Reduction of redundancy in diabetes-related gene categories utilizing conceptual similarity and gene expression information

A. Yasmine Begum
Associate professor, Department of Electronics and Instrumentation Engineering, Sree Vidyanikethan Engineering College, A rangampet, Tirupati, Andhra Pradesh

Anil Sharma
Assistant Professor, College of Computing Sciences & IT, Teerthanker Mahaveer University, Moradabad, Uttar Pradesh

M. Praveen Kumar
Assistant Professor (Senior), Control & Automation School of Electrical Engineering VIT University, Katpadi, Vellore, Tamilnadu

Adiseshaiah Sade
Assistant professor, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh

Nalini Kanta Sahoo
Principal, SRM Modinagar College of Pharmacy, SRM Institute of Science and Technology (Deemed to be University), DELHI-NCR Campus, Modinagar, Ghaziabad, Uttar Pradesh

C M Velu
Professor, Department of Computer Science Engineering, Saveetha school of Engineering, Saveetha University, SIMATS, Chennai, Tamilnadu

Abstract---Insulin resistance or insulin deficiency can cause diabetes, which has a high level of glucose in the blood (American Diabetes Association 2014). It's one of the most pressing issues in modern public health. Around 366 million individuals had diabetes in 2011, according to the "International Diabetes Federation" (IDF). This figure is expected to rise to 552 million by the end of 2030. The average score of the semantic measure and the Pearson correlation were used to create a similarity index in this study. Gene similarity can be determined by this procedure. "One" indicates perfect similarity between two genes, whereas "zero" indicates perfect dissimilarity. Algorithms based on Pearson and semantic measures were used to...
identify genes that were genetically similar. PERL was used to write the algorithm (Practical Extraction and Reporting Language). PERL found gene-to-gene similarity and gradually deleted redundant information. An algorithm known as "greedy" was employed to remove all genes that above the 0.8 threshold value. An experiment using several datasets led to a final threshold value of 0.8. The final dataset has a lot of redundant data because the threshold was raised above 0.8. In contrast, genes of great importance were omitted from the dataset when the threshold was set at 0.8.

**Keywords**---diabetes, gene, pancreatic, semantic, IDF, insulin.

**Introduction**

Using microarray technology, high-throughput research may simultaneously measure thousands of gene expressions. The use of microarray technology to analyse gene expression data will aid in the discovery of new genes. In the sick tissue, these genes are expressed in a distinct manner (Zhang 2006). Researchers are constrained by the high expense of microarray experiments. Therefore, the sample size must be reduced. A microarray experiment is difficult to interpret because of the small sample size. Effective strategies for weeding out irrelevant genes exist now[1]. These procedures are carried out with no regard for the microarray data's inherent redundancy.

Around 366 million individuals had diabetes in 2011, according to the "International Diabetes Federation" (IDF). This figure is expected to rise to 552 million by the end of 2030. Diabetes kills eight out of ten people in impoverished countries, such as India. Diabetes is predicted to become a pandemic in both developing and developed countries. This has demanded the development of a more advanced, rapid, reliable, and durable method for treating diabetes. The genesis of diabetes and the function of the pancreas are inextricably intertwined[2]. The release of insulin by the pancreas aids in the measurement of blood glucose levels in humans. When the pancreas fails, blood sugar levels become out of whack, resulting in diabetes. Type I and type II diabetes are the two most common forms. Type I diabetes is characterised by inadequate insulin synthesis by the pancreas. Type II diabetes is brought on by the body's cells becoming resistant to the effects of insulin. Because the cells are unable to take in glucose, the blood sugar level rises (American Diabetes Association 2011). Diabetes complications such as kidney, retinal, and cardiovascular disease impose a significant financial burden on nations around the world because of the associated high social and medical costs. Many complicated disorders, such as obesity, cancer, neurological diseases, and diabetes, are now being studied using DNA microarrays. Microarray is a technology that can be used to swiftly and effectively study the expression of multiple genes in a single response (Bosch et al. 2000). The DNA microarray is a powerful tool for analysing gene expression variations. Gene expression data can be used to identify genes that are overexpressed in the sick tissue (Zang et al. 2014). Gene expression in the pancreas, subcutaneous fat, muscle, liver, adipocytes and preadipocytes, visceral fat, and certain hypothalamic nuclei are monitored in normal and pathological
states to obtain data for gene expression profiling in type II diabetes and obesity. A microarray dataset has a higher number of genes, but a smaller sample size makes it difficult to analyse. As a result, most of the genes that were considered for the study were unnecessary, and as such, redundant. An increase in computational complexity and time consumption, especially when the goal is to discover the most discriminatory genes that could be implicated for causing the disease. Disease-causing genes can be identified, which can lead to a knowledge of the disease’s pathogenic process (Mohammadi et al. 2011). Gene ontology could be useful in calculating gene similarity and redundant genes. A common ancestor of the genes can be discovered by calculating semantic similarity using gene ontology information. This could lead to a better understanding of its functions[3]. At this point in time, the gene ontology consortium hasn't really begun to take shape. Because of this, the semantic similarity score alone cannot be used to find similarity and eliminate redundancy. Two genes can be linked using the Pearson correlation coefficient, which is a statistical method. Pearson's correlation coefficient will be used to evaluate how similar two genes are. Compared to the number of genes, the microarray sample size is too tiny. Gene set redundancy occurs because Pearson correlation is insufficient to identify similarity.

**Literature Survey**

Diabetes is a long-term condition that results from the body's failure to create enough insulin or the body's inability to utilise insulin properly (American Diabetes Association 2013). As a metabolic condition, it has been categorised as such. The term "metabolism" is used to define all chemical events that take place in the cells and the organism to keep it alive. Insulin is produced and released into the bloodstream by the pancreatic islet, which contains clusters of beta cells[4]. Glucose levels rise when the beta cells are unable to produce enough insulin, or when the body is otherwise unresponsive to insulin. Prediabetes or diabetes can result as a result of this disorder. In Greek, the word "diabetes" means "a syphon." Aretaeus, a Greek physician in the second century A.D., termed the condition diabainein for patients who were passing too much water (polyuria) - like a syphon (Online Etymology Dictionary). 371 million individuals were diagnosed with diabetes in 2012, according to the International Diabetes Federation (IDF). 4.8 million people died as a result of the disease[5]. There would be an increase in diabetes cases in developing countries from 84 million to 228 million, according to the report (Haslam and James 2005).

In children and adolescents, insulin-dependent diabetes mellitus, or Type I diabetes, is more common. The pancreas' defence system assaults insulin-producing cells, resulting in the body's inability to produce the insulin it needs to function (Figure.1. and Figure.2.). There is still no established cause of type I diabetes, however environmental factors, prenatal events and eating habits, and virus infections could all play a role in the development of the disease (Grant et al. 2009).
In the overall diabetic population, type I diabetes accounts for approximately 5% of cases (Centers for Disease Control and Prevention 2014). In the United States, the prevalence of type I diabetes among those under the age of 20 increased by 23% between 2001 and 2009, according to data released by the Centers for Disease Control and Prevention (CDC).

A person’s risk of developing type I diabetes is mostly influenced by their genetics. Proteins are essential for the proper functioning of the body’s cells, and they are encoded in genes. Type 1 diabetes risk has been linked to around 18 locations in the genome[6]. Insulin-Dependent Diabetes mellitus 1 (IDDM1) is one of the several areas that may include many genes. If a gene variant affects more than 1% of the population, it is referred to as "gene variant." Type I diabetes risk is connected to a variation of the HLA gene that releases antigens into white blood cells (Todd et al. 1987).

T cells go for and kill beta cells in the body. The devastation begins long before any diabetes symptoms develop. After a diagnosis, the problem persists
Type I diabetes is often difficult to detect until the beta cells have been damaged to a great extent. When this happens, the person needs to take insulin every day in order to live. Insulin is viewed as an antigen or foreign substance by the immune system of those at risk for developing type I diabetes (Harrison et al. 2004). Environmental variables are implicated in the development of type I diabetes, according to numerous research[7]. Only 13-33 percent of monozygotic twins studied had type I diabetes in both of their bodies, according to the results of the study. Type I diabetes was found in one or both siblings in the remaining 87%-67% of cases. For example, this shows a post-conceptional genetic discrepancy or a differing level of exposure to the hypothesised environmental factors (Knip et al. 2005). Some ethnic groups, such as the Caucasians, have seen a significant increase in the incidence of type I diabetes in children during the past few decades[8]. Migration studies in epidemiological studies of type I diabetes demonstrate that the prevalence of type 1 diabetes increases among populations that have moved from a low-incidence area to a high-incidence area, underscoring the importance of environmental factors (Shapira et al. 2010).

A growing number of adolescents and youngsters are developing type II diabetes, which affects adults disproportionately. Glucose levels in the blood rise because the body produces insulin, but it is either not enough or the body doesn't respond to its effects, resulting in a shortage (Grant et al. 2009) (Figure.3.). When it comes to developing type II diabetes, it's a long and gradual process. People who are overweight or obese are more likely to acquire type II diabetes because their bodies have a hard time utilising insulin appropriately[9]. Type II diabetes is strongly associated with a person's genetic make-up. Everywhere in the world, the number of persons with type II diabetes is on the rise. Western Pacific and Southeast Asian countries such as India and China are leading the current diabetes epidemic, according to a research from the World Health Organization (WHO). Type II diabetes is on the rise among youngsters in these countries, with worrisome frequency. This could have serious consequences down the road.

As part of the Diabetes Genome Anatomy Project (DGAP), researchers are trying to uncover the link between insulin action, insulin resistance, and type II diabetes genetics. The "National Institute of Diabetes and Digestive and Kidney Diseases" (NIDDK) and the "Diabetes Research Working Group" (DRWG) collaborated on this
For this initiative, researchers from five different institutions have devoted their time and resources to it. Six projects and four cores form a highly interconnected matrix. It's also a great place to connect with other initiatives and grants, as well as potential future projects. Type II diabetes susceptibility and insulin action are the two main goals of the research, which aims to identify the genes and gene products responsible for this. Furthermore, the DGAP tries to identify secondary gene expression changes as a result of diabetes' metabolic problems.

**Proposed Method**

The SOURCE server (http://smd.princeton.edu/cgi-bin/source/sourceBatchSearch) managed by Princeton University provided the gene ontology information for all the genes in the dataset. The SOURCE server received a text file with the GenBank accession number for each gene as an input. The drop-down selection for "Choose organism option" had "Homo sapiens" selected. When selecting a field for extraction, GO Annotations (full) was selected. The GO ids were then obtained by uploading the file to the server (Figure 4.).

![Figure 4. Proposed Methodology](image)

![Figure 5. SOURCE server input page](image)
Genes with the identical GO ids were eliminated using the AbleBits add-in for Microsoft Excel's duplicate removal tool. The duplicate GO ids were marked with the word "duplicate" in a status column generated by this add-in[14]. The functional similarity matrix (FunSimMat) tool's semantic similarity query option was used to compare two sets of genes' semantic similarity[15]. GO ids were entered as an input to FunSimMat, the simulation tool (Figure.6.). When the tool compared the two lists of GO ids, it computed the semantic score for each pair.

![FunSimMat Tool](image)

**Figure.6. FunSimMat Tool**

**Result and Discussion**

All the genes in the datasets were included in the SOURCE server's result file, along with their accompanying GO identifiers. Molecular function (MF), Biological process (BP), and Cellular component (CC) hierarchies were used to categorise these genes into their functional features. A total of 16167 (72.92%) of the dataset's 22172 genes were found to be involved in molecular function, 857 (3.87%) in biological function, and 633 (2.85%) in cellular component, according to the dataset "GSE7146." Among the 22172 genes, there were 4515 (20.36 percent) genes for which the gene ontology database had no information. Since the HG-U133 B gene chip array was selected for this study as opposed to the HG-U133 A for others, all datasets used in this study have shown relatively similar dispersion of genes in different hierarchies except for "Human pancreatic islets from normal and type II diabetes individuals (B)" (Figure.7 and Figure.8.). It was shown that 8664 (38.42 percent), 834 (3.7 percent), and 1016 (4.51 percent) of the 22550 genes studied were involved in the molecular function, whereas 12036 (53.37 percent) genes had no information in the gene ontology database.
The results of the preceding study were used to categorise genes into molecular functions. Using FunSimMat, researchers were able to estimate the semantic similarity between these genes. (Schlicker and Albrecht 2008) A gene's protein function is represented by its molecular function. In biology, a function or cellular component represents a sequence of events or a molecular function with a known location. Subcellular structures and macromolecular complexes are where they are most active. Semantic similarities were evaluated without taking into account genes' cellular components or their biological functions. Resnik, Jiang, and Conrath, Lin, and a mixture of Resnik and Lin (SimRel) technique were used to determine the semantic similarity (Couto et al. 2007). FunSimMat was used to generate the semantic value for all possible gene pairings in different datasets. Two genes were compared using the Pearson correlation coefficient to see if their expression patterns were similar. It was necessary to use the same gene combinations for expression similarity in order to calculate the similarity index. The FunSimMat server's guidelines for semantic similarity were designed to guide this approach. The following algorithm produced the same set of gene combinations as the semantic. As an input, the method uses semantic similarity
and a non-redundant GO keywords file. There are separate arrays for each row of the two files containing gene expression values. Individual arrays are created for each index in both arrays. As a result of this, the gene and its expression value were printed in a separate file for each array that had the same GO ID. Combining expression similarity and SimRel values yielded an average for all the distinct gene set combinations. Genes with a score greater than or equal to 0.8 were ruled out using this score (Mohammadi et al. 2011). Genes that potentially lead to type II diabetes were separated using a greedy algorithmic technique. Using the gene file’s similarity index score, the algorithm selects the best match (average score of Pearson and semantic scores). Arrays are created for each row of a file that contains a gene and its expression value. The array is broken down into its component elements, one for each index. A unique array is created for each individual gene. When the sum of the scores was more than the threshold of 0.8, the gene was removed from consideration for inclusion in the final subgroup. This file included all genes that had a similarity index score of less than 0.8 and were distinct to each other.

Conclusion

Microarray data is plagued by a large number of genes that are very redundant. Finding valuable information in the data is difficult because of these genes. The genes in the microarray data have similar molecular activities, either because they share a GO id or because they share a common ancestor. First, the redundant GO ids were removed, and then the average Pearson and semantic similarity scores were used to remove any remaining redundancy. The greedy algorithm strategy was used to complete the task. This was the final redundancy reduction stage, and it was found that about 50% of the top 10 genes on the basis of Fisher score were repeated. There are several different machine learning algorithms that can be used to assess unique genes, such as the support vector machine (SVM). Type II diabetes can be diagnosed using any one of these methods to identify a specific gene.

References