

# 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

This work was supported by the Agricultural Experiment Station at the University of the District of Columbia

## Abstract

**Problem:** Diabetes has affected 18.2 million people in the United States. Many genes have been found to play important roles in development of diabetes. Identification of these diabetes genes are very important for developing preventative and therapeutic methods.

**Methods:** 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.

**Results:** 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 are suggested as potential diabetes genes for further biological investigation.

## 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 who have diabetes. Diabetes is also one of the leading causes of death in U.S. Microarray techniques have revolutionized 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. 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. 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). 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. 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. We applied FM-test to a gene expression dataset obtained from insulin-sensitive and insulin-resistant people and identified ten significant genes (see Table 2). 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 are suggested as potential diabetes genes for further biological investigation.

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

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) biphosphate	0.999	0.0003	0.001	0.0076
M60858	Human nucleolin gene	0.935	0.0016	0.0017	0.0076
D85181	Homo sapiens mRNA for fungal sterol C-5-desaturase homolog	0.892	0.0028	0.0029	0.0147
N95610	Human alpha 2 type IX collagen (COL9A2) mRNA	0.872	0.0038	0.0066	0.0076
L07648	Human MX11 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
M04673	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 G3	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	oncosmodin	0.181	1	0.405	0.544
D50063	proteasome (prosome, macropain) 26S subunit	0.179	1	0.544	0.421

## Methods

In this section, based on the fuzzy set theory, 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 the two sets.
2. Characterize the two sets as two fuzzy sets and whose fuzzy membership functions are defined with the sample means and standard deviations.
3. Using the two fuzzy membership functions, we 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 these steps are as follows.

The sample mean  $\mu_i$  of  $S_i$  is calculated as

$$\mu_i = \frac{1}{n_i} \sum_{x \in S_i} x_i$$

The sample standard deviation  $\sigma_i$  is calculated as

$$\sigma_i = \sqrt{\frac{1}{n_i - 1} \sum_{x \in S_i} (x_i - \mu_i)^2}$$

We then characterize set  $S_i$  by a fuzzy set whose fuzzy membership function is defined as

$$f_{S_i}(x) = e^{-(x - \mu_i)^2 / (2\sigma_i^2)}$$

The function  $f_{S_i}$  maps each value  $x$  in  $S_i$  to a fuzzy membership value to quantify the degree that  $x$  belongs to. A value equal to the mean has a membership value of 1 and belongs to fuzzy set  $S_i$  to a full degree; a value that deviates from the mean has a smaller membership value and belongs to  $S_i$  to a smaller degree. The further the value deviates from the mean, the smaller the fuzzy membership value.

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.

$$c(S_1, S_2) = \frac{\sum_{x \in S_1} f_{S_2}(x) + \sum_{x \in S_2} f_{S_1}(x)}{|S_1| + |S_2|}$$

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

## Experimental Results and Discussion

To validate  $d(S_1, S_2) = 1 - c(S_1, S_2)$  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.

## 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

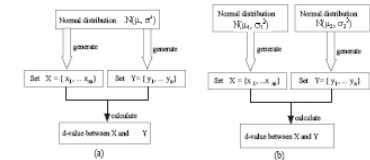


Figure 1. Random generation of d-value from normal distribution

1. We generated 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.
2. We calculated the d-value for each pair of sets.
3. We then estimated the probability density value. 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.

Figure 4 shows the relationship between d-value and p-value while Figure 5 shows how the d-value increases as the mean difference of two sets increases.

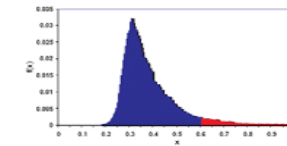


Figure 2. The probability density function of FM d-value

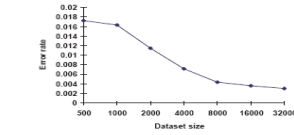


Figure 3. The impact of dataset size on error rate of PDF of FM d-value

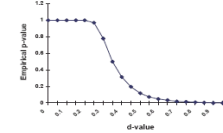


Figure 4. The relationship between FM d-value and empirical p-value

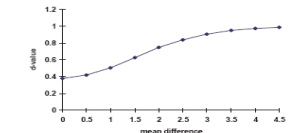


Figure 5. Relationship between the mean difference of distributions and d-value

## Conclusions

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 shown that it is effective and robust. We also applied FM-test to a real diabetes dataset and identified significant genes. While six of them have been confirmed to be associated with insulin signal and/or diabetes in the literature, one has been recommended by others, the remaining three genes are suggested as three potential diabetes genes involved in insulin resistance for further biological investigation.