group pathway gene expression matrix and experiment group pathway gene expression matrix for NMF factorization by selecting the gene expressions of its member genes from the microarray data of the two groups according to the gene Entrez IDs. After that, we utilized 1-dimension NMF (k=1) to factorize the two group pathway expression matrixes simultaneously (See method section for details) and obtained two row vector H, which were defined as control group pathway activity vector and experiment group pathway activity vector. Then, we adopted two sample t-test to test the significance of difference between two activity vectors and obtained a P-value. The pathway is considered differentially expressed if the corresponding p-value is less than 0.05. We prioritized the candidate 453 pathways according to their P-value and obtained a potentially phenotype related pathways list. Meanwhile, column vector W is actually a vector of weights (the sum of its elements is 1), which could be treated as the gene contribution scores. By ranking the weight value of each gene, we can sort out genes that have great impact on the whole pathway.

B. *Identification of smoker related pathways and genes by NMF analysis*

Cigarette smoking is the main cause of pulmonary diseases and lung cancer. The effect of smoking on gene expression has been a hot subject recently. We use the gene expression profiles obtained by research group at Boston University [17] which contained three groups of volunteers from 34 current smokers, 18 ex-smokers and 23 non-smokers. And here we identified the differentially expressed pathway between current and non-smoking groups.

We first calculated the activity levels for 453 pathways by using NMF analysis, and then used the two sample t test (see methods section for details) to identify the differentially expressed pathways (P-value<0.05). We obtained a list of 152 differentially expressed pathways between the current smoking group and non-smoking group, and listed the top 15 significantly expressed pathways (all up-regulated) in TABLE I.

Among the 152 differentially expressed pathways listed, we found that several pathways were reported to show altered activity in response to cigarette smoking in vivo or in vitro. For example, O-Glycan biosynthesis ($P = 1.281\times10^{-8}$) is linked to the increased sputum production observed in smokers [18]. Gamma-Hexachlorocyclohexane degradation pathway ($P = 8.463\times10^{-8}$) is known to contain several cytochrome P450 genes with polymorphisms that are known to alter lung cancer risk for smokers [19]. Pentose phosphate pathway ($P = 4.681\times10^{-6}$) of glucose metabolism was previously found to be activated in the endothelial cells of plasma during the treatment of exposure to cigarette smoke in vitro [20]. As known, cigarette smoking is the most common oxidant stress in daily life and can affect the antioxidant capacity in human lung cells [21]. The antioxidant function related pathways or genes may play a key role in smoking-induced human diseases. And we found that glutathione metabolism ($P = 6.950\times10^{-4}$), which is known to be a notable antioxidant pathway, may be impaired by chronic cigarette smoking and was also found altered in the endothelia cells while exposed to the cigarette smoke [20, 22].

Among these pathways, the most significantly expressed pathway is "Biosynthesis of steroids". It was seldom mentioned by foregoing methods. Steroid biosynthesis is a basic anabolic metabolic pathway, which is a common target for antibiotics and cholesterol-lowering drugs. Research

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

groups found that cigarette smoking could alter the pattern of steroid levels and inhibit steroidogenesis by reducing the interactions of the voltage-dependent anion-selective channel (VDAC) and the Steroidogenic acute regulatory protein (StAR) [23].

TABLE I. TOP 15 SIGNIFICANTLY EXPRESSED PATHWAYS IDENTIFIED BY NMFAS.

PATHWAY RANK	CATEGORY SIZE [1]	*P*-VALUE
BIOSYNTHESIS OF STEROIDS	16/17	6.006×10^{-12}
HYPOXIA AND P53 IN THE CARDIOVASCULAR SYSTEM	21/23	2.407×10^{-10}
TRYPTOPHAN METABOLISM	58/84	2.880×10^{-9}
LIMONENE AND PINENE DEGRADATION	19/28	3.518×10^{-9}
FATTY ACID METABOLISM	41/51	4.627×10^{-9}
BETA-ALANINE METABOLISM	23/24	5.634×10^{-9}
UREA CYCLE AND METABOLISM OF AMINO GROUPS	24/30	5.648×10^{-9}
VALINE, LEUCINE AND ISOLEUCINE DEGRADATION	40/50	6.464×10^{-9}
TYROSINE METABOLISM	41/59	7.381×10^{-9}
HISTIDINE METABOLISM	28/41	1.070×10^{-8}
BILE ACID BIOSYNTHESIS	33/39	1.121×10^{-8}
O-GLYCAN BIOSYNTHESIS	13/26	1.281×10^{-8}
PHENYLALANINE METABOLISM	23/29	1.526×10^{-8}
ASCORBATE AND ALDARATE METABOLISM	11/14	2.054×10^{-8}
GLYCEROLIPID METABOLISM	49/60	2.619×10^{-8}

[1] RIGHT NUMBER OF CATEGORY SIZE COLUMN IS THE TOTAL NUMBER OF GENES IN ONE PATHWAY. LEFT NUMBER REPRESENTS THE GENE NUMBER WE USED IN THE ANALYSIS (ONLY THE GENES WITH EXPRESSION INFORMATION WERE CONSIDERED IN OUR ANALYSES).

For a specific pathway, NMF analysis can also rank the importance of genes according to the NMF weights of each gene. As TABLE II listed below, NQO1, FDFT1 and SQLE are ranked in top three in the biosynthesis of steroids pathway. Among them, NQO1 is an important flavoenzyme involved in xenobiotic metabolism, which protectscells from oxidative damage and has been reported to play important role in the tumorigenesis of bladder cancer induced by smoking [24]. FDFT1 and SQLE also play important roles in the oxygenation process in sterol biosynthesis [25]. Functions of these three genes also support that biosynthesis of steroids pathway is a smoking-related pathway.

The second significant pathway is "Hypoxia and p53 in the Cardiovascular system", and this pathway is related to hypoxic stress and induction of p53 protein accumulation and p53-dependent apoptosis. The gene with the highest contribution weight in this pathway is HSPA1A, which is a member of the heat shock protein 70 family and plays a crucial role in endothelial cell apoptosis [26]. NQO1, which is also included in biosynthesis of steroids pathway, is notably related to antioxidant function of cells [27].

The third significant pathway, tryptophan metabolism, is linked to smoking initiation and progression (SI/P) and nicotine dependence [18, 28] . Aldehyde dehydrogenase isozymes 3 (ALDH3A1) is shared by several pathways with top rank, and may be upregulated in lung tissue as a result of exposure to carcinogenic aldehydes found in cigarette smoke [29]. CYP1A1 encodes a member of the cytochrome P450 superfamily of enzymes, and is involved in the metabolic process of tobacco carcinogens and could be implicated in smoking-induced lung cancer [30].

Our application of NMF to gene expression datasets of current smokers and non-smokers strongly suggested that our method is capable of identifying and characterizing the genome expression difference between two different sample groups. NMF has identified a series of pathways and genes that were also suggested by GSEA and SVD as the phenotype related molecular sets, such as glutathione metabolism, O-Glycan biosynthesis. GSEA did not found "Biosynthesis of steroids" as a notable smoking related pathway, while SVD found it a related pathway with ranking 13th. We also noted that NMF found the most effective gene in this pathway, NQO1, a well reported antioxidant gene, while SVD did not mention this gene [5].

TABLE II. TOP 5 RANKING CONTRIBUTION WEIGHTS FOR GENES IN THE THREE MOST SIGNIFICANT PATHWAYS SORTED OUT BY NMF ALGORITHM FROM SMOKING DATASET.

Pathways	Genes	NMF Weights	*P*-value[1]
Biosynthesis of steroids	NQO1	4.012×10^{-1}	4.809×10^{-11}
	FDFT1	1.264×10^{-1}	8.797×10^{-3}
	SQLE	6.330×10^{-2}	9.602×10^{-1}
	IDI1	5.419×10^{-2}	3.950×10^{-1}
	PMVK	4.900×10^{-2}	5.745×10^{-3}
Hypoxia and p53 in the Cardiovascular system	HSPA1A	3.026×10^{-1}	3.415×10^{-4}
	NQO1	2.864×10^{-1}	4.809×10^{-11}
	IGFBP3	7.036×10^{-2}	2.378×10^{-1}
	CSNK1A1	6.728×10^{-2}	3.989×10^{-3}
	CDKN1A	5.907×10^{-2}	8.979×10^{-1}
Tryptophan metabolism	ALDH3A1	4.791×10^{-1}	1.863×10^{-9}
	ECHS1	5.600×10^{-2}	2.303×10^{-1}
	WARS	3.446×10^{-2}	4.941×10^{-1}
	DHCR24	3.418×10^{-2}	5.631×10^{-1}
	CYP1A1	2.605×10^{-2}	1.846×10^{-3}

1 INDIVIDUAL P-VALUES CORRESPONDING TO T-TEST RESULTS BETWEEN THE EXPRESSION VALUES OF THE GENE IN THE SMOKER VS. NON-SMOKER SAMPLES.

C. Application of NMF analysis on the type II diabetes research

Type 2 diabetes mellitus is a common chronic disease which will induce atherosclerotic vascular disease, blindness and kidney failure. Mootha et al. [31] took use of microarrays to profile expression of 43 age-matched males, 17 with normal glucose tolerance (NGT), 8 with impaired glucose tolerance (IGT) and 18 with type 2 diabetics (DM2), in skeletal muscle tissue.

We also adopted our methods to analyze this dataset. By comparing the pathway activity of skeletal muscle tissue samples from the patients with DM2 and with NGT, GSEA found a pathway with significantly decreased expression, oxidative phosphorylation [31], while SVD reported no significant results. Using our method, we found 9 pathways with P-value less than 0.05 after multiple comparison and these pathways did make biological sense (see TABLE III, ↑ represent up-regulated).

The first ranking pathway is "Deregulation of CDK5 in Alzheimers Disease", which is seldom reported to be related to DM2 by current machine learning methods. However, ranking results of our methods suggested that this pathway might be differentially expressed in type 2 diabetic skeletal muscle tissue samples. It was reported recently that Alzheimer's disease (AD) and DM2 share several molecular

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

processes [32]. Multiple studies report that patients with diabetes have a 50-75% increased risk of developing AD compared with age- and gender-matched patients without diabetes [33]. Disturbances in insulin secretion appear to be the main common impairment that increases the risk of AD and DM2. Actually, CDK5 promotes insulin secretion and is deregulated in AD brains [34]. We also found the oxidative phosphorylation pathway differentially expressed as the 2nd rank in our results, which is consistent with the discovery by GSEA. The pathway that ranked 3rd, " Skeletal muscle hypertrophy is regulated via AKT/mTOR pathway" and the pathway that ranked the 7th,"mTOR signaling pathway" were recently recognized as playing critical roles in insulin resistance and glucose metabolism [35].

TABLE III. TOP 10 SIGNIFICANTLY EXPRESSED PATHWAYS IDENTIFIED BY NMFAS.

Pathway Rank	Size	Up/ Down	P-value
Deregulation of CDK5 in Alzheimers Disease	10/11	↑	9.878×10^{-3}
Oxidative phosphorylation	93/129	↓	1.602×10^{-2}
Skeletal muscle hypertrophy is regulated via AKT/mTOR pathway	20/20	↓	1.608×10^{-2}
Limonene and pinene degradation	19/28	↓	3.055×10^{-2}
Blockade of Neurotransmitter Relase by Botulinum Toxin	4/5	↑	3.739×10^{-2}
Activation of cAMP-dependent protein kinase, PKA	6/6	↑	3.743×10^{-2}
Selenoamino acid metabolism	21/34	↓	3.802×10^{-2}
mTOR signaling pathway	44/50	↓	3.997×10^{-2}
Presenilin action in Notch and Wnt signaling	13/14	↑	4.391×10^{-2}
Activation of Csk by cAMP-dependent Protein Kinase Inhibits Signaling through the T Cell Receptor	19/24	--	5.304×10^{-2}

TABLE IV. TOP 5 CONTRIBUTION WEIGHTS FOR GENES IN THE TOP 2 PATHWAYS SORTED OUT BY NMF ALGORITHM FROM DM2 DATASET.

Pathways	Genes	NMF Weights	P-value[1]
Deregulation of CDK5 in Alzheimers Disease	CAPNS1	5.791×10^{-1}	1.608×10^{-2}
	CSNK1A1	1.076×10^{-1}	8.411×10^{-1}
	CSNK1D	6.922×10^{-2}	6.029×10^{-1}
	APP	6.878×10^{-2}	7.943×10^{-1}
	MAPT	5.746×10^{-2}	4.736×10^{-2}
Oxidative phosphorylation	COX7A1	6.340×10^{-2}	1.578×10^{-2}
	ATP5A1	4.333×10^{-2}	9.545×10^{-2}
	NDUFA4	4.302×10^{-2}	2.376×10^{-1}
	COX7B	4.202×10^{-2}	3.964×10^{-2}
	COX6A2	3.712×10^{-2}	5.969×10^{-1}

1 INDIVIDUAL P-VALUES CORRESPONDING TO T-TEST RESULTS BETWEEN THE EXPRESSION VALUES OF THE GENE IN THE DM2 VS. NGT SAMPLES.

By viewing the global NMF weights of all the genes in the top 2 pathways (see TABLE IV), we found that most of genes in these two pathways were not significantly differentially expressed according to the comparisons. In the first pathway, CAPNS1 is the most important gene that affects the pathway activity according to its contribution weight and the only gene which is differentially expressed. It was reported CAPNS1 had been implicated in neurodegenerative processes after oxidative stress stimulation and modulated the cell survival and migration [36]. The increased activity of calcium activated neutral proteinase has been previously reported to be linked with diabetic amyotrophy in mouse skeletal muscle [37]. Regulated proteolysis of APP (amyloid beta precursor protein) has been report as a possible link between DM2 and AD [38].

Genes in the pathway "oxidative phosphorylation", such as COX7A1, COX7B and NDUFA4, have great impact on the whole pathway activity. COX7A1 was previously reported to be down-regulated in skeletal muscle from patients with DM2 [39]. COX7B, which is a subunit of the terminal component of the respiratory chain complex, is involved in the regulation of insulin secretion [37]. An interesting discovery is that CAPNS1 and COX7A1 are located immediately adjacent to each other on chromosome [40].

III. DISCUSSION

Applications of NMFAS found not only some potentially phenotype associated pathways between two different sample groups, but also some important genes which impact pathway activity greatly. By using the same gene expression datasets, we compared the performance of our algorithm with that of others on the DM2 dataset, and we found the following results. Firstly, NMF can identify more pathways than other methods. NMF has identified 9 pathways that were potentially related with DM2 while SVD and GSEA only 0 and 2 pathways respectively. By literature mining, NMF has identified 3 literature-reported DM2 related pathways, while SVD and GSEA only 0 and 1 pathways respectively. Secondly, NMF algorithm can rank the importance of each gene in the pathways, for example, gene CAPNS1and gene COX7A1, which are located immediately adjacent to each other on chromosome. Finally, the online documentation of GSEA suggests that this method may produce inflated scores when the size of gene set analyzed is smaller than 25. And the NMF method doesn't have such limitation.

We also evaluated the performance of NMFAS using an ovarian cancer dataset [43]. 16 pathways was confirmed respectively to be related with ovarian cancer survival by two groups [43,44], which were regarded as the gold standard dataset. 261 pathways were identified by NMFAS significantly related with survival ($P<0.001$), of which 10 were confirmed by the gold standard dataset, while none of pathways is identified by GSEA [45].

It is clear that the NMF analysis could suggest candidate pathways and genes related with disease phenotype and provide important clues for disease mechanism and drug response. We would not claim that NMF analysis is much 'better' than previous useful methods, but it does clearly have independent value and could provide valuable complementary results that are not suggested by the other methods. Furthermore, we noted that Zhang et al. [46] recently adopted sparse NMF to integrate gene expression and interaction datasets and effectively predicted miRNA-gene and gene-gene interactions. Although their work has a different goal and application from ours, we are inspired to use the multi-dimension NMF factorization information in the next step to analyze the disease phenotype related pathways.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

IV. MATERIAL AND METHODS

A. Gene expression datasets

The gene expression data for smokers were downloaded from the Airway Gene Expression Database (http://pulm.bumc.bu.edu/aged). All expression data has been pre-processed, including normalization and noise processing. More details can be found on AGED website. The DM2 dataset was downloaded from the Whitehead Institute Center for Genome Research website (http://www.broad.mit.edu/cancer) along with phenotype data and other information.

B. Pathway gene sets

453 pathways were downloaded from BioCarta (http://www.biocarta.com/) and KEGG (Kyoto Encyclopedia of Genes and Genomes) [41], mainly including biological processes related to metabolism pathways, biosynthesis pathways and signaling pathways, which were assembled by PLAGE website [5] .

C. Nonnegative matrix factorization algorithm

The gene expression profiles of each pathway were abstracted from the original microarray data matrix. Namely, we obtained 453 expression sub-matrices for all pathways we downloaded from the pathway database. Usually, a pathway expression profile consists of the expression levels of M genes in N samples. The pathway expression profile is then represented by a data $M \times N$ matrix V.

Considering the non-negativity of microarray data matrix, we used NMF to factorize the matrix V into the product of two positive matrixes.

$$V_{m\times n} \approx W_{m\times k} \cdot H_{k\times n} \tag{1}$$

Where, $W_{m\times k}$ is k dimensional column vector, and row vector $H_{k\times n}$ also has k dimension. We define the $H_{k\times n}$ as metasamples, $W_{m\times k}$ as the expression weight for the metasamples [13]. Inspired by Tomfohr et al. [5], we set the k equal to 1, and define the element of metasample $H_{1\times n}$ as the pathway activity, namely each pathway has a different activity in each sample. Each element of $W_{m\times 1}$ can be regarded as the contribution weight of each gene to the pathway activity. Once we determined the contribution weight of every gene in a pathway, we could find out which gene impacts the pathway activity mostly.

The NMF algorithm needs an initial vector w and h to start the iteration process. We set random initial w and h. The divergence function has been proved a more powerful rule for updating than residual minimizing equation, and K-L divergence function has been proved more effective when used it to detect the complex pattern of gene expressions in biological systems [13]. We iteratively updated W and H in each step to minimize the K-L divergence.

$$D(V\|WH) = \sum_{i,j}[V_{ij}\log\frac{V_{ij}}{(WH)_{i,j}} - V_{i,j} + (WH)_{i,j}] \tag{2}$$

And that equals to update the coupled divergence equations as described in (3).

$$\begin{cases} \left.\begin{aligned} W_{ia} &\leftarrow W_{ia}\sum_{\mu}\frac{V_{i\mu}}{(WH)_{i\mu}}H_{a\mu} \\ H_{a\mu} &\leftarrow H_{a\mu}\sum_{i}W_{ia}\frac{V_{i\mu}}{(WH)_{i\mu}} \end{aligned}\right\} \text{Divergence update rule} \\ \left.\begin{aligned} W_{ia} &\leftarrow \frac{W_{ia}}{\sum_{j}W_{ja}} \end{aligned}\right\} \text{Liu's Normalization} \end{cases} \tag{3}$$

Following the multiplicative updating rule listed above, NMF would converge to different local optimal matrix factorizations after repeated iterations. That is to say, NMF is not deterministic and different runs yield different results. After looking up to multiple NMF algorithms, we found that NMF with repeating the K-L divergence update rules and Liu's normalization [42] during each iteration step could eventually yield unique factorization result. Elements of W are values in the interval (0, 1) after the iteration in (3), which could be regarded as the impact factor of each gene to the pathway activity. We define the vector W as the gene weight vector and H as the pathway activity vector. We then compare the pathway activity vector H of experiment group with the control group. Algorithm introduced above is programmed in MATLAB m files.

D. Prioritization of the underlying pathways

When two pathway activity vectors were obtained from two different groups, we adopted the P-value from two sample t-test to prioritize the candidate pathways. If the P-value of the comparison between the two groups exceeds the alpha level, the pathway activities of the two groups would be regarded as significantly different. The calculation of the t statistic is

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n} + \frac{s_y^2}{m}}} \tag{5}$$

Where $\bar{x}$ and $\bar{y}$ are sample means, s_x and s_y are the sample standard deviations of x and y, n and m are the sample sizes of x and y, respectively. If the activity means of case group are greater than control group, pathway would be regarded as up-regulated.

ACKNOWLEDGMENT

H.G. thanks Lin Hou for valuable comments and language corrections. H.G. thanks Chunyuan Yang for helpful discussion.

This work was funded by the Chinese National Key Program of Basic Research (2011CB910202) and National S&T Major Project(2008ZX10002-016, 2009ZX09301-002).

REFERENCES

[1] D. di Bernardo, et al., "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks," Nat Biotechnol, vol. 23, pp. 377-83, Mar 2005.

[2] H. Zhu and L. Li, "Biological pathway selection through nonlinear dimension reduction," Biostatistics, Jan 20 2011.

[3] M. B. Eisen, et al., "Cluster analysis and display of genome-wide expression patterns," Proc Natl Acad Sci U S A, vol. 95, pp. 14863-8, Dec 8 1998.

[4] P. Tamayo, et al., "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," Proc Natl Acad Sci U S A, vol. 96, pp. 2907-12, Mar 16 1999.

[5] J. Tomfohr, et al., "Pathway level analysis of gene expression using singular value decomposition," BMC Bioinformatics, vol. 6, p. 225, 2005.

[6] R. K. Curtis, et al., "Pathways to the analysis of microarray data," Trends Biotechnol, vol. 23, pp. 429-35, Aug 2005.

[7] S. Draghici, et al., "Global functional profiling of gene expression," Genomics, vol. 81, pp. 98-104, Feb 2003.

[8] J. J. Goeman, et al., "A global test for groups of genes: testing association with a clinical outcome," Bioinformatics, vol. 20, pp. 93-9, Jan 1 2004.

[9] B. R. Zeeberg, et al., "High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID)," BMC Bioinformatics, vol. 6, p. 168, 2005.

[10] A. Subramanian, et al., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," Proc Natl Acad Sci U S A, vol. 102, pp. 15545-50, Oct 25 2005.

[11] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," Nature, vol. 401, pp. 788-91, Oct 21 1999.

[12] R. Peter, et al., "Evaluation of SVD and NMF Methods for Latent Semantic Analysis," International Journal of Recent Trends in Engineering, vol. 1, pp. 308-310, 2009.

[13] J. P. Brunet, et al., "Metagenes and molecular pattern discovery using matrix factorization," Proc Natl Acad Sci U S A, vol. 101, pp. 4164-9, Mar 23 2004.

[14] Y. Gao and G. Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization," Bioinformatics, vol. 21, pp. 3970-5, Nov 1 2005.

[15] P. M. Kim and B. Tidor, "Subsystem identification through dimensionality reduction of large-scale gene expression data," Genome Res, vol. 13, pp. 1706-18, Jul 2003.

[16] K. Devarajan, "Nonnegative matrix factorization: an analytical and interpretive tool in computational biology," PLoS Comput Biol, vol. 4, p. e1000029, 2008.

[17] A. Spira, et al., "Effects of cigarette smoke on the human airway epithelial cell transcriptome," Proc Natl Acad Sci U S A, vol. 101, pp. 10143-8, Jul 6 2004.

[18] T. M. Dwyer, "Cigarette smoke-induced airway inflammation as sampled by the expired breath condensate," Am J Med Sci, vol. 326, pp. 174-8, Oct 2003.

[19] G. S. Eichler, et al., "The LeFE algorithm: embracing the complexity of gene expression in the interpretation of microarray data," Genome Biol, vol. 8, p. R187, 2007.

[20] A. A. Noronha-Dutra, et al., "Effect of cigarette smoking on cultured human endothelial cells," Cardiovasc Res, vol. 27, pp. 774-8, May 1993.

[21] J. D. Morrow, et al., "Increase in circulating products of lipid peroxidation (F2-isoprostanes) in smokers. Smoking as a cause of oxidative damage," N Engl J Med, vol. 332, pp. 1198-203, May 4 1995.

[22] S. Teramoto, et al., "Effect of age on alteration of glutathione metabolism following chronic cigarette smoke inhalation in mice," Lung, vol. 174, pp. 119-26, 1996.

[23] M. Bose, et al., "Cigarette smoke decreases mitochondrial porin expression and steroidogenesis," Toxicol Appl Pharmacol, vol. 227, pp. 284-90, Mar 1 2008.

[24] D. W. Nebert, et al., "NAD(P)H:quinone oxidoreductase (NQO1) polymorphism, exposure to benzene, and predisposition to disease: a HuGE review," Genet Med, vol. 4, pp. 62-70, Mar-Apr 2002.

[25] K. S. Dolt, et al., "Transcriptional downregulation of sterol metabolism genes in murine liver exposed to acute hypobaric hypoxia," Biochem Biophys Res Commun, vol. 354, pp. 148-53, Mar 2 2007.

[26] M. He, et al., "Functional SNPs in HSPA1A gene predict risk of coronary heart disease," PLoS One, vol. 4, p. e4851, 2009.

[27] A. T. Dinkova-Kostova and P. Talalay, "NAD(P)H:quinone acceptor oxidoreductase 1 (NQO1), a multifunctional antioxidant enzyme and exceptionally versatile cytoprotector," Arch Biochem Biophys, Mar 31 2010.

[28] J. Wang and M. D. Li, "Common and unique biological pathways associated with smoking initiation/progression, nicotine dependence, and smoking cessation," Neuropsychopharmacology, vol. 35, pp. 702-19, Feb 2010.

[29] M. Patel, et al., "ALDH1A1 and ALDH3A1 expression in lung cancers: correlation with histologic type and potential precursors," Lung Cancer, vol. 59, pp. 340-9, Mar 2008.

[30] D. P. Ng, et al., "CYP1A1 polymorphisms and risk of lung cancer in non-smoking Chinese women: influence of environmental tobacco smoke exposure and GSTM1/T1 genetic variation," Cancer Causes Control, vol. 16, pp. 399-405, May 2005.

[31] V. K. Mootha, et al., "PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," Nat Genet, vol. 34, pp. 267-73, Jul 2003.

[32] L. Li and C. Holscher, "Common pathological processes in Alzheimer disease and type 2 diabetes: a review," Brain Res Rev, vol. 56, pp. 384-402, Dec 2007.

[33] B. Kim, et al., "Increased tau phosphorylation and cleavage in mouse models of type 1 and type 2 diabetes," Endocrinology, vol. 150, pp. 5294-301, Dec 2009.

[34] L. Lilja, et al., "Cyclin-dependent kinase 5 promotes insulin exocytosis," J Biol Chem, vol. 276, pp. 34199-205, Sep 7 2001.

[35] A. Tzatsos and K. V. Kandror, "Nutrients suppress phosphatidylinositol 3-kinase/Akt signaling via raptor-dependent mTOR-mediated insulin receptor substrate 1 phosphorylation," Mol Cell Biol, vol. 26, pp. 63-76, Jan 2006.

[36] P. Pamonsinlapatham, et al., "Capns1, a new binding partner of RasGAP-SH3 domain in K-Ras(V12) oncogenic cells: modulation of cell survival and migration," Cell Signal, vol. 20, pp. 2119-26, Nov 2008.

[37] S. Kobayashi, et al., "Diabetic state-induced activation of calcium-activated neutral proteinase in mouse skeletal muscle," Endocrinol Jpn, vol. 36, pp. 833-44, Dec 1989.

[38] E. Kojro and R. Postina, "Regulated proteolysis of RAGE and AbetaPP as possible link between type 2 diabetes mellitus and Alzheimer's disease," J Alzheimers Dis, vol. 16, pp. 865-78, 2009.

[39] T. Ronn, et al., "Age influences DNA methylation and gene expression of COX7A1 in human skeletal muscle," Diabetologia, vol. 51, pp. 1159-68, Jul 2008.

[40] C. Drogemuller and T. Leeb, "Molecular characterization of the porcine gene CAPNS1 encoding the small subunit 1 of calpain on SSC6q1.1-->q1.2," Cytogenet Genome Res, vol. 98, pp. 206-9, 2002.

[41] H. Ogata, et al., "KEGG: Kyoto Encyclopedia of Genes and Genomes," Nucleic Acids Res, vol. 27, pp. 29-34, Jan 1 1999.

[42] W. Liu, et al., "Non-negative matrix factorization for visual coding," in Proceedings of the IEEE International Confer-ence on Acoustics, Speech, and Signal Processing, 2003, pp. 293-296.

[43] A. P. Crijns, et al., "Survival-related profile, pathways, and transcription factors in ovarian cancer," PLoS Med, vol. 6, p. e24, Feb 3 2009.

[44] H. K. Dressman, et al., "An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer," J Clin Oncol, vol. 25, pp. 517-25, Feb 10 2007.

[45] S. Lee and J. Kim, "A comparative study on gene-set analysis methods for assessing differential expression associated with the survival phenotype," BMC Bioinformatics, vol. 12, p. 377, 2011.

[46] S. Zhang, et al., "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules," Bioinformatics, vol. 27, pp. i401-9, Jul 1 2011.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE