Identifying Informative Genes for Type 1 Diabetes using Genetic Algorithm-Neural Network (GANN)

Dong Ling Tong

Artificial Intelligence Lab, Faculty of Engineering and Computing, First City University College, Petaling Jaya, Malaysia

ABSTRACT

Diabetes mellitus is a major health burden worldwide. Approximately 400 million cases worldwide on diabetes were reported in 2014 and a projection death toll due to diabetes is doubled up in-between 2005 and 2030. The rate of newly diagnosed diabetes cases continues to grow in conjunction to global aging. Numerous computational models have been developed to investigate diabetes disease. However, many emphasize on the risk factors of the disease, little attention was paid on the development of genetic mutation of the disease. This paper aims to explore molecular relationship of the mutated genes in diabetes using a hybrid computational method. A set of correlated genes were extracted using a hybrid genetic algorithm and neural network model from a high dimensional microarray data and validated using various classification methods. Results demonstrated the effectiveness of the hybrid model in selecting informative genes for the disease.

Keywords: *Classification, Feature selection, Genetic algorithm, Informative genes, Microarray, Neural network, Type 1 diabetes.*

I. INTRODUCTION

Body weight has becoming a primary concern of most people today. Uncontrollable weight growth leads to the retention of sugar level on the blood vessels and subsequently promotes complications in many parts of the body. These complications including pancreatic failure, stroke, kidney failure, heart attack, nerve damage, vision loss and amputation. Diabetes mellitus is a metabolic disorder disease caused by the long-term high blood glucose in blood vessels and is one of the global burden of non-communicable diseases. Approximately 400 million cases worldwide have been reported in 2014 and the projected death toll due to diabetes is doubled up in-between 2005 and 2030 [1].

With the significant advances in computational technology, extensive studies on diabetes have been conducted and better prediction accuracy is reported when using computational model compared to manual prediction. Zheng et al. [2] reported an average area under the curve (AUC) performance of 98% was observed when using computational methods to predict diabetic medical data compared to only 71% in AUC when doing a manual prediction. Nilashi et al. [3] proposed a combined computational method to study risk factors of the Pima Indians diabetes dataset obtained from the UCI repository. In their study, self-organizing map (SOM) was used to cluster the data, followed by noise removal using principal component analysis (PCA) and artificial neural network (ANN) for classification. They reported high classification accuracy of 92.28% using these combined methods. Using the similar dataset, Wu et al. [4] achieved the accuracy of 95.42% using K-means clustering, coupled with logistic regression classification. Meanwhile, Alghamdi et al. [5] used synthetic minority oversampling technique (SMOTE) algorithm to handle data imbalance issue on their diabetic prediction. A high AUC performance of 92% was reported using the ensemble of 3 decision tree models, i.e. random forest (RF), naive Bayes (NB) tree and logistic model tree (LMT). Many of these studies emphasize on classification of clinical data [2-6], in the light to identify key risk factors in diabetic patients, little attention was paid on the genetic mutation in diabetes pathology to understand the development and progression of the disease.

This study aims to explore significant gene sets for Type 1 diabetes using a hybridized computational model called genetic algorithm-neural network (GANN). This study utilizes the universal computational power of ANN to calculate the fitness function of GA while uses the evolutionary capability of GAs to optimize ANN models. For gene validation, 5 commonly applied classifiers, such as multilayer perceptron (MLP), support vector machine(SVM), naive Bayes (NB), random forest (RF) and logistic model tree (LMT) have been employed using the WEKA data mining suite. Through the identified genes, the genetic pathology of diabetes can be revealed and precautions can be taken to delay further proliferation of the disease.

II. DATASET

In this study, the GSE9006 microarray data obtained from GEO (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc= GSE9006) was used. GSE9006 is the peripheral blood mononuclear cell (PBMC) microarray data reported by Kaizer et al. [7]. This dataset contains 117 samples collected from children diabetes patients aged in-between 2 and 18 years. The dataset was categorized into 3 groups: 24 samples in the healthy group, 81 in Type 1 diabetes (T1D) and 12 in Type 2 diabetes (T2D). Out of 81 T1D samples, 43 were newly diagnosed samples and 38 were returning diagnostic samples (19 were returning diagnostic 1 month after the initial diagnostic and another 19 were 4-month after the initial diagnostic).

The samples were hybridized according to Affymetrix protocol, yielding a total of 44760 probe sets stored in Affymetrix U133A and U133B gene chips. Detailed information on the sample preparation can be found in the original study [7].

To maximize the chance of deriving a panel of informative genes that are associated to T1D, only the newly diagnosed T1D samples and healthy samples in the U133A gene chip. In other words, a total of 67 samples (i.e. 43 T1D sample, 24 Healthy samples) comprising 22283 probe sets was analyzed in this study.

III. METHODOLOGY

A common challenge when working with microarray studies is the curse of data dimensionality and sample sparsity. To ensure that the identified genes are "true" markers for diabetes, the data was first split into training and blind testing sets. Table 1 shows the distribution of samples in the training and blind sets. GANN was then used to extract a more focused subset of genes from the training set. Using the blind testing set, these selected genes were validated using WEKA classifiers. Fig. 1 shows the schematic workflow of our analysis.

Table 1. Sample distribution for training and blind sets.

Group	Training set	Blind set
Healthy	18	6
T1D	29	14

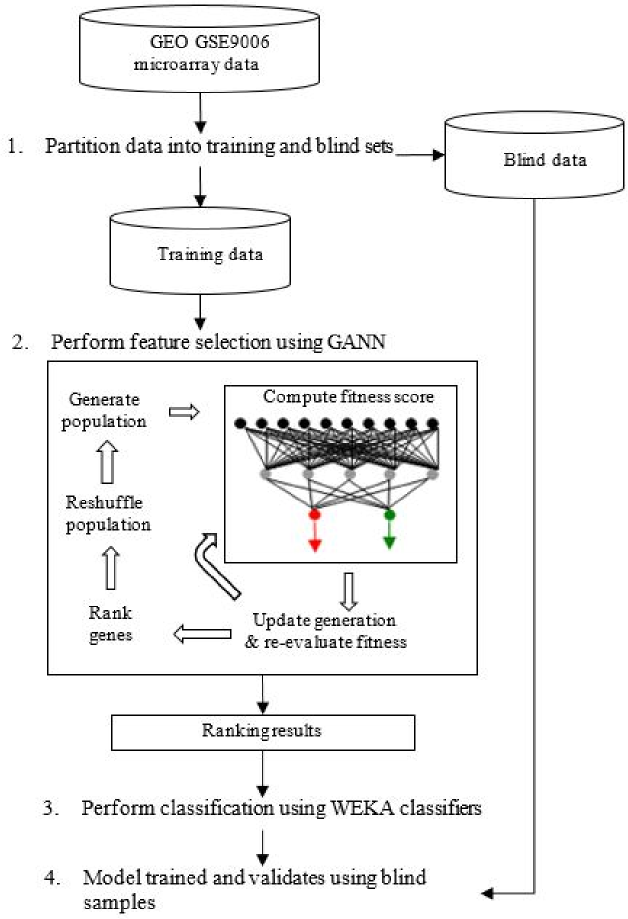
GEO GSE9006
microarray data
1. Partition data into training and blind sets
Blind data
Training data
2. Perform feature selection using GANN
Generate population
Compute fitness score
Reshuffle population
Rank genes
Update generation & re-evaluate fitness
Ranking results
3. Perform classification using WEKA classifiers
4. Model trained and validates using blind samples

Fig. 1 The schematic workflow of this study.

A. Feature selection using genetic algorithm neural-network (GANN)

Genetic algorithm-neural network (GANN) is the hybrid genetic algorithm (GA) and artificial neural network (ANN) program that was developed for microarray analysis [8-9]. It is a form of co-evolution algorithm based on 2 distinct objectives, i.e. to identify high statistical significant feature set for high dimension and imbalance data. GANN utilized the universal computational power of ANN to compute GA fitness score and at the same time, the weights of ANN is also optimized by GA.

During gene selection process, a population of 300 chromosomes (i.e. genes) was randomly selected from a pool of entire chromosomes (i.e. 22283 genes). The fitness for each chromosome is calculated based on the true positives (TPs) of samples computed using feedforward neural networks. The chromosome with least fitness value is replaced with better fit offspring chromosome. The process of adjusting fitness values and network weights iterated 40000 times and a list of ranked genes based on the cumulative TPs is produced. The population is then shuffled at the pool of entire chromosomes and the entire process of fitness and weight computation/adjustment repeats with a new population. The entire gene selection process is terminated when either one of the following termination criteria is satisfied:

(a) the maximum number of population reshuffle, i.e. 5000 times;

(b) the maximum number of evaluations, i.e. 40000 generations; or

(c) the pre-set threshold for TP

Table 2 summarizes the GANN parameters used in this study. GANN has been applied to identify significant feature subset on various biological data, including bioassay [9], mass spectrometry [10], cytometry [11-12], and microarray [13]. Detailed information on the selection performance of GANN can be found in our previous study [14].

B. Classification using WEKA classifiers

WEKA is the open-source data mining suite developed by The University of Waikato, aims to make machine learning techniques available to public to analyze their data [15-16]. Five (5) statistically different classifiers from WEKA version 3.8 was used in this study. These classifiers are NB, MLP, SMO, RF, and LMT.

NB is a probabilistic model that measure how likely each hypothesis to be happened based on the assumption that parameters that driven the occurrence of the hypothesis are not dependent. MLP is a 3-layered backpropagation perceptron-based model that employs a sigmoid activation function. SMO is an implementation of SVMs that compute the upper bound of support vector weights using a sequential minimal optimization algorithm. RF is an ensemble learning model that use many uncorrelated decision trees to derive a common solution to the problem. LMT is a decision tree model that coupled with logistic regression functions in the model.

The default parameter settings in WEKA were used in these classifiers. All classifiers are trained using 10-fold cross validation on the training set of the dataset and validated using the independent blind set of samples.

Table 2. Summary of the GANN parameters.

Parameter	Setting
Initialization	
Population size	300
Chromosome size	10
Chromosome encoding	Real-number representation
Fitness computation	
Fitness function	Total number of correctly labeled samples
Selection type and size	Tournament, size = 2
ANN type and architecture	Feedforward, architecture = 10-5-2
Total nodes	67, including 7 bias nodes
Activation function	Sigmoidal
Fitness evaluation	
Crossover type and rate	Single-point, rate = 0.5
Mutation rate	0.1
Termination criteria	
Evaluation size	40000
Whole cycle repeat	5000

IV. RESULTS AND DISCUSSION

Approximately 5.4 hours (19520 s) was used by the GANN model to rank all 22283 genes for 47 training samples. Amongst the first 20 ranked genes by GANN, 7 were consistent with all the 7 over-expressed genes reported by Kaizer et al. [7]. This indicates that GANN is capable of selecting genes that show biological relevant to the disease of study. These genes are IL1B (205067_at, 39402_at), EGR2 (205249_at), EGR3 (206115_at), PTGS2 (204748_at), CCR1 (205098_at) and FOSB (202768_at). Amongst these overlapped genes, PTGS2, IL1B, and CCR1 are the first 3 highly ranked in GANN with corrected p-value less than 6 x 10^{-6}. This indicates that GANN able to extract statistical significant genes from the high dimensional microarray data. Table 3 shows the first 20 ranked genes selected by GANN based on the training set of the entire dataset.

To show the regulation behavior of these selected genes to the disease, a heat map is generated. Fig. 2 presents the heat map for the selected genes by GANN. A clear gene behavior in the healthy and T1D groups was observed. Amongst these 20 genes selected by GANN, 14 show upregulation behavior on T1D. These genes are PTGS2, IL1B, CCR1, SGK1 (201739_at), EGR2, EGR1 (201694_s_at), FOSB, EGR3, IGH (217022_s_at), TRIB1 (202241_at), DUSP6 (208893_s_at), IGKC (211644_x_at) and HLA-DRB4 (209728_at). Out of these genes, IL1B, EGR2, EGR3, PTGS2, CCR1 and FOSB were also reported in the original study as over-expressed genes on T1D [7].

Table 3. Genes selected by GANN. Overlapping genes with the original study is indicated with *.

Accession ID	Gene Symbol	p-value corrected with Benjamini-Hochberg FDR
204748_at	PTGS2*	6.36 x 10^{-6}
205067_at	IL1B*	1.20 x 10^{-7}
39402_at	IL1B*	2.10 x 10^{-7}
205098_at	CCR1*	1.20 x 10^{-7}
201739_at	SGK1	2.10 x 10^{-7}
203305_at	F13A1	1.26 x 10^{-4}
205249_at	EGR2*	5.66 x 10^{-6}
200665_s_at	SPARC	1.36 x 10^{-4}
214146_s_at	PPBP	3.42 x 10^{-5}
204439_at	IFI44L	5.51 x 10^{-4}
201694_s_at	EGR1	3.62 x 10^{-6}
202768_at	FOSB*	1.20 x 10^{-5}
203680_at	PRKAR2B	6.44 x 10^{-5}
209728_at	HLA-DRB4	2.16 x 10^{-2}
206115_at	EGR3*	2.99 x 10^{-6}
217022_s_at	IGH	2.67 x 10^{-4}
202241_at	TRIB1	1.08 x 10^{-5}
204115_at	GNG11	3.42 x 10^{-5}
208893_s_at	DUSP6	2.39 x 10^{-5}
211644_x_at	IGKC	1.49 x 10^{-2}

Meanwhile, genes F13A1 (203305_at), GNG11 (204115_at), SPARC (200665_s_at), PRKAR2B (203680_at), IFI44L (204439_at), and PPBP (214146_s_at) were inhibited in most of the T1D samples, indicating that they are not diabetes-specific genes. However, these genes may be use as pre-screening genes for early diagnostic of T1D.

To examine capability of the GANN model in extracting statistical significant genes, experiments were run to evaluate the extracted genes and a comparison analysis was carried out based on the use of the entire dataset and the genes selected by GANN. Table 4 presents the summary of the classification results based on the independent test performance.

Table 4. Summary of the classification results (%). Optimal classification results achieved by classifiers in boldface.

Number of genes	NB	MLP	SMO	RF	LMT
22283	70.2	Out of memory	90	70	90
First 20	95	90	100	100	100
First 15	100	80	100	100	100
First 10	100	90	100	**100**	95
First 5	**100**	**95**	**100**	95	**100**

The performance of all classifiers improved when using the selected genes by GANN compared to using the entire genes in the test set. The performance of all classifiers further improved with lesser number of identified genes. Most of the classifiers achieved optimal classification results using only the first 5 ranked genes by GANN. Amongst these classifiers, NB, SMO and LMT achieved 100% classification accuracies with only 5 genes. Meanwhile RF required at least 10 genes to achieve optimal classification. MLP performs the worst among the classifiers in this study, with the optimal classification result of 95% using 5 genes. Overall, all classifiers except MLP achieved more than 90% classification accuracy using the genes selected by GANN. This shows that they are strong predictors for classifiers and different combination of these predictors can optimize performance of classifiers.
The first 5 selected genes by GANN are PTGS2, IL1B, CCR1 and SGK1. Over-expression of these genes in T1D (see Fig. 2) indicates that they are genuine markers for T1D and they are also strong predictors (see Table 4) to discriminate T1D from healthy patients. These genes were suppressed in the healthy group.

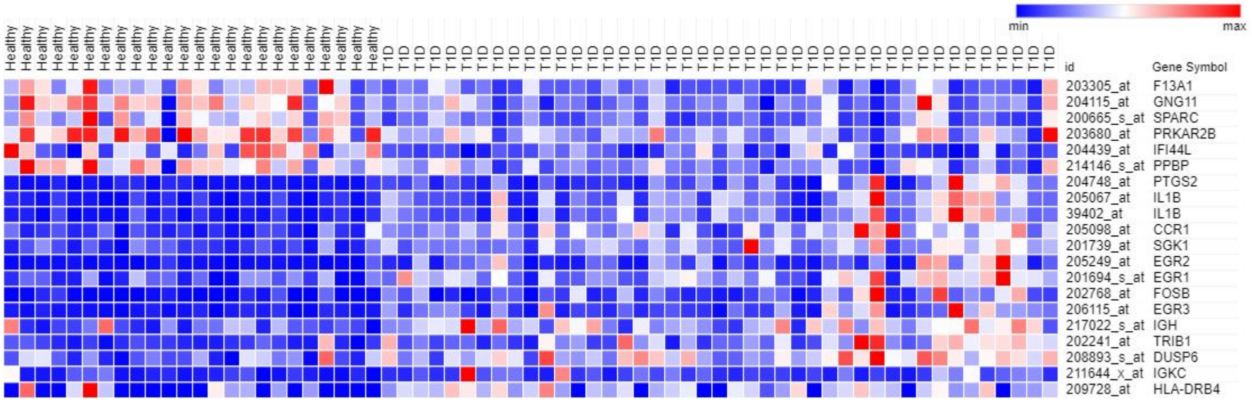

Fig. 2 Heat map for the top-20 genes. Over-expressed genes colored in red. Suppressed genes colored in blue.

V. CONCLUSION

This paper presented a computational framework to study molecular relationship of genes in Type 1 diabetes (T1D). A feature selection framework based on a hybrid GA and ANN (GANN) was used to select a panel of significant genes from the high dimensional PBMC microarray data. These genes were validated using WEKA classifiers.

Significant expression behavior on these genes were observed. The results showed that the GANN model capable to extract genes that were expressed in T1D and also the inhibited genes. This provides some insight on the molecular development of T1D. However, a biological assessment of these genes is highly recommended to confirm their molecular functionality in the development of T1D.

A significant improvement on the performance of all classifiers was observed using the gene panel selected by GANN than using all the genes in the dataset. Better, or equivalent classification results were achieved when the number of the genes reduced down to 5 genes. This confirmed that the selected genes are statistical significant and are strong predictors for T1D classification.

REFERENCES

[1] World Health Organization, "*Global Report On Diabetes*". World Health Organization, 2016.

[2] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "*A machine learning-based framework to identify type 2 diabetes through electronic health records''*, Int. J. Med. Inform 2017, vol. 97, pp. 120-127.

[3] M. Nilashi, O. Ibrahim, M. Dalvi, H. Ahmadi, and L. Shahmoradi, "*Accuracy improvement for diabetes disease classification: A case on a public medical dataset*", Fuzzy Inf. Eng. 2017, vol. 9, pp. 345-357.

[4] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "*Type 2 diabetes mellitus prediction model based on data mining*", Informatics in Medicine Unlocked 2018, vol. 10, pp. 100-107.

[5] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "*Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exerclse Testing (FIT) project*", PLoS One 2017, e0179805.

[6] X-H. Meng, Y-X. Huang, D-P. Rao, Q. Zhang, and Q. Liu, "*Comparison of three data mining models for predicting diabetes or prediabetes by risk factors*", Kaohsiung J. Med. Sci. 2013, vol. 29, pp. 93-99.

[7] E. C. Kaizer, C. L. Glaser, D. Chaussabel, J. Banchereau, V. Pascual, and P. C. White, "*Gene expression in peripheral blood mononuclear cells from children with diabetes*", J. Clin. Endocrinol. Metab. 2007, vol. 9, pp. 3705-3711.

[8] D. L. Tong, "*Hybridising genetic algorithm-neural network (GANN) in marker genes detection*", In: Proceedings of the Eight International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009, pp. 1082-1087.

[9] D. L. Tong, and A. C. Schierz, "*Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data*", Artif. Intell. Med. 2011, vol. 53, pp. 47-56.

[10] D. L. Tong, D. J. Boocock, C. Coveney, J. Saif, S. G. Gomez, S. Querol, R. Rees, and G. R. Ball, "*A simpler method of preprocessing MALDI-TOF MS data for differential biomarker analysis: Stem cell and melanoma cancer studies*", Clin. Proteomics 2011, vol. 8, 14.

[11] D. L. Tong, and G. Ball, "*Pattern recognition of multidimensional PBMC flow cytometry histograms for prostate cancer identification*", In: Proceedings of the International Work-Conference on Bioinformatics and Biomedical Engineering, Granada, 18-20 March 2013, pp. 509-516.

[12] D. L. Tong, G. R. Ball, and G. Pockley, "*gEM/GANN: A multivariate computational strategy for auto-characterizing relationships between cellular and clinical phenotypes and predicting disease progression time using high-dimensional flow cytometry data*", Cytometry A, 2015, vol. 87, pp. 616-623.

[13] D. L. Tong, D. J. Boocock, G. K. R. Dhondalay, C. Lemetre, and G. R. Ball, "*Artificial neural network inference (ANNI): A study on gene-gene interaction for biomarkers in childhood sarcomas*", PLoS One 2014, vol. 9, e102483.

[14] D. L. Tong, and R. Mintram, "*Genetic algorithm-neural network (GANN): A study of neural network activation functions and depth of genetic algorithm search applied to feature selection*", Int, J. Mach. Learn. & Cyber. 2010, vol. 1, pp. 75-87.

[15] E. Frank, M. A. Hall, and I. H. Witten, The WEKA Workbench. Online appendix for "*Data Mining: Practical Machine Learning Tools and Techniques*", Morgan Kaufmann, Fourth Edition, 2016.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "*The Weka data mining software: An update*", SIGKDD Explorations 2009, vol. 11.