International Journal of Advanced Computer and Mathematical Sciences
ISSN 2230-9624. Vol 5, Issue2, 2014, pp36-42
http://bipublication.com

Selection of Discriminatory Gene Set for Type II Diabetes Using Fisher Linear Discriminant

Atul Kumar*[1], D. JeyaSundara Sharmila[1], Rajiv Kant[2].

[1] Department of Bioinformatics, Karunya University, Coimbatore, Tamil Nadu, India
[2]Department of Biotechnology, Karunya University, Coimbatore, Tamil Nadu, India
*Corresponding Author: E mail id: atulkumar@karunya.edu, Tel: +919488523540

ABSTRACT

Genes are also known to play a role in the occurrence of infectious diseases like tuberculosis and AIDS as well as some non-communicable diseases like cancer and diabetes. A discriminative gene can act as a target which is a molecular structure that will undergo a specific interaction with drugs because they are administered to treat or diagnose a disease. One of the best and most accurate methods for identifying disease-causing genes is monitoring gene expression values in different samples using microarray technology. The main problem of microarray data is its limited number of samples with respect to number of genes. Fisher's linear discriminant is a classification method that projects high-dimensional data onto a line and performs classification in this one-dimensional space. In this work the Fisher criteria is used for selection for discriminatory gene set.

Keyword: Fisher Criteria, Type II Diabetes, Microarray, Gene Expression, Discriminatory Gene Set, Filter Methods, Wrapper Methods, Embedded Methods

[I] INTRODUCTION

Genes are specific lengths of DNA that determine the order of amino acids used to make protein. Dysfunctional gene behaviour is commonly termed as a mutation. These mutations are responsible for causing illnesses. Moreover, if the gene mutations exist in the egg or sperm cell, children can inherit the defective gene from their parents. Diseases can occur due to a defect in a single gene or in a set of genes. Genes are also known to play a role in the occurrence of infectious diseases like tuberculosis and AIDS as well as some non-communicable diseases like cancer and diabetes. A discriminative gene can act as a target which is a molecular structure that will undergo a specific interaction with drugs because they are administered to treat or diagnose a disease. The interaction has a connection with the clinical effects. Though scientists have now identified many gene variants that increase susceptibility to type II diabetes, the majority have yet to be discovered. The known genes appear to affect insulin production rather than insulin resistance. One of the best and most accurate methods for identifying disease-causing genes are monitoring gene expression values in different samples using microarray technology.
Transcribing a gene's DNA sequence into RNA is known as Gene Expression. A gene's expression level indicates the approximate number of copies of that gene's RNA produced in a cell and it is correlated with the amount of the corresponding proteins made [18].

Simultaneous monitoring of thousands of gene expressions has been made possible due to recent advent of microarray technology. Analysing gene expression data can indicate the genes which are differentially expressed in the diseased tissues [22]. The main problem of microarray data is its limited number of samples with respect to number of genes. Many of these genes have no role in creation of the disease of interest; therefore identification of disease-causing genes can determine not only the cause of the disease, but also its pathogenic mechanism [12]. Diagnostic tests and classification of patients can be done by using marker genes, which will reduce laboratory costs and increase the accuracy. Identification and selection of a subset of genes as disease causing genes is called gene selection.

Different methods have been proposed in the literature for gene selection. They can be organized in three categories: filter methods, wrapper methods and embedded methods [16]. Filter methods evaluate the goodness of the genes looking only at the intrinsic characteristics of the data, based on the relation of each single gene with the class label by the calculation of simple statistical criteria [9]. Some of the methods are parametric and some are nonparametric. Parametric methods have strict assumptions on the analysed data, including: normal distribution, homogeneous variances between data groups and continuous measures with equal intervals. Non-parametric methods do not require above assumptions, so they are computationally easier and quicker but statistically less powerful [21].

There is a large variety of parametric methods such as Signal to Noise Ratio (SNR) [7] and Fisher [15]. Filter methods are fast and simple.Fisher's linear discriminant is a classification method that projects high-dimensional data onto a line and performs classification in this one-dimensional space. The projection maximizes the distance between the means of the two classes while minimizing the variance within each class. This defines the Fisher criterion, which is maximized over all linear projections, g:

$$Fisher(g) = \frac{(m_1(g) - m_2(g))^2}{(s_1^2(g) + s_2^2(g))}$$

where $m_1(g)$ and $m_2(g)$ are means of gene g expression across the diabetic and normal samples respectively. $s_1(g)$ and $s_2(g)$ are standard deviations of gene g across the diabetic and normal samples respectively. The Fisher criterion score gives higher values to features whose means differ greatly between the two classes relative to their variances. Those genes with the highest scores are considered as the most discriminatory genes.

[II] MATERIALS AND METHODS

141 samples from different tissues of *Homo sapiens* were collected from GEO database [4] and Diabetes Genome Anatomy Project (DGAP). Out of these, 71 samples are of normal human beings and 70 are of diabetic humans (Table 1). Normalization of data and model based expression was computed using the dchip software. Then Fisher Criteria was applied on the samples for gene classification.

Table 1: Set of data samples taken for studies

GEO Accession	Data	No of Samples		No of Genes	Country
		Normal	Diabetic		
GSE7146	Effect of insulin infusion on human skeletal muscle	6	6	22215	Sweden
GSE12643	Transcription profiling of myotubes from patients with type 2 diabetes	10	10	12558	Denmark
GSE20966	Gene expression profiles of beta-cell enriched tissue obtained by Laser Capture Microdissection from subjects with type 2 diabetes	10	10	61308	USA
GSE23343	Expression data from human liver with or without type 2 diabetes	7	10	54600	Japan
GSE25724	Expression data from type 2 diabetic and non-diabetic isolated human islets	7	6	22184	Italy
DGAP	Human pancreatic islets from normal and Type 2 diabetic subjects (A)	7	5	22191	Caucasian and Asian

DGAP	Human pancreatic islets from normal and Type 2 diabetic subjects (B)	7	5	22550	
DGAP	Human skeletal muscle - type 2 diabetes	17	18	22177	Sweden

[III] RESULTS AND DISCUSSIONS

All the samples collected from GEO database were subjected to fisher criteria. The top 10 genes ranked according to the Fischer Score are shown in the tables in this section. The results obtained are notable and discussed herewith.

In the data GSE7146 (Effect of insulin infusion on human skeletal muscle) [14] gene G0S2 (G0/G1switch 2) showed the highest value of Fischer score i.e. 10.1864072 (Table 2). The gene G0S2 is a target for PPAR [20] and regulates human adipocyte lipolysis by affecting activity and localization of adipose triglyceride lipase [17]

Table 2: Fischer Scores for Top 10 genes of data GSE7146

Gene	Mean Normal	Standard Deviation Normal	Mean Diseased	Standard Deviation Diseased	Fisher Score
G0S2	872.96	399.7819673	2394.8	259.8766914	10.1864072
SLC22A6	3.88	1.39250015	8.87	1.128964127	7.74831057
CDC6	6.07	1.343785697	13.45	2.396250543	7.21595278
SCNN1G	2.11	1.161540644	5.3	0.59462313	5.97626198
LOC441601 /// LOC652471	2.48	1.42021712	8.69	2.134126675	5.86837431
DNAJC1	47.23	4.740292888	35.49	2.984915409	4.39219807
HSPA14	69.42	1.953929374	53.66	7.368942258	4.27359313
KIAA0692	17.51	3.360115573	8.76	2.58512411	4.25980434
TLE1	124.54	11.03881108	92.64	10.94223743	4.2121727
UBXD8	8.44	1.081864132	12.49	1.677092723	4.11805467

Gene LPAR1 (lysophosphatidic acid receptor 1) in the data GSE12643 (Transcription profiling of myotubes from patients with type 2 diabetes) [6] showed a Fischer value of 2.045876 (Table 3) which is highest among the all the genes in that data sample. Gene LPAR1 is a G protein-coupled receptor that binds the lipid signalling molecule lysophosphatidic acid [1] and regulates blood glucose by stimulating myotube and adipocyte glucose uptake [19].

Table 3: Fischer Scores for Top 10 genes of data GSE12643

Gene	Mean Normal	Standard Deviation Normal	Mean Diseased	Standard Deviation Diseased	Fisher Score
LPAR1	256.57	76.02597	451.53	113.1309	2.045876
PNRC1	145.62	34.22019	207.98	27.46885	2.019553
CREBL2	105.55	13.56716	147.91	29.93188	1.661476
PCNA	406.1	56.32556	520.08	70.27516	1.601673
RABGAP1L	273.1	60.26868	186.13	33.39923	1.593105
CHN2	26.14	7.309533	37.63	5.783126	1.519677
C10orf10	392.74	146.0305	735.46	239.1661	1.495784
GUCY2F	16.5	2.707892	11.32	3.481438	1.379339
MAPK8	43.09	7.965332	56.57	8.630498	1.317391
PPL	412.92	242.2957	864.25	310.8722	1.311235

Among all the genes of the data GSE20966 (Gene expression profiles of beta-cell enriched tissue obtained by Laser Capture Microdissection from subjects with type II diabetes) [10] gene FXYD3 (FXYD domain containing ion transport regulator 3) showed the highest Fischer score of 4.662484 (Table 4). FXYD3 acts as a pancreatic beta cell-specific biomarker [5].

Table 4: Fischer Scores for Top 10 genes of data GSE20966

Gene	Mean Normal	Standard Deviation Normal	Mean Diseased	Standard Deviation Diseased	Fisher Score
FXYD3	123.25	26.25503	222.22	37.80294	4.623825
ZNF688	127.02	21.39657	185.24	17.95785	4.343943
MDFIC	68.99	25.6878	200.23	60.25393	4.01453
MDFIC	309.58	104.6787	643.52	131.7585	3.937998
Hs.148558.0	25.81	5.019425	13.92	3.66561	3.65952
BIVM	31.44	3.974955	42.79	4.685031	3.412537
LAMC1	60.09	7.412706	84.83	11.78551	3.157486
PSMD7	19.42	2.495623	13.18	2.48051	3.144932
Hs.24340.0	39.74	9.776042	18.54	7.342271	3.006691
Hs2.310261.1	84.47	8.956178	65.85	5.953416	2.997714

In the data GSE23343 (Expression data from human liver with or without type 2 diabetes) [11] the gene KIAA0090 showed the highest value of score of Fischer score which is 4.662484 (Table 5). It is a Human gene coding for a protein of unknown function. As this gene is from human liver it can be targeted as a potential drug target for type II Diabetes.

Table 5: Fischer Scores for Top 10 genes of data GSE23343

Gene	Mean Normal	Standard Deviation Normal	Mean Diseased	Standard Deviation Diseased	Fisher Score
PRDX4	2132.03	158.8653	578.88	155.8583	48.70332
KIAA0146	53.11	2.659303	32.52	1.971142	38.69096
SERPINI1	233.57	30.28421	29.04	13.82267	37.74817
PNPLA4	105.49	8.314443	44.34	5.567555	37.34556
USO1	1003.95	59.61715	226.23	115.4329	35.83446
PDPN	8.7	0.440379	2.16	1.018442	34.74094
ARFGAP3	729.95	70.52113	241.17	45.47178	33.9311
IQCA1	21.91	1.951826	35.66	1.386711	32.98028
NFATC4	47.14	5.351033	85.53	4.072358	32.59329
SSX3	19.5	3.435176	47.18	3.443148	32.38891

The gene PRDX4 (Peroxiredoxin 4) in the data GSE25724 (Expression data from type 2 diabetic and non-diabetic isolated human islets) [3] have the highest Fischer score among all the samples taken for this work. The Fischer score for PRDX4 is 48.70332 (Table 6). PRDX4 protects against nonalcoholicsteatohepatitisand Type II Diabetes in a Nongenetic Mouse Model [13]. Overexpression of peroxiredoxin 4 protects against high-dose streptozotocin-induced diabetes by suppressing oxidative stress and cytokines in transgenic mice [2].

Table 6: Fischer Scores for Top 10 genes of data GSE25724

Gene	Mean Normal	Standard Deviation Normal	Mean Diseased	Standard Deviation Diseased	Fisher Score
KIAA0090	60.85	6.654775	96.77	15.24609	4.662484
PLIN5	104.64	13.50832	46.02	23.76857	4.597556
CLCN3	47.59	9.695399	77.33	10.23498	4.450028
ARL8A	130.72	40.57515	257.43	48.20727	4.043899
MUC4	12.7	2.267814	6.03	2.669137	3.626633
LYG1	39.84	5.180169	26.04	5.068539	3.625755
FEM1B	363.14	49.40298	496.81	49.95285	3.619911
ANAPC10	70.71	16.78853	115.25	16.6563	3.547036
WNK1	360	57.28857	560.42	90.60813	3.495373
TP73	18.88	2.139326	10.1	4.187096	3.486824

Two different data (A and B) of Causian and Asian population from Human pancreatic islets from normal and Type 2 diabetic subjects [8] showed gene DECR2 (sample A) and gene THRAP6 (sample B) as highest scored genes with the Fisher score as 6.528318 and 8.040487 respectively (Table 7 & 8). Gene DECR2 (2,4-dienoyl CoAreductase 2) is an Auxiliary enzyme of beta-oxidation and target for PPARα, whereas gene THRAP6 is Thyroid hormone receptor associated protein 6.

Table 7: Fischer Scores for Top 10 genes of data from DGAP (Sample A)

Gene	Mean Normal	Standard Deviation Normal	Mean Diseased	Standard Deviation Diseased	Fisher Score
DECR2	92.26	14.10227	151.98	18.63961	6.528318
BUB3	68.25	9.129948	107.04	12.20484	6.47685
RANBP9	103.88	14.23112	52.28	14.48722	6.456192
ZC3H14	60.2	10.39184	98.96	11.79194	6.081351
TMEM111	5.18	1.88501	11.55	1.880005	5.724989
FAM82C	125.77	13.7613	164.92	8.88765	5.711365
CYP7A1	5.09	1.259913	1.66	0.74611	5.487204
GPR132	7.48	2.453086	1.55	0.689435	5.41586
Hs.247983.0	21.57	7.094177	39.17	3.500317	4.949858
KIAA0507	7.34	2.057121	12.81	1.411701	4.806842

Table 8: Fischer Scores for Top 10 genes of data from DGAP (Sample B)

Gene	Mean Normal	Standard Deviation Normal	Mean Diseased	Standard Deviation Diseased	Fisher Score
THRAP6	46.06	7.683657	21.07	4.316347	8.040487
LPHN3	1.98	1.064765	4.97	0.607149	5.950729
VPS36	33.65	5.716146	51.61	5.072118	5.523248
CAPS	58.05	15.44427	101.54	10.24684	5.505828
Transcribed locus	6.25	1.868827	12.26	1.772772	5.443681
Transcribed locus, moderately similar to XP_517655.1 PREDICTED	17.6	5.439143	33.2	4.19659	5.156407
MAP2K5	9.26	3.692054	19.73	2.881298	4.997957
ZNF559	5.03	2.658344	14.65	3.403501	4.962003
ZNF638	7.3	1.123299	10.16	0.625564	4.947947
ZNF605	20.99	5.484764	36.12	4.200536	4.796368

The last data which was from skeletal muscle of Swedish male having Type II Diabetes (Vamsi et al. 2003) have gene ZAK with the highest Fischer score of 1.021514 (Table 9). ZAK is sterile alpha motif and leucine zipper containing kinase AZK.

Table 9: Fischer Scores for Top 10 genes of data from DGAP (Swedish Male)

Gene	Mean Normal	Standard Deviation Normal	Mean Diseased	Standard Deviation Diseased	Fisher Score
ZAK	39.44	8.512428	25.97	10.25466	1.021514
ANKHD1 /// MASK-BP3	40.28	5.993019	30.87	7.38196	0.97941
ProSAPiP1	12.13	4.774689	6.61	2.959105	0.96566
DAZ1 /// DAZ3 /// DAZ2 /// DAZ4	5.63	2.141676	3	1.664714	0.940046
PCDHB3	32.62	6.813426	24.72	4.703889	0.910439
ZNF688	51.71	14.86344	35.47	10.3075	0.806127
EIF2C4	40.05	9.556071	29.9	6.143436	0.798251
CADPS2	29.88	5.574188	24.08	3.349427	0.795455
COX7A1	4549.36	688.5083	3751.08	573.7689	0.793336
ZNF267	13.98	4.824634	8.9	3.046019	0.792694

[IV] CONCLUSION

Among all the 8 data taken for this study from different tissue samples of human, the genes in the data GSE25724 have the maximum values for Fischer score (Table 6). The genes from this data samples may be discriminatory for Type II diabetes and may act as a potential drug target

and few genes can be novel genes for Type II diabetes. Genes from all other data samples have a moderate to low Fischer score but they may also be taken as drug target by analysing it carefully. Though Filter methods (Fischer) are fast and simple but they do not consider the correlation of genes and lead to redundancy in the selected gene sets. A further refinement in classification of genes can be done by introducing a redundancy reduction stage and the subjecting the final gene set obtained to SVMRFE approach for final selection of gene set [12].

[V] ACKNOWLEDGEMENT

The authors are very grateful to Mr. Sachidanand Singh, Assistant Professor, Department of Bioinformatics, Karunya University for his help in providing the lab space and permitting to use his server. His continuous guidance and encouragement has led to successfully completion of work.

[VI] REFERENCES

1. Choi JW, Herr DR, Noguchi K, Yung YC, Lee C-W, Mutoh T, Lin M-E, Teo ST, Park KE, Mosley AN, Chun J (2010) LPA Receptors: Subtypes and Biological Actions. Annual Review of Pharmacology and Toxicology 50 (1): 157 – 186.
2. Ding Y, Yamada S, Wang KY, Shimajiri S, Guo X, Tanimoto A, Murata Y, Kitajima S, Watanabe T, Izumi H, Kohno K, Sasaguri Y (2010) Overexpression of peroxiredoxin 4 protects against high-dose streptozotocin-induced diabetes by suppressing oxidative stress and cytokines in transgenic mice. Antioxidant and Redox signalling 13(10): 1477 – 1490
3. Dominguez V, Raimondi C, Somanath S, Bugliani M, Loder MK, Edling CE, Divecha N, da Silva-Xavier G, Marselli L, Persaud SJ, Turner MD, Rutter GA, Marchetti P, Falasca M, Maffucci T (2011) Class II phosphoinositide 3-kinase regulates exocytosis of insulin granules in pancreatic beta cells. The Journal of Biological Chemistry 286 (6): 4216 – 4225
4. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Research 30 (1): 207 - 210
5. Flamez D, Roland I, Berton A, Kutlu B, Dufrane D, Beckers MC, De Waele E, Rooman I, Bouwens L, Clark A, Lonneux M, Jamar JF, Goldman S, Maréchal D, Goodman N, Gianello P, Van Huffel C, Salmon I, Eizirik DL (2010) A genomic-based approach identifies FXYD domain containing ion transport regulator 2 (FXYD2) γa as a pancreatic beta cell-specific biomarker. Diabetologia 53(7): 1372 – 1383
6. Frederiksen CM, Højlund K, Hansen L, Oakeley EJ, Hemmings B, Abdallah BM, Brusgaard K, Beck-Nielsen H, Gaster M (2008) Transcriptional profiling of myotubes from patients with type 2 diabetes: no evidence for a primary defect in oxidative phosphorylation genes. Diabetologia 51(11): 2068 - 2077
7. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 286: 531 - 537
8. Gunton JE, Kulkarni RN, Yim S, Okada T, Hawthorne WJ, Tseng YH, Roberson RS, Ricordi C, O'Connell PJ, Gonzalez FJ, Kahn CR (2005) Loss of ARNT/HIF1beta mediates altered gene expression and pancreatic-islet dysfunction in human type 2 diabetes. Cell 122 (3): 337 - 349
9. Inza I, Larranaga P, Blanco R, Cerrolaza AJ (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artificial Intelligence in Medicine 31(1): 91 - 103
10. Marselli L, Thorne J, Dahiya S, Sgroi DC, Sharma A, Bonner-Weir S, Marchetti P, Weir GC (2010) Gene expression profiles of Beta-cell enriched tissue obtained by laser capture microdissection from subjects with type 2 diabetes. PLOS Med 5 (7): 1 – 13
11. Misu H, Takamura T, Takayama H, Hayashi H, Matsuzawa-Nagata N, Kurita S, Ishikura K, Ando H, Takeshita Y, Ota T, Sakurai M, Yamashita T, Mizukoshi E, Yamashita T, Honda M, Miyamoto K, Kubota T, Kubota N, Kadowaki T, Kim HJ, Lee IK, Minokoshi Y, Saito Y, Takahashi K, Yamada Y, Takakura N, Kaneko S (2010) A liver-derived secretory

protein, selenoprotein P, causes insulin resistance. Cell Metabolism 12 (5): 483 – 495
12. Mohammadi A, Saraee MH, Salehi M (2011) Identification of disease-causing genes using microarray data mining and Gene Ontology. BMC Medical Genomics 4: 12 – 19
13. Nabeshima A, Yamada S, Guo X, Tanimoto A, Wang KY, Shimajiri S, Kimura S, Tasaki T, Noguchi H, Kitada S, Watanabe T, Fujii J, Kohno K, Sasaguri Y (2013) Peroxiredoxin 4 Protects Against NonalcoholicSteatohepatitis and Type 2 Diabetes in a Nongenetic Mouse Model. Antioxidant and Redox signalling
14. Parikh H, Carlsson E, Chutkow WA, Johansson LE, Storgaard H, Poulsen P, Saxena R, Ladd C, Schulze PC, Mazzini MJ, Jensen CB, Krook A, Björnholm M, Tornqvist H, Zierath JR, Ridderstråle M, Altshuler D, Lee RT, Vaag A, Groop LC, Mootha VK (2007) TXNIP regulates peripheral glucose metabolism in humans. PLOS Med 4 (5): 868 – 879
15. Pavlidis P, Weston J, Cai J, Grundy WN (2001) Gene functional analysis from heterogeneous data. Research in Computational Molecular Biology (RECOMB), New York, ACM Press 249-255
16. Saeys Y, Inza I, Larranaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19): 2507 - 2517
17. Schweiger M, Paar M, Eder C, Brandis J, Moser E, Gorkiewicz G, Grond S, Radner FP, Cerk I, Cornaciu I, Oberer M, Kersten S, Zechner R, Zimmermann R, Lass A (2012) G0/G1 switch gene-2 regulates human adipocyte lipolysis by affecting activity and localization of adipose triglyceride lipase. Journal of lipid research 53 (11): 2307 – 2317
18. Weaver RF (2003) Molecular Biology. Boston McGraw-Hill
19. Yea K, Kim J, Lim S, Park HS, Park KS, Suh PG, Ryu SH (2008) Lysophosphatidic acid regulates blood glucose by stimulating myotube and adipocyte glucose uptake. Journal of Molecular Medicine 86 (2): 211 - 220
20. Zandbergen F, Mandard S, Escher P, Tan NS, Patsouris D, Jatkoe T, Rojas-Caro S, Madore S, Wahli W, Tafuri S, Müller M, Kersten S (2005) The G0/G1 switch gene 2 is a novel PPAR target gene. The Biochemical Journal 392 (2): 313 – 324
21. Zhang A (2006) Advanced Analysis of Gene Expression Microarray Data. Danvers. World Scientific Publishing Co.
22. Zhang Z (2006) The Use of Microarray Data Integration to Improve Cancer Prognosis. University of North Carolina

[VII] CONFLICT OF INTEREST STATEMENT

I, the corresponding author certify that there is no conflict of interest with any financial organization regarding the material discussed in the manuscript.