Multi-dimensional Cluster Misclassification Test for Pathway Differential Analysis of Diabetes

Lily R. Liang*
Department of Computer Science and Information Technology
University of the District of Columbia, USA
lliang@udc.edu

Vinay Mandal
Department of Computer Science
Wayne State University, USA

Yi Lu
Department of Computer Science
Prairie View A&M University, USA

Deepak Kumar*
Department of Biology and Environmental Sciences
University of the District of Columbia
dkumar@udc.edu

Abstract

Gene pathway can be defined as a group of genes that interact with each other to perform some biological processes. Along with the efforts to identify the individual genes that play vital roles in a particular disease, there is a growing interest in identifying the roles of gene pathways in such diseases. This paper proposes an innovative method which measures the significance of the gene pathways in a particular disease using the concept of fuzzy set theory. Experiments have been conducted on published diabetes gene expression dataset together with a list of predefined pathways from KEGG. Results on the real world diabetes dataset identified OXPHOS pathway involved in oxidative phosphorylation in mitochondria and other mitochondrial related pathways to be deregulated in diabetes patients. Our results support the previously supported notion that mitochondrial dysfunction is an important event in insulin resistance and type-2 diabetes.

Keywords: *fuzzy, pathway analysis, multi-dimensional cluster misclassification*

1. Introduction

Current microarray technologies conduct simultaneous studies of gene expression data of thousands of genes under different conditions. In most of these studies, expression data are analyzed using various standard statistical methods to identify a list of genes responsible for a particular condition. However, in these approaches, interplay among genes is not taken into account. Since organisms behave as complex systems, functioning through chemical networks and interaction of various molecules (also known as pathways), this interplay of genes may provide invaluable insight to the understanding of various diseases. Thus, along with the efforts to identify the individual genes that play vital roles in a particular disease, there is a growing interest in identifying the roles of different pathways in such diseases.

Biological pathway is a group of related genes coding for proteins that interact with each other to perform some biological processes. According to the biological processes they are involved with, pathways can be divided into several categories, such as metabolic pathways and regulatory pathways. Metabolic pathways are series of chemical reactions occurring within a cell, catalyzed by enzymes, resulting in either the formation of a metabolic product to be used or stored by the cell, or the initiation of another metabolic pathway. Regulatory pathways represent protein-protein interactions.

During the past few years, many signaling and metabolic pathways have been discovered experimentally and have been integrated into pathway databases, such as KEGG [2] and Biocarta [1]. Various statistical techniques have been developed to analyze microarray expression data for the relevance of predefined pathways to a particular disease. These techniques include gene set enrichment analysis [8][10], pathway level analysis of gene expression using singular value decomposition by Tomfohr et al. [12], and hypothesis testing [11] by Tian et al. These approaches are reviewed in detail in the related works section.

* Corresponding authors

Table 1. An example of a five-gene pathway

Gene ID	S_1					S_2					CM *d*-value	*P*-value		
												CM-test	t-test	Rank Sum test
1	750	559	649	685	636	310	359	135	97	178	1	0.001	0.000	0.008
2	391	379	268	323	380	774	506	416	468	449	1	0.005	0.029	0.008
3	598	424	695	451	141	342	260	266	229	234	0.904	0.018	0.077	0.152
4	233	216	193	394	327	436	980	363	424	416	0.905	0.017	0.071	0.015
5	305	221	241	183	158	201	176	189	177	250	0.812	0.143	0.448	0.693

Generally speaking, these approaches can be divided into two categories:

- Conduct statistical differential analysis at the individual gene level, and integrate the result statistics of the genes in the same pathway;
- Obtain activity level indices of each pathway for different sample groups and conduct differential analysis of these indices.

For the first category, when the statistics at individual gene level miss significant genes, the effectiveness of the pathway analysis will be affected. An example is given in the related works section. For the second approach, extracting pathway activity level indices from gene expression data may cause loss of information.

In this paper, we propose an innovative approach for pathway analysis, Multi-dimensional Cluster Misclassification test (MCM-test), based on our previous works [5] [6] [7]. The differential analysis is done directly at the pathway level without calculating individual gene differential statistics or extracting activity level of pathways. This allows keeping the maximum amount information for the pathway differential analysis. The fuzzy concept also made the approach more tolerant to individual data item noise.

Diabetes is a group of diseases characterized by high levels of blood glucose resulting from defects in insulin production, insulin action, or both. It is one of the most common diseases, affecting 18.2 million people in the United States, or 6.3% of the population [4]. Hence, identifying active pathways in diabetes is a critical task for understanding this disease. Several pathway analysis works have been proposed in this area [8] [12] [11]. In this paper, we apply our proposed MCM-test on diabetes data to identify pathways involved in diabetes.

Contributions. In this paper, we propose an innovative fuzzy-set-theory based approach to measure the significance of gene pathways and apply it on identifying significant pathways for diabetes. Experiments have been conducted on both synthetic data sets and real world data. Results on real world diabetes data identified several pathways that were deregulated in diabetes patients. The top 3 pathways identified were related to mitochondrial functions in accordance with previous diabetes studies. Mitochondrial dysfunction is known to be related to insulin resistance and type-2 diabetes. Our data suggests that the method can be successfully used in pathway level differential analysis of gene expression datasets.

The paper is organized as following: In Section 2, related works in the literature are reviewed; we propose our methodology in Section 3 and analyze it for its theoretical justification in Section 4; experiment results are presented in Section 5, and we give our conclusions and discuss future works in Section 6.

2. Related works

In GSEA [8], a differential statistic is calculated first for each gene from their expression data of two different groups of samples. Then the genes are ordered according to the statistic values. A running sum of weights is calculated from the ordered list for a particular pathway. The maximum value of this running sum is called the enrichment score of that pathway. To measure the significance of this score, a null distribution of enrichment scores is generated by

permuting the sample labels. This approach falls into the first category as stated above, i.e., a statistical analysis at individual gene level is performed followed by an integration of these statistics of genes in the same pathway.

In [11], a hypothesis testing framework for pathway differential analysis is proposed. T-test and Wilcoxon rank test are recommended to measure the difference of expressions of a single gene between two groups of samples. Then this statistic is accumulated over each gene in a particular pathway and standardized by the total number of genes in this pathway. The significance of the result is then interpreted by rejecting two null hypotheses, each with a null population generated by permuting sample labels or gene indices. In this approach, a statistical analysis at individual gene level is still required for the pathway analysis. This approach also belongs to the first category above.

In [12], singular value decomposition is used to obtain pathway activity levels from the gene expression matrix. T-test is applied to the pathway activity levels of the two different sample groups to measure the difference. Significance of the measurement is also obtained by permuting the sample labels. In this approach, no differential analysis at individual gene level is required. However, an extraction of pathway activity level prior to the differential analysis is required. During this extraction process, since only the first eigenvector of singular value decomposition is used, some information of expressions is lost. This approach belongs to the second category stated above.

As discussed above, either t-test or rank sum test is used as a core step by [8] [11] to identify individual genes which are expressed differently from two different sample groups. And thus these methods inevitably inherit the disadvantage of t-test and rank sum test. While the t-test cannot distinguish two sets with close means even though the two sets are significantly different from each others and is very sensitive to extreme values, the rank sum test is not sensitive to absolute values. In turn, those pathways contain genes which can not be identified by t-test or rank sum test but actually are significantly different from two different sample groups will be affected. For example, as showed in Table 1, the expression of Gene 3 are significantly different under two conditions, but was not identified by t-test. Thus, a pathway involving this gene is less likely to be identified by the first category of analysis that uses t-test at the gene level.

In our proposed MCM-test, instead of identifying individual genes first [8] [11] [12], the differential analysis is done directly at the pathway level without individual gene differential statistic. The intuition behind this is based on the fact that genes for each patient interplay with each other. All expression values of genes which belong to a pathway of a particular patient are treated as a vector. MCM-test does not extract activity level of pathways. This allows keeping the maximum amount of information for the pathway differential analysis. Moreover, the fuzzy concept makes the approach more tolerant to individual data item noise.

3. Methodology

In [5] [6] [7], we proposed two fuzzy-set-theory based methods, CM-test and FM-test, to identify the individual genes that expressed significant differences under two conditions. In this paper, by extending the cluster misclassification concept to a multi-dimensional space, we propose a new approach for pathway analysis, the Multi-dimensional Cluster Misclassification (MCM-test). Comparing with CM-test and FM-test, MCM-test looks for a group of genes significant under two conditions instead of identifying significant individual genes under two conditions. In this approach, the expression values of a group of Q genes for a particular sample under a particular condition are considered as a Q-dimension vector. The differential analysis is done at the vector level, without individual gene differential statistic.

In this section, we first introduce the concept of fuzzy membership function of vectors, then the detail of MCM-test.

3.1 Fuzzy membership function of vectors

In fuzzy set theory, the degree for one variable to belong to a fuzzy set is defined by a function. Figure 1 shows a typical fuzzy membership function used for a single variable.

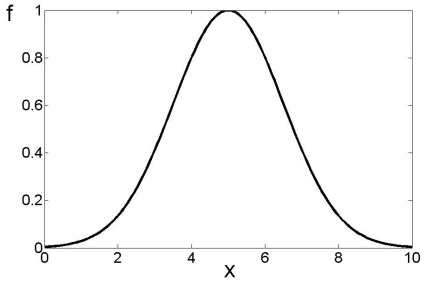

Fig.1. A sample fuzzy membership function of a variable x

For a vector which has two dimensions, the degree that it belongs to a set of vectors can be defined by a

three-dimensional function, with the third dimension being the measure of the membership. Figure 2 shows a sample fuzzy membership function for a vector *(x, y)*.

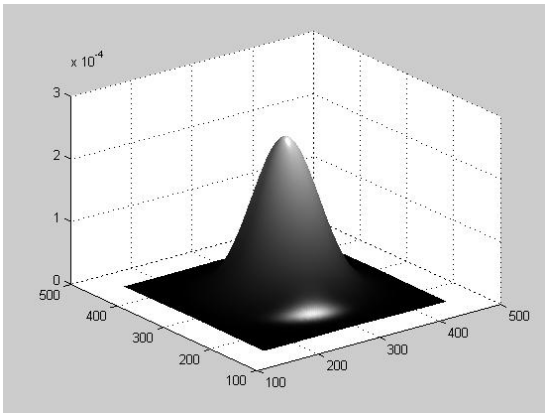

Fig.2. A sample fuzzy membership function of vector (x, y)

For vectors with *n* dimensions, their fuzzy membership function will be *n+1*-dimensional, with one dimension measuring the fuzzy membership.

3.2 Our approach

Consider a pathway that consists of *Q* genes, the problem now is to determine how these *Q* genes are expressed differently under two conditions. To perform this, we consider the expression values of the *Q* genes for a particular sample under a particular condition as a *Q*-dimension vector. Then the expression values of a pathway under one condition *j* can be modeled as set S_j (*j*=1, 2) of points in a *Q*-dimension space. The idea is to consider the two sets of points S_1 and S_2 as samples from two different fuzzy sets. We then examine the membership value of each element with respect to these two fuzzy sets and determine the *d*-value between the two sets of samples.

The mean $\vec{\mu}_j$ of the expression values of set S_j is:

$$\vec{\mu}_j = \frac{1}{N_j} \sum_{\vec{x}_n \in S_j} \vec{x}_n \qquad (1)$$

where,

$$\vec{\mu}_j = \begin{bmatrix} \mu_{j1} \\ \mu_{j2} \\ \dots \\ \mu_{jQ} \end{bmatrix} \text{ and } \vec{x}_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \dots \\ x_{nQ} \end{bmatrix}$$

N_j is the number of samples in S_j, $\vec{x}_n$ is vector made by the expression values of the n-th sample under condition j

We then characterize set S_j (*j*=1, 2) by a fuzzy set FS_j (j=1, 2) whose fuzzy membership function is defined as:

$$f_{FS_j}(\vec{x}\ ) = \exp\left(-\frac{1}{2} (\vec{x}_n - \vec{\mu}_j)^T \Sigma_j^{-1} (\vec{x}_n - \vec{\mu}_j) \right) \quad (2)$$

where,

$$\Sigma_j == \frac{1}{N_j - 1} \sum_{\vec{x}_n \in S_j} (\vec{x}_n - \vec{\mu}_j)(\vec{x}_n - \vec{\mu}_j)^T \quad (3)$$

Given an element $\vec{e}$ in S_1, we calculate its element misclassification degree with respect to FS_2 as

$$m(\vec{e}, FS_2) = Max(f_{FS_2}(\vec{e}) - f_{FS_1}(\vec{e}), 0) \quad (4)$$

We denote the misclassified elements in S_1 with respect to FS_2 as $M_{FS2}(S_1) = \{\vec{e} \mid \vec{e} \in S_1 \wedge m(\vec{e}, FS_2) > 0\}$. Similarly, we denote the misclassified elements in S_2 with respect to FS_1 as $M_{FS1}(S_2) = \{\vec{f} \mid \vec{f} \in S_2 \wedge m(\vec{f}, FS_1) > 0\}$. We denote the number of misclassified elements in S_1 and S_2 with respect to each other as $\#M(S_1, S_2) = |M_{FS2}(S_1)| + |M_{FS1}(S_2)|$. We then define the convergence degree (*c*-value) of S_1 and S_2 as a linear interpolation of the number of misclassified elements and the mutual misclassification degrees as follows.

$$c(S_1, S_2) = \beta * T_1 + (1 - \beta) * T_2 \qquad (5)$$

where,

$$T_1 = \frac{\#M(S_1, S_2)}{S_1 + S_2} \qquad (6)$$

and

$$T_2 = \frac{\sum_{\vec{e} \in S_1} m(\vec{e}, S_2) + \sum_{\vec{f} \in S_2} m(\vec{f}, S_1))}{S_1 + S_2} \quad (7)$$

Then, the divergence between S_1 and S_2 can be calculated using the following:

$$d(S_1, S_2) = 1 - c(S_1, S_2) \quad (8)$$

In our method, to negate the effect of outliers, we used α-trimmed mean instead of normal mean. The α-trimmed mean is calculated by ordering the sample under consideration and taking away the smallest and largest α values from the ordered sample. The mean of the remaining values in the sample is α-trimmed mean of the sample. For instance, if we have a sample of (3, 17, 25, 29, 23, 53, 22, 31, 45, 81, 90, 1), the 2-trimmed mean is calculated by removing the smallest two values (1, 3), and largest two values (81, 90) from the sample set. The mean of the remaining values (30.625) becomes the 2-trimmed mean of the sample.

For computational simplicity, an Epanechnikov function shown as following can be used instead of the Gaussian function of equation (2):

$$f_{FS_j}(\vec{x}_n) = Max\{0, 1 - \frac{\| \vec{x}_n - \vec{\mu}_j \|^2}{\sum_{q=1}^{Q} \sigma_q^2}\} \quad (9)$$

where,

$$\sigma_q = \sqrt{\frac{1}{N_j - 1} \sum_{\vec{x}_n \in S_j} (\vec{x}_n - \vec{\mu}_j)^2} \quad (10)$$

4. Analysis

One dimension: a special case. In this section we analyze MCM-test for it theoretical justification. For the sake of clarity, we start with one dimension, the simplest and special case of multi-dimension. The one dimensional MCM-test corresponds to differential analysis of a single gene.

In figure 3, two distributions, D_1 and D_2 are displayed in blue and red respectively, with mean $\mu_1 = 600$ and standard deviation $\sigma_1 = 50$ for D_1 and $\mu_2 = 700$, $\sigma_2 = 100$ for D_2. The visualization tells us that they are different as they cover different areas and have different shapes. The CM-test, which can be considered as a special case of the MCM-test differentiate them by measuring the differences on the Y axis, which is a combined result of the location difference together with the difference of the variances.

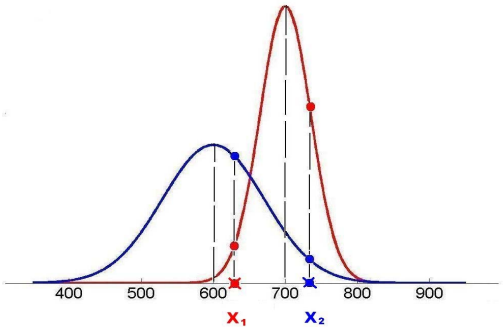

Fig. 3. One dimension Gaussian distributions, μ1 = 600, μ2 = 700, 1 = 50, 2 = 100.

MCM-test uses the probability distribution functions of these two distributions as their fuzzy membership functions respectively, and takes advantage of the membership differences of "misclassified" samples. As shown in figure 3, a sample x_1 of D_2 has a higher degree of belonging to D_1, thus is "misclassified" in MCM-test. This misclassification degree is aggregated with all the other "misclassified" samples of D_2 that are misclassified. Similarly, x_2 of D_1 has a higher degree for D_2, thus is also misclassified. This misclassification degree is also aggregated with all the other misclassified samples of D_1.

MCM-test collects all the misclassified degrees and the number of misclassified samples and form them into an index to measure the divergence of these two distributes. With the mean difference between these two distributions increases, the number of misclassified samples, as well as the aggregated misclassification degree decreases. Thus the MCM *d*-value will decrease. Our experiment on synthetic data proved this in Section 5.

Two and higher dimensions. Figure 4 (a) illustrates samples of two distributions, each of which is a 2-D Gaussian function. In pathway analysis, the *X* and *Y* axis can be the expression data of two individual genes respectively. Figure 4 (b) shows the probability density functions of these two distributions, which can be used as their fuzzy membership functions after multiplying a constant.

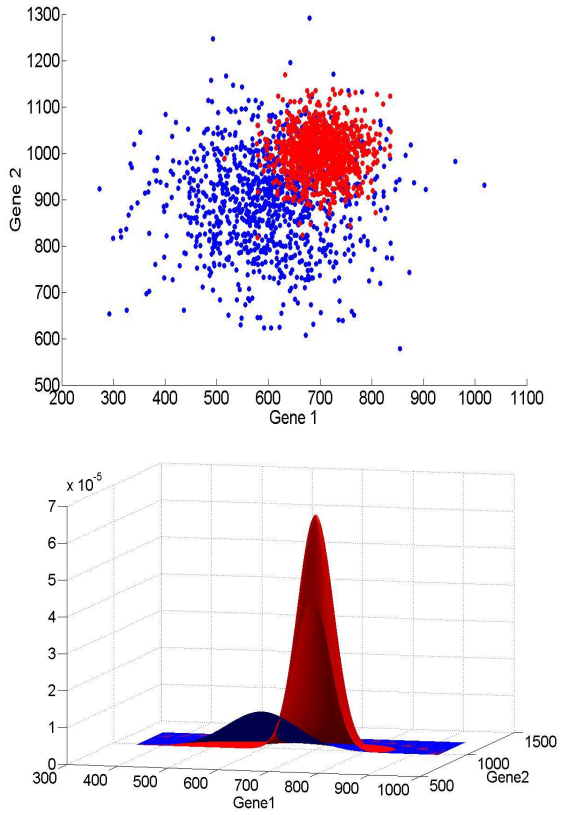

Fig. 4. (a) 1000 samples of 2-D Gaussian distribution μ1x = 600, μ1y = 900, 1x = 1y = 50 and 1000 samples of 2-D Gaussian distribution μ2x = 700, μ2y= 1000, 2x = 2y = 100. (b) Probability density functions of the two distributions.

Distributions of higher dimensions are hard to visualize. But the idea of the misclassification test stays the same. In multi-dimension space, each sample is a vector. And their misclassification degrees are used to measure the divergence of their distributions.

5. Experiment results

To investigate our approach, we conducted experiments on synthetic datasets to find the relationship between MCM *d*-value and the mean difference. We then used the MCM-test on the real world diabetes dataset analyzed by Tomfohr et al. [12] and GSEA [8].

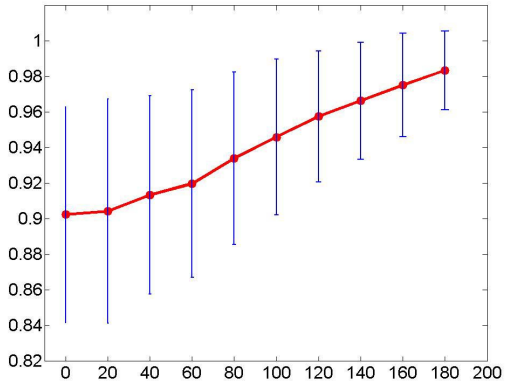

Fig. 5. Relationship between d-value and mean difference

Relationship between MCM d-value and mean difference of the distributions. Suppose two sets S_1 and S_2 are drawn from two different distributions, then a good divergence value will satisfy the following property: the less the overlap, the higher the *d*-value. To validate that our MCM-test has this property, we performed the following steps:

- We generated two datasets from two different distributions. Let $N(\mu, \sigma)$ and Let $N(\mu + x, \sigma)$ be two Gaussian distributions where μ is the mean and σ is the variance. For this experiment, we varied the mean difference as follows: x = 0, 20, 40, 60, 80, 100, 120, 140, 160, and 180.
- We generated 100 pairs of sets of values, with the first set having 17 values generated from $N(\mu, \sigma)$ and the second set also having 17 values generated from $N(\mu + x, \sigma)$. The number 17 was chosen to mimic the real world diabetes dataset used for the analysis in this paper.
- We analyzed this 100 pair of sets of values with MCM-test and obtained the *d*-value.
- We iterated the steps 2 and 3 for 1000 times for each x and averaged the *d*-values over all the iterations.
- We performed the steps 2, 3 and 4 for different values of x. Finally the curve of figure 5 was drawn. Errors of the d-values are also shown.

From figure 5, we can see that the MCM-test has the desired property: the larger the mean difference between two sets, the larger the divergence *d*-value.

Table 2. The results from MCM-test on diabetes dataset

Pathway Name	MCM-test p-value	No. of genes hits in dataset	Actual no. of genes in pathway	Percentage of gene hits
OXPHOS	0.04995	106	114	92.98
c20 U133 probes	0.013	215	270	79.62
human mitoDB	0.029	436	594	73.4
c33 U133 probes	0.021	245	362	67.67
MAP00252	0.022	23	35	65.71
c34 U133	0.012	274	452	60.62
c21 U133	0.026	166	287	57.84
c8 U133	0.013	164	288	56.94

Analyzing the diabetes dataset with MCM-test. The diabetes dataset contains the transcriptional profiles of smooth muscle biopsies of diabetic and normal individuals. In the expression dataset, for each gene, there are 17 expression values from subjects with type 2 diabetes (DM2), 17 expression values from subjects with normal glucose tolerance (NGT) and 10 expression values from subjects with impaired glucose tolerance (IGT). For our analysis, we only used the 34 expression values from subjects with type 2 diabetes and subjects with normal glucose tolerance. Furthermore, we used about 150 pathways obtained from KEGG (Kyoto Encyclopedia of Genes and Genomes) [2].

Filtering and analysis. The expression values in the dataset which are too small, i.e., less than 100 are considered to be the result of noise. So, to minimize the effect of these low values, we only included the genes which have at least one of the expression values greater than 100. Out of the 22,283 genes in the dataset, 10,983 met the filtering criteria. The *d*-value for each pathway was calculated as described in the methodology section before. The *p*-value for the pathway was calculated using permutation test. We permuted the genes 1000 times, each time selecting the same number of genes as that of the pathway under consideration. We then calculated the *d*-value of each pathway obtained thus and the *p*-value for the pathway was the fraction of times the *d*-values of the pathways obtained by 1000 permutation equaled or exceeded the original *d*-value.

Result and discussion. The pathways are ordered in the ascending order of their *p*-values. The significant pathways, i.e., the pathways with *p-value* less than 0.05, are then ordered according to the percentage of the genes in the pathway which were represented in the dataset. Table 2 shows the result after sorting.

Using our method, we identified the deregulation of mitochondrial pathways in the dataset which is in accordance with previous studies. The first cluster of genes involved was from the mitochondrial *OXPHOS* pathway. The *OXPHOS* pathway was well represented in the data with *93%* of genes (106 out of 114) present in the dataset. Oxidative phosphorylation in mitochondria provides energy in the form of ATP generation. In muscle cells, mitochondrial dysfunction has been linked to insulin resistance and type-2 diabetes [3]. The involvement of genes coded by mitochondria in insulin resistance is also well known. The depletion of cellular mitochondrial DNA has been shown to cause insulin resistance in experimental model [9]. Reduced mitochondrial oxidative phosphorylation leads to the accumulation of intracellular triglycerides resulting in insulin resistance. The next 2 clusters, *c20_U133* which is a manually curated cluster of genes coregulated with OXPHOS [8] and the mitochondrial gene cluster *human_mitoDB_6_2002* reinforce that muscle mitochondrial dysfunction is linked to type-2 diabetes.

6. Conclusion and future works

In this paper, we proposed an innovative approach for measuring the significance of gene pathway with microarray data. Experiments have been conducted on both synthetic data sets and real world data. Results on real world diabetes data identified several number of gene pathways. Of note, our top significant pathways were related to mitochondrial function which is well known to be involved in insulin resistance and type-2 diabetes. Our data suggests that the method can be successfully used in pathway level differential analysis of gene expression datasets.

We plan to apply this approach to expression datasets of other diseases to identify pathways relevant to those diseases. We are also planning to further develop this approach for differential analysis in other application domains. As measuring the differences of two groups of data are essential to most researches, our approach can provide a solution to this general and most critical problem.

Acknowledgement

We would like to thank Ying Wang and Togba Liberty for generating some of the figures used in this paper. This work was supported by the Agriculture Experiment Station at the University of the District of Columbia (Project No.: DC-0LIANG; Accession No.: 0203877)

References

[1] Biocarta. http://www.biocarta.com/

[2] "Kyoto Encyclopedia of Genes and Genomes," http://www.genome.jp/kegg/

[3] Morino K, Petersen KF, and Shulman GI, "Molecular mechanisms of insulin resistance in humans and their potential links with mitochondrial dysfunction," *Diabetes*, 55 Supp 2:S9–S15, 2006

[4] M. Kanehisa. Kegg, "From genes to biochemical pathways," In S. Letovsky, editor, Bioinfomatics: Databases and Systems, pages 63–76. Kluwer Academic Publishers, 1999

[5] Lily R. Liang, Shiyong Lu, Yi Lu, Puneet Dhawan, and Deepak Kumar, "CM-test: An innovative divergence measurement and its application in diabetes gene expression data analysis," *In Proc. of the IEEE International Conference on Granular Computing*, pages 262–268, Atlanta, Georgia, USA, May 2006

[6] Shiyong Lu, Lily R. Liang, Xuena Wang, Yi Lu, Vinay Mandal, Dorrelyn Patacsil, and Deepak Kumar, "FM-test: A fuzzy-set-theory-based approach to differential gene expression data analysis," *BMC Bioinformatics*, 7, 2006

[7] Yi Lu, Shiyong Lu, Lily R. Liang, and Deepak Kumar, "FM-test: A fuzzy-set-theory based approach for discovering diabetes genes," *In Proc. of the IEEE International Symposium of Computations in Bioinformatics and Bioscience*, pages 48–55, Hangzhou, Zhejiang, P.R. China, June 2006

[8] V. K. Mootha, C. M. Lindgren, K. F. Eriksson, A. Subramanian, S. Sihag, J. Lehar, P. Puigserver, E. Carlsson, M. Ridderstrale, and et al, "Laurila, E. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nature Genet*, 34:267–273, 2003

[9] Seung Y. Park and Wan Lee, "The depletion of cellular mitochondrial DNA causes insulin resistance through the alteration of insulin receptor substrate-1 in rat myocytes," *Diabetes Res Clin Pract*, 17462778, April 2007

[10] A.S. 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 genomewide expression profiles," *PNAS* ,102:15545-15550, 2005

[11] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, "Discovering statistically significant pathways in expression profiling studies," *PNAS,* 102: 13544–13549, 2005

[12] J. Tomfohr, J. Lu, and T. B. Kepler, "Pathway level analysis of gene expression using singular value decomposition," *BMC Bioinformatics*, 6, 2005