FM-test: A Fuzzy Set Theory Based Approach for Discovering Diabetes Genes[*]

Yi Lu, Shiyong Lu
Wayne State University
{luyi, shiyong}@wayne.edu

Lily R. Liang, Deepak Kumar
University of the District of Columbia
{lliang, dkumar}@udc.edu

Abstract

Diabetes is a disorder of metabolism that has affected 18.2 million people in the United States. In recent years, researchers have identified many genes that play important roles in the onset, development and progression of diabetes. Identification of these diabetes genes offers better understanding of the molecular mechanisms underlying pathogenesis, which is essential for developing preventative and therapeutic methods. In this paper, we propose an innovative approach, fuzzy membership test (FM-test), based on fuzzy set theory to identify diabetes associated genes from microarray gene expression profiles. A new concept of FM d-value is defined to quantify the divergence of two sets of values. Experiments were conducted to study the distribution of d-values and the relationship between the d-value and the significance level of p-value. We applied FM-test to a gene expression dataset obtained from insulin-sensitive and insulin-resistant people and identified ten significant genes. Six of the ten have been confirmed to be associated with diabetes in the literature and one has been suggested by other researchers. The remaining three genes, $D85181$,$M95610$ *and* $U06452$*, are suggested as potential diabetes genes for further biological investigation.*

1 Introduction

Diabetes is a group of diseases characterized by high levels of blood glucose resulting from defects in insulin production, insulin action, or both. There are 18.2 million people in the United States, or 6.3% of the population, who have diabetes. Diabetes is also one of the leading causes of death in U.S. In 2000, it contributed to 213,062 deaths. The risk for death among people with diabetes is about 2 times of that among people without diabetes [1]. The direct and indirect cost of diabetes in the United States for 2002 totaled $132 billion, among which, $92 billion are direct medical costs and $40 billion are indirect costs of disability, work loss, premature mortality etc[1].

Microarray techniques have revolutionalized genomic research by making it possible to monitor the expression of thousands of genes in parallel. As the amount of microarray data being produced in an exponential rate, there is a great demand for efficient and effective expression data analysis tools. The gene expression profile of a cell determines its phenotype and responses to the environment. These responses include its responses towards environmental factors, drugs and therapies. Gene expression patterns can be determined by measuring the quantity of the end product, protein, or the mRNA template used to synthesize the protein. Comparison of gene expression profiling in diabetes patients versus the normal counterpart people will enhance our understanding of the disease and identify leads for therapeutic intervention. Several important breakthroughs and progress in the gene expression profiling of diabetes have been made [10, 14, 13]. Patterns of gene expression have been proposed and associated with diabetes [15, 16]. More interestingly, researchers have identified many genes that play important roles in the onset, development, and progression of diabetes. Identification of these diabetes genes offers a route to better understanding of the molecular mechanisms underlying pathogenesis, a necessary prerequisite for the rational development of improved preventative and therapeutic methods.

One effective approach of identifying genes that are associated with diabetes is to measure the divergence of two sets of values of gene expression, one from a group of people that are insulin resistant (IR), the other from a group that are insulin sensitive (IS) [17]. A motivating example is shown in Table 1, which records the microarray gene expression values of five genes for two groups of people: five insulin-sensitive humans and five insulin-resistant humans. In order to identify the genes that are associated with diabetes, one needs to determine for each gene whether or not the two sets of expression values are significantly different from each other. One popular method is t-test [11], which uses the difference of the means of the two sets to measure the divergence. In Table 1, the first four genes are iden-

[*] This work was supported by the Agricultural Experiment Station at the University of the District of Columbia (Project No.: DC-0LIANG; Accession No.: 0203877)

Table 1. The microarray gene expression values for five genes under two conditions

Gene ID	IR					IS					d-value	FM-test p-value	t-test p-value	rank sum p-value
1	750	559	649	685	636	310	359	135	97	178	0.999	0.001	0.008	0.00
2	123	142	11	406	220	305	398	707	905	688	0.756	0.012	0.011	0.031
3	246	213	232	134	67	86	79	77	94	61	0.725	0.017	0.021	0.098
4	200	191	220	83	197	49	81	116	111	135	0.708	0.019	0.024	0.058
5	598	424	695	451	141	342	260	266	229	234	0.674	0.025	0.077	0.152

tified by t-test as significant genes (with p-value ≤ 0.05). However, t-test cannot distinguish two divergent sets with close means and is very sensitive to extreme values. As a result, t-test fails to recognize genes 5 as significant genes although their expression values under the two conditions are significantly different from each other. Another popular method is Wilcoxon rank sum test [11], which uses the sum of ranks for one of the sets to measure the divergence. In Table 1, the first two genes are identified by Wilcoxon rank sum test as significant genes (with p-value ≤ 0.05). Although Wilcoxon rank sum test overcomes the limitation of t-test on the sensitivity to extreme values, it is not sensitive to absolute values. As a result, Wilcoxon rank sum test fails to identify gene 3, 4 and 5 as significant genes although the two sets for these three genes are significantly different from each other.

In this paper, based on the fuzzy set theory [8], we propose an innovative approach that overcomes the above limitations of t-test and rank sum test. The basic idea is to consider the two sets of values as samples from two different fuzzy sets. We examine each element in one fuzzy set for its membership value with the other fuzzy set. Each such membership value is considered as a bond between the two sets. An aggregation of the values of all elements represent the overall bond between the two sets. By standardizing the aggregation and then subtracting the result from 1, we measure the divergence of the original two sets. The weaker this overall bond is, the more divergent the two sets are, and the more significant the corresponding gene is.

The main contributions of this paper are:

1. We propose an innovative approach based on the fuzzy set theory, the *fuzzy membership test* (FM-test), which quantifies the divergence of two sets directly.

2. We validated FM-test on synthetic datasets and show that it is effective and robust.

3. We apply FM-test to a real diabetes gene expression dataset and identified 10 significant genes. Six among these ten have been known to be associated with diabetes, one is suggested by other researchers as potential diabetes genes, and we suggest the remaining three genes for further biological investigation.

The rest of the paper is organized as follows. Section 2 briefly reviews t-test and Wilcoxon rank sum test and their limitations. Section 3 presents our fuzzy-set-theory-based method, FM-test. Section 4 provides our experimental results on both synthetic datasets and a real dataset of gene expression profiles. In the end, Section 5 concludes the paper and points out some potential future work.

2 Related work

Many genes have been identified to be important in the onset, development, and progression of diabetes. One effective approach of identifying genes that are associated with diabetes is to measure the divergence of two sets of values of gene expression, each from a group of people with a particular condition [17]. Two most popular methods to measure the divergence of two sets of values are t-test [11] and Wilcoxon rank sum test [11],

The statistical method t-test assesses whether the means of two groups are statistically different from each other. Given two sets S_1 and S_2, the t-value is calculated as

$$t(S_1, S_2) = \frac{\mid \mu_{S_1} - \mu_{S_2} \mid}{\sqrt{\frac{\sigma^2_{S_1}}{|S_1|} + \frac{\sigma^2_{S_2}}{|S_2|}}} \tag{1}$$

where μ_S and σ_S are the sample mean and standard deviation of S, respectively.

The limitation of t-test is that it cannot distinguish two sets with close means even though the two sets are significantly different from each other. Another limitation of t-test is that it is very sensitive to extreme values.

Another popular statistical method is Wilcoxon rank sum test, which can be used to test the null hypothesis that two sets S_1 and S_2 have the same distribution. We first merge the data from these two sets and rank the values from the lowest to the highest with all sequences of ties being assigned an average rank. The Wilcoxon test statistic W is the sum of the ranks from set S_1. Assuming that the two sets have the same continuous distribution (and no ties occur), then W has a mean and standard deviation given by

$$\mu = \frac{m * (m + n + 1)}{2} \tag{2}$$

$$\sigma = \sqrt{\frac{m * n * (m + n + 1)}{12}}, \quad (3)$$

where $m = \mid S_1 \mid$ and $n = \mid S_2 \mid$.

We test the null hypothesis H_o: no difference in distributions. A one-sided alternative is H_a: S_1 yields lower measurements. We use this alternative if we expect or see that W is unusually lower than its expected value μ. In this case, the p-value is given by a normal approximation. We let $N \sim N(\mu, \sigma)$ and compute the left-tail $Pr(N \leq W)$ (using continuity correction if W is an integer).

If we expect or see that W is much higher than its expected value, then we should use the alternative H_a: first S_1 yields higher measurements. In this case, the p-value is given by the right-tail $Pr(N \geq W)$. If the two sums of ranks from each set are close, then we could use a two-sided alternative H_a: there is a difference in distributions. In this case, the p-value is given by twice the smallest tail value $2Pr(N \leq W)$ if $W < \mu$, or $2Pr(N \geq W)$ if $W > \mu$.

Although rank sum test overcomes the limitation of t-test on sensitivity on extreme values, it is not sensitive to absolute values. This might be advantageous to some applications but not to others.

3 Methodology

In this section, based on the fuzzy set theory [8], we present our innovative approach, the fuzzy-set-theory-based method test (FM-test), to quantify the divergence of two sets of values directly and robustly.

Let S_1 and S_2 be two sets of values of a particular feature for two groups of samples under two different conditions. The basic idea is to consider the two sets of values as samples from two different fuzzy sets. We examine the membership value of each element with respect to the other fuzzy set. By calculating the average of membership values, we measure the divergence of the original two sets. In particular, we perform the following steps:

1. Compute the sample mean and standard deviation of S_1 and of S_2 respectively.

2. Characterize S_1 and S_2 as two fuzzy sets FS_1 and FS_2 whose fuzzy membership functions, $f_{FS_1}(x)$ and $f_{FS_2}(x)$, are defined with the sample means and standard deviations. The fuzzy membership function $f_{FS_i}(x)$ $(i = 1, 2)$ maps each value x to a fuzzy membership value that reflects the degree of x belonging to FS_i $(i = 1, 2)$.

3. Using the two fuzzy membership functions, $f_{FS_1}(x)$ and $f_{FS_2}(x)$, quantify the convergence degree of two sets.

4. Define the divergence degree (FM d-value) between the two sets based on the convergence degree.

The details of each step is elaborated in the sequel.

3.1 Fuzzy Sets and Membership Functions

The sample mean μ_1 of S_1 is calculated as

$$\mu_1 = \frac{1}{n_1} \sum_{x_i \in S_1} x_i \quad (4)$$

where n_1 is the number of elements in S_1, and the sample standard deviation σ_1 of S_1 is calculated as

$$\sigma_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{x_i \in S_1} (x_i - \mu_1)^2} \quad (5)$$

For gene 5 in Table 1, we have $\mu_1 = 461.8$, $\sigma_1 = 210.59$, $\mu_2 = 266.2$, and $\sigma_2 = 45.29$. We then characterize set S_1 by a fuzzy set FS_1 whose fuzzy membership function is defined as

$$f_{FS_1}(x) = e^{-(x-\mu_1)^2/(2\sigma_1^2)} \quad (6)$$

The function $f_{FS_1}(x)$ maps each value x in S_1 to a fuzzy membership value to quantify the degree that x belongs to FS_1. A value equal to the mean has a membership value of 1 and belongs to fuzzy set FS_1 to a full degree; a value that deviates from the mean has a smaller membership value and belongs to FS_1 to a smaller degree. The further the value deviates from the mean, the smaller the fuzzy membership value. Similarly, the fuzzy membership function for S_2 is defined as

$$f_{FS_2}(x) = e^{-(x-\mu_2)^2/(2\sigma_2^2)} \quad (7)$$

where μ_2 and σ_2 are the mean and standard deviation of S_2 respectively.

For gene 5 in Table 1, we have $f_{FS_1}(x) = e^{-(x-461.8)^2/88696.3}$ and $f_{FS_2}(x) = e^{-(x-266.2)^2/4102.4}$. With these two fuzzy membership functions, the fuzzy membership values for each element with respect to the two sets can be calculated. For example, $f_{FS_1}(598) = 0.81$ and $f_{FS_2}(598) = 2.2E^{-12}$.

3.2 Our Proposed Method: FM-test

Since the fuzzy membership functions can overlap, one element can belong to more than one fuzzy set with a respective degree for each. For an element in S_1, we measure the degree that it belongs to FS_1 by applying its value to f_{FS_1}. Similarly we can apply its value to f_{FS_2} to measure the degree that it belongs to FS_2. The idea of FM-test is to consider the membership value of an element in S_1 with respect to S_2 as one bond between S_1 and S_2, and vice versa, then the aggregation of all these bonds reflects the overall bond between these two sets. The weaker this overall bond is, the more divergent these two sets are. The strength of the overall bond between two sets is quantified by their c-value, which aggregates the mutual membership values of elements in S_1 and S_2 and is defined as follows.

Definition 3.1 (FM c-value) Given two sets S_1 and S_2, the convergence degree between S_1 and S_2 in FM-test is defined as

$$c(S_1, S_2) = \frac{\sum_{e \in S_1} f_{F(S_2)}(e) + \sum_{f \in S_2} f_{F(S_1)}(f)}{\mid S_1 \mid + \mid S_2 \mid} \tag{8}$$

□

Now we define the divergence value in FM-test (FM d-value) as follows.

Definition 3.2 (FM d-value) Given two sets S_1 and S_2, the FM d-value between S_1 and S_2 is defined as

$$d(S_1, S_2) = 1 - c(S_1, S_2) \tag{9}$$

□

For gene 5 in Table 1, $c(S_1, S_2) = 0.326$, thus the divergence value is $1 - c(S_1, S_2) = 0.674$. We calculated all the p-values for the five genes in Table 1 for the three methods. One interesting observation is that, while both t-test and Wilcoxon rank sum test fail to recognize gene 5 as a significant gene since their p-values are greater than 0.05, our FM-test identifies gene 5 as a significant gene with a p-value of 0.025. The reason of the failure of t-test and Wilcoxon rank sum test is due to their sensitivity to the extreme value 141 in the first set of the gene.

Given a calculated FM d-value D for two sets S_1 and S_2, to interpret D in terms of "significantly divergent" or not, we need to know the cutoff value δ of D, so that when $D \geq \delta$, the two sets are interpreted as significantly divergent. In the context of FM-test, we like to test the following null hypothesis H_o: S_1 and S_2 originate from the same distribution. Then the p-value is defined as the probability $\{Pr(d(S_1, S_2) \geq D \mid S_1$ and S_2 were randomly sampled from the same distribution$\}$. As a convention of statistical analysis, if $p - value \leq 0.05$, then it is a strong evidence to reject the null hypothesis, and accepts that the two sets are significantly divergent, while the p-value reflects the significance. It has been very common to use Monte Carlo procedures to calculate the empirical p-value which approximates the exact p-value without relying on asymptotic distributional theory or on exhaustive enumeration. Davison and Hinkley [5] present the formula for obtaining an empirical p-value as $(n + 1)/(N + 1)$, where N is the number of samples in the data set, and n is the number of those samples which produce the statistical value greater than or equal to the specified value.

We perform the following steps to calculate the p-value of two sets S_1 and S_2 with their FM d-value D: (1) Estimate the distribution that S_1 and S_2 are drawn from a normal distribution $N(\mu, \sigma)$, where μ and σ are estimated using the sample mean and standard deviation of $S_1 \cup S_2$; (2) Randomly draw N pairs of sets from $N(\mu, \sigma)$, then calculate the FM d-value for each pair; (3) Calculate the empirical p-value as $(n + 1)/(N + 1)$, where n is the number of pairs whose FM d-values are equal or greater than D. The cutoff FM d-value we obtain in this way is introduced in the next section.

4 Experimental Results and Discussion

To validate our approach, first, we investigated the distribution of FM d-value on a set of synthetic datasets. Second, we conducted experiments on a synthetic dataset to study the relationship between FM-test d-value and its empirical p-value. Third, on another synthetic dataset, we studied the relationship between FM d-value and the mean difference of distributions. Finally we conducted FM-test on a real microarray dataset of diabetes gene expressions to identify genes that are related to diabetes and insulin metabolism.

4.1 The probability distribution of FM d-value

Suppose two sets S_1 and S_2 are randomly drawn from the same normal distribution, what is the probability distribution of FM d-value? To answer this question, we conducted the following simulation:

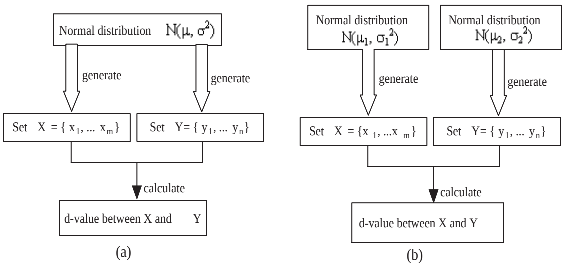

Figure 1. Random generation of d-value from normal distribution

1. We generated N = 64000 pairs of sets of values, with each set containing 5 values. As shown in Figure 1(a), each value in the two data sets is randomly generated from the same normal distribution $N(0, 1)$.

2. We calculated the d-value for each pair of sets.

3. We then estimated the probability density value $f(d) = \frac{|\{i|d-\delta<d_i\leq d+\delta\}|}{N*2\delta}$ where $\delta = 0.005$. The value is essentially the fraction of the FM d-values falling in region $[d-\delta, d+\delta]$ divided by the region length 2δ. The probability density function of the d-distribution was drawn in Figure 2.

4. Finally, in order to understand the effect of the number of pairs used for simulation, i.e., the size of the dataset, on the approximation error of the d-distribution, we generated datasets with different data sizes. For each data size, we generated 10 datasets, and thus derived 10 probability density functions. The maximum standard deviation for all d-values is recorded as *the error rate* for that data size. As shown in Figure 3, the error rate decreases as the size of the dataset increases.

From Figure 2, we can see that most FM d-values fall into the range from 0.2 to 0.5, and very few fall into the range greater than 0.6, or less than 0.2. In particular, when $d \geq 0.6056$, $p-value \leq 0.05$. This is reflected in the red-shared area in Figure 2 with $\int_{0.6056}^{1.0} f(x)dx = 0.05$. Therefore, given two sets S_1 and S_2 drawn from the same normal unit distribution, the chance that the pair has a FM d-value equal to or greater than 0.6056 is very low. On the other hand, if we observe that two sets have a d-value equal to or greater than 0.6056, then there is a strong evidence that these two sets are drawn from two different distributions, and thus considered as significantly divergent.

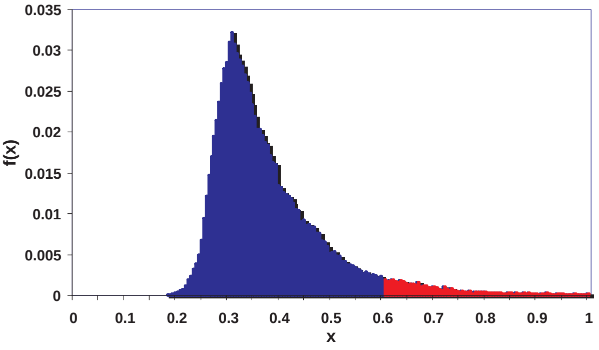

Figure 2. The probability density function of FM d-value

Figure 3 shows the effect of data size on the error rate of the derived probability density function. As the data size increases, the error rate decreases. We can see from Figure 3 that, after the number of pairs of sets in a dataset is greater than 8000, the trend of the error rate becomes stable. Thus, to obtain a reliable empirical p-value for FM-test, the data size should be greater than 8000.

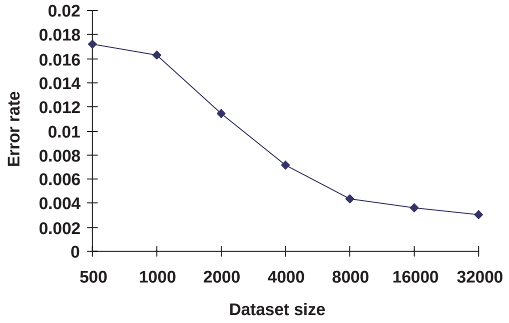

Figure 3. The impact of dataset size on error rate of PDF of FM d-value

4.2 Relationship between FM d-value and empirical p-value

Suppose two sets S_1 and S_2 are drawn from the same normal distribution, then what is the probability that they have a FM d-value equal to or greater than a particular D? is the D increases, will this probability decrease? To answer these questions, we studied the relationship between FM d-value and empirical p-value as follows:

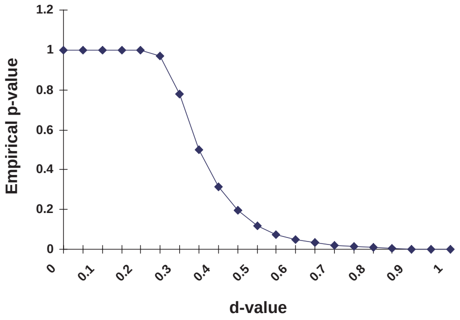

Figure 4. The relationship between FM d-value and empirical p-value

1. Based on the above experimental result, we know that we need at least 8000 pairs of sets to obtain a reliable empirical p-value. Therefore, in this experiment, we generated 10000 pairs of sets of values, with each set containing 5 values. Each value is randomly generated from the unit normal distribution $N(0,1)$.

2. We calculated the d-value for each pair of sets.

3. For each pair of sets S_1 and S_2 with d-value D, we calculated its empirical p-value as $n+1/10001$ where n is the number of pairs in these 10000 pairs that have a d-value equal to or greater than D.

4. We drew the relationship between d-value and empirical p-value in Figure 4.

From Figure 4, we can see that as d-value increases, the p-value decreases. In particular, when $d \geq 0.6056$, we have $p \leq 0.05$.

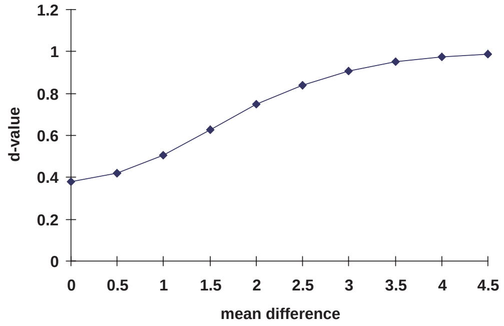

Figure 5. Relationship between the mean difference of distributions and d-value

4.3 Relationship between the mean difference of distributions and d-value

Suppose two sets S_1 and S_2 are drawn from two different distributions, then a good divergence measurement should satisfy the following property: the less overlap between these two distributions, the greater the d-value. We validated that our FM-test has this property as follows:

1. As shown in Figure 1(b), two data sets are generated from two distribution. Let $N(0,1)$ and $N(x,1)$ be two normal distributions, where x is the mean difference between these two distributions. In this experiment, we consider x = 0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, respectively.

2. We generated 1000 pairs of sets of values, with the first set containing 5 values that are randomly generated from $N(0,1)$, and the second set containing 5 values that are randomly generated from $N(x,1)$.

3. We calculated the d-value for each pair. Let the average of these 1000 d-values be d. We then plotted (x,d) in Figure 5.

4. We repeated step 2 and 3 for different x. Finally, the curve was drawn in Figure 5.

Figure 5 confirmed the desirable property of FM-test: the larger the mean difference between the two distributions, the greater the d-value.

4.4 Applying FM-test to Diabetes Gene Expression Analysis

A dataset of microarray gene expression for a total of 10831 genes downloadable from [17] is used in this experiment. For each gene, there are ten expression values, five from a group of insulin-sensitive (IS) people and five from a group of insulin-resistant (IR) people. Only the genes that have no null expression values are included in this analysis. We also require that, for a gene to be included, at least five out of its ten expression values are greater than 100. This eliminates the genes whose expression values are noisy and not reliable.

The results of FM-test are compared with the results of t-test and rank sum test. As we can seen in Table 2, although the orders of ranking are different for different methods, all three methods identify these genes as significantly differentially expressed between the IS and IR groups. Furthermore, 10 worst ranked genes in FM-test shown in Table 2 are also consistent with the result of the other two methods. However, gene U49835 is identified by FM-test as the 21st ranked significant gene with p-value 0.0258, neither t-test (with p-value 0.0768) nor rank sum test (with a p-value 0.1522) identifies this gene as significant.

To study the relevance of genes in insulin metabolism and diabetes, the 10 best ranked differentially regulated genes shown in Table 2 were further searched in published literature. Human phosphatidylinositol (4,5)bisphosphate 5-phosphatase homolog (gene U45973) was found to be differentially expressed in insulin resistance cases. Over-expression of inositol polyphosphate 5-phosphatase-2 SHIP2 has been shown to inhibit insulin-stimulated phosphoinositide 3-kinase (PI3K) dependent signaling events. Analysis of diabetic human subjects has revealed an association between SHIP2 gene polymorphism and type 2 diabetes mellitus. Also knockout mouse studies have shown that SHIP2 is a significant therapeutic target for the treatment of type-2 diabetes as well as obesity [6]. Csermely et al. reported that insulin mediates phosphorylation/dephosphorylation of nucleolar protein nucleolin (gene $M60858$) by stimulating casein kinase II, and this may play a role in the simultaneous enhancement in RNA efflux from isolated, intact cell nuclei [4]. c-myc is an oncogene that codes for transcription factor Myc that along with other binding partners such as MAX plays an important role widely studied in various physiological processes including tumor growth in different cancers. Myc modulates the expression of hepatic genes and counteracts the obesity and insulin resistance induced by a high-fat diet in trans-

Table 2. Ten best-ranked and worst-ranked genes identified by FM-test

Probe Set	Gene Description	d-value	Empirical p-value	t-test p-value	rank sum p-value
U45973	Human phosphatidylinositol (4,5) bisphosphate	0.999	0.0003	0.0001	0.0076
M60858	Human nucleolin gene	0.935	0.0016	0.0017	0.0076
D85181	Homo sapiens mRNA for fungal sterol-C5-desaturase homolog	0.892	0.0028	0.0029	0.0147
M95610	Human alpha 2 type IX collagen (COL9A2) mRNA	0.872	0.0038	0.0066	0.0076
L07648	Human MXI1 mRNA	0.858	0.0043	0.0052	0.0076
L07033	Human hydroxymethylglutaryl-CoA lyase mRNA	0.855	0.0046	0.0054	0.0076
X53586	Human mRNA for integrin alpha 6	0.851	0.0047	0.0075	0.0076
X81003	Homo sapiens HCG V mRNA	0.791	0.0089	0.0077	0.0076
X57959	ribosomal protein L7	0.767	0.0108	0.0109	0.0313
U06452	melan-A	0.756	0.0126	0.0118	0.0311
X82324	POU domain, class 3, transcription factor 4	0.206	0.9987	0.407	1
M14764	nerve growth factor receptor (TNFR superfamily, member 16)	0.204	0.9989	0.652	1
M64673	heat shock transcription factor 1	0.204	0.9990	0.652	0.844
U20657	ubiquitin specific peptidase 4 (proto-oncogene)	0.197	0.9993	0.642	0.844
D17793	aldo-keto reductase family 1, member C3	0.196	0.9999	0.471	0.839
D78014	dihydropyrimidinase-like 3	0.194	1	0.620	0.548
AB002314	PDZ domain containing 10	0.191	1	0.367	0.545
L20348	oncomodulin	0.181	1	0.405	0.544
D50063	proteasome (prosome, macropain) 26S subunit	0.179	1	0.544	0.421

genic mice overexpressing c-myc in liver [3]. Max interactor protein, MXI1 (gene L07648) competes for MAX thus negatively regulates MYC function and may play a role in insulin resistance. In the presence of glucose or glucose and insulin, leucine is utilized more efficiently as a precursor for lipid biosynthesis by adipose tissue. It has been shown that during the differentiation of 3T3-L1 fibroblasts to adipocytes, the rate of lipid biosynthesis from leucine increases at least 30-fold and the specific activity of 3-hydroxy-3-methylglutaryl-CoA lyase (gene L07033), the mitochondrial enzyme catalyzing the terminal reaction in the leucine degradation pathway, increases 4-fold during differentiation [7]. Schottelndreier et al. [12] have described a regulatory role of integrin alpha 6 (gene $X53586$) in Ca2+ signaling, that is known to have a significant role in insulin resistance [9]. HCGV gene product (gene $X81003$) is known to inhibit the activity of protein phosphatase-1, which is involved in diverse signaling pathways including insulin signaling [18]. Human ribosomal protein L7 (Gene X57959) plays a regulatory role in eukaryotic translation apparatus. It has been shown to be an autoantigen in patients with systemic autoimmune diseases, such as systemic lupus erythematosus [2]. Identification of this gene in our analysis and by [17] suggests a possible role of this gene in insulin resistance. Published reports on these genes indicate their roles in insulin signaling and warrant further investigations on their functions in insulin resistance cases. We further recommend genes D85181, M95610 and U06452 as candidate genes for future research in this area.

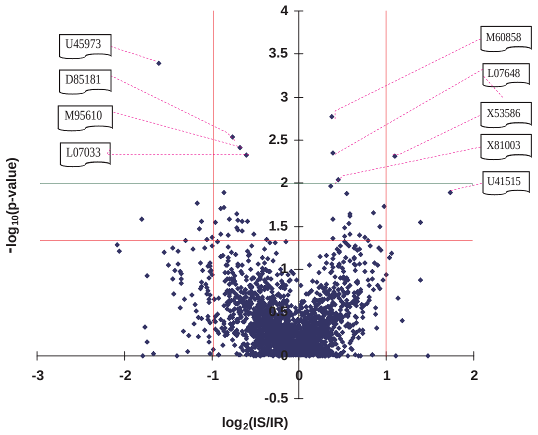

Figure 6. The volcano plot for the diabetes dataset

In order to compare the fold change of expression levels between the IS and IR groups to the statistical significance p-values, we presented all the genes in the diabetes dataset with a volcano plot shown in Figure 6. The volcano plot arranges the genes along dimensions of biological and statistical significance. The X axis is the fold change between the two groups, which is on a log scale $log_2(\bar{IS}/\bar{IR})$, where $\bar{IS}$ is the mean of expressions in the IS group, and $\bar{IR}$ is the mean of the expressions in the IR group. In this way, up and down regulation appear symmetric. The Y axis represents the p-value for our FM-test, which is on a negative log scale $-log_{10}(p - value)$, so that smaller p-values appear higher up. The X axis indicates biological impact of the change; the Y axis indicates the statistical evidence, or reliability of the change. As shown in Figure 6, gene $U45973$ is identified by FM-test as the most statistically significant gene and it is over-expressed in the IR group; gene $X53586$ is identified by FM-test as the 7th statistically significant gene and it is over-expressed in the IS group. Although genes $M60858$, $D85181$, $M95610$, $L07648$, $L07033$, and $X81003$ have been identified by FM-test among the top ten significant genes, they are not overexpressed in either groups. Finally, gene $U41515$ is identified by FM-test as the 11th significant gene and it is over-expressed in the IS group.

In summary, out of the top 10 genes identified by FM-test, we could find 6 of them in published literature about their association with insulin metabolism and diabetes. Among the remaining four genes, gene $X57959$ has been recommended by [17] as a candidate gene for diabetes, we recommend that gene $D85181$, $M95610$ and $U06452$ could serve as candidate genes for future research in this area.

5 Conclusions and Future Work

We proposed an innovative approach based on the fuzzy set theory, FM-test, that quantifies the divergence of two sets directly. We have validated FM-test on synthetic datasets and show that it is effective and robust. We also applied FM-test to a real diabetes dataset and identified 10 significant genes. While six of them have been confirmed to be associated with insulin signaling and/or diabetes in the literature, one has been recommended by others, the remaining three genes, $D85181$, $M95610$ and $U06452$, are suggested as three potential diabetes genes involved in insulin resistance for further biological investigation. Further investigation is needed to identify the properties of distribution of FM d-value and the equation to calculate its p-value. FM-test will soon be freely available at website `http://database.cs.wayne.edu/bioinformatics`.

References

[1] *American Diabetes Association*. `http://www.diabetes.org/`.

[2] Characterization of eukaryotic protein l7 as a novel autoantigen which frequently elicits an immune response in patients suffering from systemic autoimmune disease. *Immunobiology*, Dec. 1994.

[3] Overexpression of c-myc in the liver prevents obesity and insulin resistance. *FASEB J.*, (12):1715–7, Sept. 2003.

[4] P. Csermely, T. Schnaider, B. Cheatham, M. Olson, and C. Kahn. Insulin induces the phosphorylation of nucleolin. a possible mechanism of insulin-induced rna efflux from nuclei. *J Biol Chem*, 268(13):9747–52, 1993.

[5] A. Davison and D. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, Cambridge, 1997.

[6] J. Dyson, A. Kong, F. Wiradjaja, M. Astle, R. Gurung, and C. Mitchell. The SH2 domain containing inositol polyphosphate 5-phosphatase-2: SHIP2. *Int J Biochem Cell Biol*, 37(11):2260–5, 2005.

[7] F. Frerman, J. Sabran, J. Taylor, and S. Grossberg. Leucine catabolism during the differentiation of 3t3-l1 cells. expression of a mitochondrial enzyme system. *J Biol Chem*, 258(11):7087–93, 1983.

[8] G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice-Hall, Upper Saddle River, CA, 1995.

[9] R. Kulkarni, M. Roper, G. Dahlgren, D. Shih, L. Kauri, J. Peters, M. Stoffel, and R. Kennedy. Islet secretory defect in insulin receptor substrate 1 null mice is linked with reduced calcium signaling and expression of sarco(endo)plasmic reticulum ca2+-atpase (serca)-2b and -3. *Diabetes*, 53(6):1517–25, 2004.

[10] S. Rome, K. Clement, R. Rabasa-Lhoret, E. Loizon, C. Poitou, G. Barsh, J. Riou, M. Laville, and H. Vidal. Microarray profiling of human skeletal muscle reveals that insulin regulates approximately 800 genes during a hyperinsulinemic clamp. *J Biol Chem*, 278(20):18063–8, 2003.

[11] B. Rosner. *Fundamentals of Biostatistics*. Duxbury Press, Pacific Grove, CA, fifth edition, 2000.

[12] H. Schottelndreier, B. Potter, G. Mayr, and A. Guse. Mechanisms involved in alpha6beta1-integrin-mediated ca(2+) signalling. *Cell Signal*, 13(12):895–9, 2001.

[13] S. E. SE, Q. Ruan, C. Collins, P. Yang, R. McIndoe, A. Muir, and J. She. Molecular pathways altered by insulin b9-23 immunization. *Ann N Y Acad Sci*, 1037:175–185, Dec. 2004.

[14] A. Shalev, C. Pise-Masison, M. Radonovich, S. Hoffmann, B. Hirshberg, J. Brady, and D. Harlan. Oligonucleotide microarray analysis of intact human pancreatic islets: identification of glucose-responsive genes and a highly regulated tgfbeta signaling pathway. *Endocrinology*, (9):3695–8, Sept. 2002.

[15] A. Shalev, C. Pise-Masison, M. Radonovich, S. Hoffmann, B. Hirshberg, J. Brady, and D. Harlan. Oligonucleotide microarray analysis of intact human pancreatic islets: identification of glucose-responsive genes and a highly regulated tgfbeta signaling pathway. *Endocrinology*, (9):3695–8, Sept. 2002.

[16] R. Sreekumar, P. Halvatsiotis, J. Schimke, and K. Nair. Gene expression profile in skeletal muscle of type 2 diabetes and the effect of insulin treatment. *Diabete*, (6):1913–20, Jun 2002.

[17] X. Yang, R. Pratley, S. Tokraks, C. Bogardus, and P. Permana. Microarray profiling of skeletal muscle tissues from equally obese, non-diabetic insulin-sensitive and insulin-resistant pima indians. *Diabetologia*, 45:15841593, 2002.

[18] J. Zhang, L. Zhang, S. Zhao, and E. Lee. Identification and characterization of the human hcg v gene product as a novel inhibitor of protein phosphatase-1. *Biochemistry*, 37(47):16728–34, 1998.