Identifying Variant Positions of TCF7L2 Gene Related to T2DM

K. Vijayalakshmi¹, Prof. M. Padmavathamma²

Research Scholar, Dept. of Computer Science, S.V. University, Tirupati, Andhra Pradesh, India¹
Professor & Head, Dept. of Computer Science, S.V. University, Tirupati, Andhra Pradesh, India²

ABSTRACT: Type2 Diabetes Mellitus (T2DM) is associated with impaired insulin secretion. It is the most leading public health problem around the globe, particularly the developing nations in South Asian countries which include India. The pathogenesis of T2DM primarily initiates with inadequacy of pancreatic islet β-cells. Worldwide there are number of Gene Variants that have been associated to T2D. But these genetic variants are differing among the ethnic groups due to different environmental factors and genetic background. By the consequent research by Geneticist, many genes are identified by region wise. Several recent studies by GWAS in Asian Indian populations TCF7L2 is the first largest gene identified. In this research work we collected tagging SNP protein sequence of TCF7L2 gene (i.e. rs7903146, rs12255372, rs7901695). In this paper we propose DR-VL and GA-VL approaches to identify different transcript variant positions due to single mutation. By identifying these phenotype changes we can predict the disease at earlier and take precautions to reduce the high risk.

KEYWORDS: T2DM, TCF7L2, rs7903146, rs12255372, rs7901695

I. INTRODUCTION

T2DM is an endocrinical metabolic disorder that has become more widespread in these modern years, with significant undesirable effects on human health. T2DM is characterized by insulin resistance (IR) and deficient β-cell function. Usually, T2DM is associated with older age, obesity, family history, previous history of GDM, physical inactivity etc. Almost every organ of the body is affected with long-term complications like blindness, heart and blood vessel disease, stroke, kidney failures, amputations and nerve damage. Simultaneously, Genes play a vital role in the development of Diabetes Mellitus (DM)[5,6]. Insulin secretion is regulated by different hormonal stimuli that activate respective molecular pathway within pancreatic β cells. The most two important physiological stimuli are glucose & incretin. The SNPs of each gene effects on incretin sensitivity or incretin secretion, proinsulin conversions etc. With continuous research on Pathomechanisms of T2DM genes, they identified 27 confirmed potential susceptibility genes for T2DM [2]. In that TCF7L2 is the large, strongly associated gene with 3 different common variants .We are concentrating to identify various transcript position from tagging SNPs of protein sequences of the above with computational approach.

II. RELATED WORK

“Pathomechanisms of Type 2 DiabetesGenes” by Harald Staiger, Fausto Machino et.al Type2 Diabetes Mellitus is a complex metabolic disease that is caused by insulin resistance and β-cell dys-function .And also this disease has an evident genetic component and represents a polygenic disease. It was crucially influenced by the development of affordable high density SNPs. The related genes are identified by two approaches i.e. candidate gene approach and hypothesis free GWA scans and finally different gene variants affecting insulin secretions are identified. “The Genetics of Type2 Diabetes Mellitus: A Review” by Sunitha Singh identified different variants that have been associated to T2DM. By conducting survey on various populations based on genetic factors, various genes are identified that are related to T2DM. Across all the GWAs completed, TCF7L2 clearly shows the larger effect in the development of T2DM. KCNQ1 has been shown to be second largest effect next to TCF7L2. G.R.Chandak, C.S. Janipali, S.Bhaskar et.al in their paper “Common variants in the TCF7L2 gene are strongly associated with type 2 diabetes mellitus in the Indian population” mentioned that Indian subjects replicates the strong
association of TCF7L2 variants with T2DM. It also provides evidence for pathogenesis of T2DM by influencing both insulin secretion and insulin resistance. They mention the most common variants i.e. rs7903146, rs12255372, rs7901695 related to this gene are mostly affecting T2DM. “Gene Expression Analysis for Type-2 Diabetes Mellitus- A study on Diabetes with and without parental history” by Chandrasekhar Vasamsetty.et.al diabetes is a group of metabolic disease primarily due to life study factors and genetics. They focus on identifying and classifying genes that cause T2DM with 2 different samples in 2 stages i.e. by Multivariate Outlier Detection and Univariate Outlier Detection. ShitalC.Shah, Andrew Kusiak in their paper “Data mining and genetic algorithm based gene/SNP selection “mentioned that genomic studies provide large volumes of data with no. of SNPs ranging into thousands. The analysis of SNPs permits to determine the most significant features that reflect the best interactions between genotypic & phenotypic data. They proposed 2 approaches for identification of significant genes that are related to drug therapy. They are Weighted Decision tree based gene selection (WDTGS) and Genetic algorithm based gene selection (GAGS).

III. GENETICS OF T2DM

Nowadays, considerable progress was made in the identification of Type2 Diabetes risk genes. And also the recent GWA studies revealed convincing evidence for the contribution of genes to the pathogenesis of T2DM. Several findings indicate that genetics is an important contributing factor for causing this disease. According to the commonly accepted common variant-common disease hypothesis type 2 diabetes is caused by the simultaneous occurrence of common DNA sequence variations in many genes. Each of these DNA alterations is supposed to exert only moderate effects on the affected genes’ function and/or expression, but in their sum, these variations confer an increased susceptibility toward the unfavorable environmental factors also. Single nucleotide polymorphisms (SNPs) exchanges of single base pairs that changes approximately 90% of the sequence variation within the human genome. There are two main approaches that lead the search for type 2 diabetes genes. One is candidate gene approach other is hypothesis-free GWA scan. In Candidate gene approach, the common genetic variants examined and these variants’ allele frequencies are finally analyzed for being changed in type 2 diabetes cases compared with healthy controls. The positional genetic copies of genes located between or near diabetes-linked chromosomal symbols are compared. But this is difficult due to the non-Mendelian mode of inheritance of human type 2 diabetes; and the size of the chromosomal areas identified in this way that often encompass up to hundreds of genes. However the general diabetes-associated variants in TCF7L2 on chromosome 10q25.3 (10-23) were recognized by this manual approach and simulated in several populations and ethnicities, and were confirmed in prospective studies of meta-analyses and also furthermore set by GWA studies also. At present TCF7L2 represents the most realistic diabetes risk gene with an overall allelic relative risk.

IV. TCF7L2 GENE

TCF7L2 variants are the strongest well known genetic risk factors for T2DM. It is a key component and involved in the development of a wide variety of cell lineages and organs. These variants influence T2DM influences in adipogenesis, myogenesis, and pancreatic islet development, as well as in beta-cell survival and insulin secretary granule function. Finally, TCF7L2 polymorphisms have been associated with impaired insulin secretion, glucose production, and glucose tolerance via direct effects on pancreatic islet beta cells [25]. Although genetic tests for TCF7L2 could help to predict the incidence and the rate of onset of T2DM. The strongest predictors continue to be positive family history, increased body mass index, increased blood pressure, and increased serum levels of triglycerides, and liver enzymes, all of which precede inception of metabolic syndrome. The analytical power of TCF7L2 variants disappears with lifestyle modifications and the data suggest that genetic susceptibility to T2DM is determined by TCF7L2 variants might attest an actionable sign for early intervention and disease prevention.
Figure 1: TCF7L2 Gene

DNA Structure and Protein Translation

DNA, or deoxyribonucleic acid, is the heritable material located in the cell nucleus in humans and almost all other organisms. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people [3]. DNA bases pair up with each other, A with T and C with G, to form units called base pairs. Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate are called a nucleotide. Nucleotides are arranged in two long strands that form a spiral called a double helix. The structure of the double helix is somewhat like a ladder, with the base pairs forming the ladder’s rungs and the sugar and phosphate molecules forming the vertical sidepieces of the ladder.

Figure 2: DNA Structure

Transcription is the process through which a DNA sequence is enzymatically copied by an RNA polymerase to produce a complementary RNA. Or, in other words, the transfer of genetic information from DNA into RNA. In the case of protein-encoding DNA, transcription is the beginning of the process that ultimately leads to the translation of the genetic code (via the mRNA intermediate) into a functional peptide or protein. Translation is the process that takes the information passed from DNA as messenger RNA and turns it into a series of amino acids bound together with peptide bonds. It is essentially a translation from one code (nucleotide sequence) to another code (amino acid sequence).

Proposed Work

In this research we identified TCF7L2 gene is the most susceptible gene in Asian Indians [1]. According Harald Staiger, Fausto Machino, Andreas Fritsche et.al words Type 2 Diabetes is a multifactorial disorder that is caused by insulin resistance and β cell dysfunction. And they summarize the current knowledge about risk genes and their variant’s pathomechanism. With their efforts of research, identified 27 confirmed and potential susceptibility genes of Type 2 Diabetes [2]. These are different gene variants that affect insulin secretions and insulin sensitivity of each identified gene. Glucose and Incretin are the two physiological stimuli that regulate insulin secretion that activate respective molecular pathways within the pancreatic β cell. The SNP of each gene effects on incretin sensitivity, incretin secretions, proinsulin conversions etc and effect on β cell functionality. So, mainly we concentrated on TCF7L2 gene, gathered information related to from NCBI public repository.
National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health. It has a responsibility for making available the GenBank DNA sequence and a series of databases relevant to biotechnology and biomedicine. In this Gene Expression Omnibus (GEO) is one of the public functional genomics data repositories supporting MIAME-compliant data submissions. This database stores curated gene expression data sets, as well as original Series and Platform records in the Gene Expression Omnibus (GEO) repository. This repository is used to search terms to locate experiments of our own interest.

To assess the method, we used data set of different size. We collected 50 different transcript variants related to rs7903146, rs12255372, rs7901695 SNPs [4]. There we are considering the protein sequence of each gene variant which is having different transcript positions related both diabetic and Non-Diabetic patients. To identify these disparity positions, we propose the following Iterative model as shown below

![Iterative Model to identify Transcript Variant Position](image)

Figure3: Iterative Model to identify Transcript Variant Position

In the proposing model the collected tagging SNP protein sequences are evaluated and variant position changes are identified by following two proposing approaches namely DR-VL and GA-VL. In the evaluation process the data set is underdone 10-Fold-Cross Validation to make training and testing data set.

6.1: Decision Rule Variant Locator (DR-VL):

The data set contains relatively large sequence of normal, diabetic patients. We compared normal tagging SNP with diseased CDS one by one and identified 12 different transcript positions from the collected dataset. To identify those positions we use the following Rule:

\[
\begin{align*}
\text{IF} & \quad \text{CDS}_P=1 \& \& \text{CDS}_V=k | \text{CDS}_P=118 \& \& \text{CDS}_V=l | \text{CDS}_P=136 \& \& \text{CDS}_V=h | \text{CDS}_P=161 \& \& \text{CDS}_V=q | \\
& \quad \text{CDS}_P=37 \& \& \text{CDS}_V=q \ldots \\
\text{THEN} & \quad \text{DECISION}=\text{HIGH RISK OF DIABETIC}; \\
\text{ELSE} & \quad \text{DECISION}=\text{NORMAL};
\end{align*}
\]
6.2: Genetic Algorithm Variant Locator (GA-VL):

In this approach we are randomly running data set by dividing into training and testing data using 10-Fold–Crossvalidation. Optimized variant positions are identified by running the following genetic algorithm steps.

Step 1: Randomly generate Test data and training data using 10-fold–cross validation
Step 2: Initialize fittest individual X from CDS as test data
Step 3: Best individual Y is selected using Tournament selection
    Step 3.1. Apply one-bit crossover operation on X, Y
    Step 3.2. Evaluate X[i], Y[i] where i=0 to N
    Step 3.3. If X[i] != Y[i]
        Identify and Store CDS_V and CDS_L
    Else
        Continue;
    Step 3.4 Do mutation on the results to produce new offspring
Step 4: Repeat until reach termination

Where CDS is tagging Codon Sequence
N is the total no. of individual Codon sequences in data set
CDS_V is Codon variable
CDS_L is Location of the variable

By computing the above steps we get the same positions as well as some other optimized locations using GA-VL.

VII. ANALYSIS OF COMPUTATIONAL RESULTS

By performing Meta data analysis of collected information we measure computational accuracy of each model by using the following formula.

\[
\text{Accuracy} = \frac{(TN + TP)}{(TN+TP+FN+FP)} \times 100%
\]

We get more performance accuracy with GA-VL (0.9642) when compared to the DR-VL (0.9264). The total no. of occurrences of each transcript location and changes are tabulated, diagrammatically shown using bar chart as shown below.
VIII. CONCLUSION

We conclude that Genes play an active role in the strengthening of Type2 Diabetes along with rapid changes in life style patterns, environmental changes, increasing physical inactivity. We chosen TCF7L2 gene and its common variant SNPs like rs7903146, rs12255372, rs7901695 to build a model to identify the transcript change in the diabetic patient in both ways using DR-VL and GA-VL. We analysed the testing accuracy of the both developing approaches.

REFERENCES

5. Florez JC. 2008. Newly identified loci highlight _cell dysfunction as a key cause of type 2 diabetes: where are the insulin resistance genes? Diabetologia 51:1100--1110
25. “Gene Expression Analysis For Type 2 Diabetes Mellitus – A Study On Diabetes With And Without Parental History,” by