

CAPITAL UNIVERSITY OF SCIENCE AND
TECHNOLOGY, ISLAMABAD



**Construction Of Biological
Association Network For
Connected Components Analysis
Using Graph Theory**

by

Attiya Kanwal

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

in the

Faculty of Health and Life Sciences

Department of Bioinformatics and Biosciences

2019

**Construction Of Biological Association Network
For Connected Components Analysis Using
Graph Theory**

By

Attiya Kanwal

(PI123001)

Foreign Evaluator 1

University/Institute

Foreign Evaluator 2

University/Institute

Dr. Sahar Fazal

(Thesis Supervisor)

Dr. Sahar Fazal

(Head, Department of Bioinformatics and Biosciences)

Dr. Muhammad Abdul Qadir

(Dean, Faculty of Health and Life Sciences)

**DEPARTMENT OF BIOINFORMATICS AND BIOSCIENCES
CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY
ISLAMABAD**

2019

Copyright © 2019 by Attiya Kanwal

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.

(Attiya Kanwal)

Registration No: PI123001

With love, First and foremost I would like to dedicate this work to the Almighty ALLAH, Who bestowed me with the courage and faith to complete this work. I have to thank my parents for their prayers, love and support throughout my life, my husband for his continuous inspiration and my sons, Mohammad Ashaz Ahmed and Mohammad Asher Ahmed.



**CAPITAL UNIVERSITY OF SCIENCE & TECHNOLOGY
ISLAMABAD**

Expressway, Kahuta Road, Zone-V, Islamabad
Phone: +92-51-111-555-666 Fax: +92-51-4486705
Email: info@cust.edu.pk Website: <https://www.cust.edu.pk>

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in the thesis, entitled “**Construction of Biological Association Network for Connected Components Analysis Using Graph Theory**” was conducted under the supervision of **Dr. Sahar Fazal**. No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the **Department of Bioinformatics & Biosciences, Capital University of Science and Technology** in partial fulfillment of the requirements for the degree of Doctor in Philosophy in the field of **Bioinformatics**. The open defence of the thesis was conducted on **October 09, 2019**.

Student Name : Ms. Attiya Kanwal
(PI-123001)



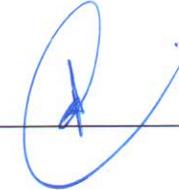
The Examination Committee unanimously agrees to award PhD degree in the mentioned field.

Examination Committee :

(a) External Examiner 1: Dr. Jamil Ahmad
Professor
University of Malakand, Swat



(b) External Examiner 2: Dr. Amir Ali Abbasi
Associate Professor
NCB, QAU, Islamabad



(c) Internal Examiner : Dr. Arshia Amin Butt
Assistant Professor
CUST, Islamabad



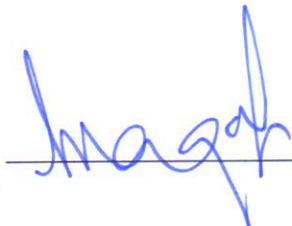
Supervisor Name : Dr. Sahar Fazal
Associate Professor
CUST, Islamabad



Name of HoD : Dr. Sahar Fazal
Associate Professor
CUST, Islamabad



Name of Dean : Dr. Muhammad Abdul Qadir
Professor
CUST, Islamabad



AUTHOR'S DECLARATION

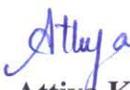
I, **Ms. Attiya Kanwal (Registration No. PI-123001)**, hereby state that my PhD thesis titled, '**Construction of Biological Association Network for Connected Components Analysis Using Graph Theory**' is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/ world.

At any time, if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my PhD Degree.

09-10-2019

Dated:

October, 2019


(Ms. Attiya Kanwal)

Registration No : PI-123001

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled “**Construction of Biological Association Network for Connected Components Analysis Using Graph Theory**” is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

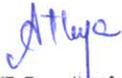
I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/ cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of PhD Degree, the University reserves the right to withdraw/ revoke my PhD degree and that HEC and the University have the right to publish my name on the HEC/ University Website on which names of students are placed who submitted plagiarized thesis.

09-10-2019

Dated:

October, 2019


(Ms. Attiya Kanwal)

Registration No : PI-123001

List of Publications

It is certified that following publication(s) and submissions have been made out of the research work that has been carried out for this thesis:-

1. Attiya Kanwal and Sahar Fazal. “Construction and analysis of protein-protein interaction network correlated with ankylosing spondylitis”, *Gene*, vol.638, 2018, pp. 41-51.
2. Attiya Kanwal, Sahar Fazal, Aamir Iqbal Bhatti, Mukhar Ullah and Muhammad Arslan Khalid. “PubMedInfo Crawler: An innovative extraction process that leads towards biological information mining”, *Meta Gene*, vol.20, 2019, Article. 100550.

Attiya Kanwal

(PI123001)

Acknowledgements

All praises to ALLAH Almighty, the Most Generous and Empathetic and His Holy Prophet Mohammad (Peace be upon Him), the most perfect and glorious among and of ever born on the surface of earth, who is forever a torch of supervision and acquaintance for humankind. They gave me potency and proficiency to complete this goal.

Secondly, I am feeling proud to express some obsessions about my admirable research supervisor Dr. Sahar Fazal, Associate Professor, Head of Department of Department of Biosciences, Capital University of Science and Technology and Co-supervisor Dr. Aamer Iqbal Bhatti, Professor of Department of Electrical Engineering, Capital University of Science and Technology, who continually and convincingly conveyed a spirit of adventure in regard to research. Their vigor, confidence, intelligence and incessant support at every step during the course of this whole project enabled me to attain my goals. I must say that without their support and kind effort, this job would have been impossible.

No words can express my deepest gratitude towards my family who have done and still doing their best to embellish me with the jewelry of education, and for all they did for me, no doubt everything. I owe my deep appreciation for their unmatched support, love and prayers.

I offer my regards and my gratitude to my friends and to all members of CASPR (Controls and signal processing Research group) who hold me up in any respect, who in one way or the other were a source of motivation for me and gave me sincere suggestions during the completion of this dissertation.

Abstract

The concept of big data has been around for years. Big data analysis approaches can play an imperative part in health research. This can be a helpful asset for researchers because it can reveal veiled acquaintance from a massive sum of data. In order to get insights of metabolic pathways at molecular level there is a need to present a unit framework model that serves as the cutting edge technology of system biology. In this thesis data mining techniques were used to extract data from online databases automatically by developing a tool (PubMed Info Extractor). On inputting queries, relevant information about the association and interactions among biological entities have been found in the retrieved collection. The approach has been applied on the case study of T2DM. Cell boundaries of the found gene components were also identified. Only those components were selected that show highly expressed genes in the pancreatic endocrine cells and skeletal muscle cell for determining the disrupted pathway after analyzing the normal functional pathway of T2DM. Further, bioinformatics tools were used for the topological analysis of obtained patterns. From a network generated by using the association rule mining technique showing associations, their nature of relationship, seven strongly connected components have been identified. These components represent P53, HNF1Alpha, HNF1Beta, INSR, INS, IL-6 and GnRH as the regulators or initiators of seven different biological regulatory pathways. This research is an evidence for the association of T2DM with the genes that are involved in different pathways of cancer cell metabolism, growth regulation, proliferation control etc along with insulin signaling pathway, mTOR pathway, MODY pathway, glycolysis, lipid homeostasis, Age-rage signaling pathway, MAPK pathway, p53 pathway. Self inhibition of ngn3 is also acknowledged in these components. In diabetic patients, pancreatic islets in case of fasting lessen PKA and mTOR activity and induce Sox2 and Ngn3 expression and insulin production. Self inhibition of Ngn3 can therefore affect the insulin production. This research gives a unit framework model of system biology which gives better understanding of intrinsic disease mechanism.

This research regarding T2DM which will facilitate the researchers to comprehend the system of T2DM disease mechanism and how to cure it with respect to personalized drugs.

Contents

Author’s Declaration	iv
Plagiarism Undertaking	v
List of Publications	vi
Acknowledgements	vii
Abstract	viii
List of Figures	xiv
List of Tables	xvi
Abbreviations	xxii
1 Introduction	1
1.1 Background	1
1.1.1 Association mining	2
1.1.2 Genetic interaction networks	4
1.1.3 Connected components in association networks	4
1.1.4 Metabolic syndrome	5
1.1.4.1 Diabetes Mellitus	6
1.1.4.2 Ankylosing Spondylitis	11
1.2 Applications	12
1.3 Ongoing and near future trends of data mining	13
1.4 Problem in focus	16
1.5 Research problem	17
1.6 Research objectives	19
1.7 Research philosophy	20
1.8 Research hypothesis	21
1.9 Research methodology	22
1.9.1 Data Retrieval through PubMed Info Extraction (PMIE)	22
1.9.2 Association Rules Generation	22

1.9.3	Extraction of connected components within the network for pathways crosstalk	23
1.9.4	Topological analysis of normal and disrupted pathways	23
1.10	Summary	25
2	Literature Review	26
2.1	Motivation of current research	27
2.2	Theory and background information	28
2.2.1	Association studies using data mining techniques	28
2.2.2	Connected components in biological networks	29
2.2.3	Topological analysis of genetic interaction networks	31
2.3	Contemporary research work	32
2.3.1	Data mining techniques	32
2.3.2	Related techniques on extracting strongly connected components in a graph	36
2.3.3	Topological analysis of genetic networks	38
2.3.4	Limitations and bottlenecks of existing approaches	40
2.4	Current research vs. existing work	41
2.5	Summary	42
3	Methodology details	44
3.1	Introduction	44
3.2	Tools and equipment	45
3.2.1	Hardware specifications	45
3.2.2	Software(s), tool(s) used	45
3.2.2.1	Windows platform	45
3.2.2.2	Language used in project	46
3.2.2.3	Biological database	46
3.2.3	Bioinformatics tools	47
3.3	Data collection from biological database	47
3.3.1	System requirements	48
3.3.2	Getting keywords	48
3.3.3	Connecting to PubMed	48
3.3.4	Extracting articles details	49
3.4	Association rules mining	49
3.4.1	Data collection	49
3.4.2	Identification of hypothetical association terms	50
3.4.3	Processing phase: segmentation	51
3.4.4	Processing phase: sentence splitting technique	52
3.4.5	Processing phase: text documents representation	52
3.4.6	Mining association rules	53
3.4.7	Transitive association rule mining	53
3.4.8	Nature of relationship	54
3.5	Extraction of strongly connected components within a network	55
3.5.1	Extraction of connected components within association graph	56

3.5.2	Identification of cell boundaries of components in a network	57
3.5.3	Analysis of functional pathway of genes	58
3.5.4	Analysis of components pathways	58
3.5.5	Selection of the components involved in specific cells	58
3.5.6	Identification of disrupted pathways	59
3.6	Topological analysis of biological networks	59
3.6.1	Extraction of genes associated with disease of interest from the literature	59
3.6.2	Scanning protein-protein interactions	60
3.6.3	Construction of PPIs network and extracting the giant component from the extended network	60
3.6.4	Topological analysis of protein interaction network	61
3.6.5	Shrinking a network	61
3.6.6	Identification of structural holes	62
3.6.7	Scale-freeness topology of the networks	62
3.6.8	Detection of hub nodes	63
3.7	Summary	63
4	Results and analysis	64
4.1	Data collection from biological database for T2DM	64
4.1.1	Getting Keywords related to T2DM	65
4.1.2	Connecting to PubMed	65
4.1.3	Extracting articles details	67
4.2	Association rules mining techniques to generate T2DM network	73
4.2.1	Phase 1: Data collection	73
4.2.2	Phase 2: Identification of hypothetical association terms for T2DM	74
4.2.3	Phase 3: Processing phase	74
4.2.3.1	Segmentation	74
4.2.3.2	Sentence splitting technique	75
4.2.3.3	Text documents representation	76
4.2.4	Mining association rules	76
4.2.4.1	Biological interpretation of T2DM network:	80
4.2.4.2	Self inhibition of Ngn3:	83
4.2.5	Transitive association rule mining	84
4.2.6	Nature of relationship between genes responsible for T2DM	85
4.3	Extraction of strongly connected components within a T2DM network	86
4.3.1	Extraction of connected components within association graph	86
4.3.2	Identification of cell boundaries of components genes in a network	95
4.3.3	Analysis of functional T2DM pathway	97
4.3.4	Analysis of components pathways	100
4.3.5	Selection of the components involved in specific cells	103
4.3.6	Identification of disrupted pathway	110

4.4	Topological analysis of Ankylosing Spondylitis protein-protein interactions network	113
4.4.1	Key nodes in the PPI network:	116
4.4.2	Structural holes and scale freeness of the network topology: .	122
5	Conclusion and recommendations	125
5.1	Conclusion	125
5.2	Future work	128
	Bibliography	130
	Appendix A	156

List of Figures

2.1	Connected graph in an undirected network.	29
2.2	Connected components in an undirected network.	29
2.3	Connected graph in a directed network.	30
2.4	Strongly connected component is a directed graph.	30
3.1	Architecture Model of current research.	45
3.2	Architecture Model of PubMed Information Extractor	50
3.3	Conceptual model of mining association rules	51
3.4	Flow diagram for extraction of strongly connected components	55
4.1	Screenshot of the Output of the PubMed Info Extractor (PMIE) against Type 2 Diabetes for the year 2017	66
4.2	Screenshot of the Titles of the output of figure 5.1 with URLs	67
4.3	Screenshot of the paper details against specific query	68
4.4	Graphical representations of the data against PubMed queries	68
4.5	Paper details accuracy results against the query terms for year 2013	71
4.6	Paper details accuracy results against the query terms for year 2014	71
4.7	Paper details accuracy results against the query terms for year 2015	72
4.8	Paper details accuracy results against the query terms for year 2016	72
4.9	Paper details accuracy results against the query terms for year 2017	72
4.10	Screenshot of Gene variants and notable protein coding genes for association rules	75
4.11	Graphical representation of association rules showing association between genes	79
4.12	Association among GWAS known gene variants of T2DM and their consequents by using Gephi [161]	80
4.13	Transitive Association among gene variants of T2DM	85
4.14	Association among gene variants of T2DM along with nature of relationship	86
4.15	Extracted component 1 within the association network in which P53 act as a gene regulator and involve in the activation or inhibition of various other genes of different biological pathways.	88
4.16	Extracted component 2 within the association network in which HNF1-Alpha act as a regulator and involve in the activation or inhibition of various other genes of different biological pathways.	89

4.17	Extracted component 3 within the association network in which HNF1-Beta act as a regulator and involve in the activation or inhibition of various other genes of MODY, T2DM, Insulin secretion pathways [163].	90
4.18	Extracted component 4 within the association network in which INSR act as a regulator and involve in the activation or inhibition of various other genes of Insulin secretion KEGG pathway, Insulin signaling KEGG pathway, Apoptosis KEGG pathway, Glycolysis/Gluconogenesis KEGG pathway [163].	90
4.19	Extracted component 5 within the association network in which TSH act as a regulator and involve in the activation or inhibition of various other genes for the regulation of lipolysis in adipocytes KEGG Pathway, cGMP-PkG signaling KEGG Pathway, Insulin signaling KEGG pathway [163].	91
4.20	Extracted component 6 within the association network in which IL-6 act as a regulator and involve in the activation or inhibition of various other genes of Non-alcoholic fatty liver disease KEGG pathway, TNF signaling KEGG pathway, Adipo-cytokine signaling KEGG pathway, PPAR signaling KEGG pathway, P13K-AKT signaling KEGG pathway, Apoptosis KEGG pathway [163].	91
4.21	Extracted component 7 within the association network in which GnRH act as a regulator and involve in the activation or inhibition of various other genes [163].	95
4.22	Details retrieved from The Human Protein Atlas against each components gene [124]	97
4.23	Expression of genes in different human cells	97
4.24	Schematic depiction of the insulin signal transduction cascade. [194]	98
4.25	Pathway crosstalk between component 3 and component 6	111
4.26	Protein-Protein interactions of seed proteins from STRING database.	113
4.27	Extended network includes one giant network and its small components based on Energy level	114
4.28	Overview of the Protein-Protein interactions derived from seed proteins respectively scanned from STRING database	115
4.29	Network showing the node with highest betweenness centrality and closeness centrality.	115
4.30	Partitioning of networks on the basis of degree of each node.	120
4.31	Vectors of the nodes on the basis of size to show the most required nodes in the network.	121
4.32	Topology of the backbone network	121
4.33	Structural holes in a network	122
4.34	Scale freeness topology of a network	123

List of Tables

3.1	Relationship thesaurus used in the current research	55
3.2	Algorithms for extracting strongly connected components	56
4.1	Last five years results evaluation by different users against possible T2DM queries	70
4.2	Articles retrieved from PubMed using PubMed Info Extractor against specific queries of Type II Diabetes Mellitus disease.	73
4.3	INS Subgroup and its association with other genes	77
4.4	INSR Subgroup and its association with other genes	78
4.5	IRS1/2 Subgroup and its association with other genes	78
4.6	Genes involved in identified components	92
4.7	Pathways involved in extracted components	101
4.8	Pathways involved in extracted components	102
4.9	Protein expression of component 1's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].	104
4.10	Protein expression of component 2's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].	105
4.11	Protein expression of component 3's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].	106
4.12	Protein expression of component 4's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].	107
4.13	Protein expression of component 5's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].	108
4.14	Protein expression of component 6's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].	109
4.15	Protein expression of component 7's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].	110
4.16	The general network measurements for networks	116
4.17	The list of genes extracted from literary database showing associa- tion with essential Ankylosing spondylitis [123], [124], [126], [159]	117

A.1	Common risk gene variants for T2DM, identified by GWAS	157
A.2	Akt Subgroup and its association with other genes	159
A.3	mTOR Subgroup and its association with other genes	159
A.4	TGFBR Subgroup and its association with other genes	159
A.5	TGFBR Subgroup and its association with other genes	160
A.6	THADA Subgroup and its association with other genes	160
A.7	P300 Subgroup and its association with other genes	160
A.8	P300-SP1 Subgroup and its association with other genes	160
A.9	P53 Subgroup and its association with other genes	160
A.10	IGFBP3 Subgroup and its association with other genes	161
A.11	FN1 Subgroup and its association with other genes	161
A.12	STAT1 Subgroup and its association with other genes	161
A.13	MIZ1 Subgroup and its association with other genes	161
A.14	TGFBR1/2 Subgroup and its association with other genes	162
A.15	SMAD2 Subgroup and its association with other genes	162
A.16	FST Subgroup and its association with other genes	162
A.17	Act Subgroup and its association with other genes	162
A.18	TRAF2 Subgroup and its association with other genes	163
A.19	PIP3 Subgroup and its association with other genes	163
A.20	ADAMTs9 Subgroup and its association with other genes	163
A.21	GCK Subgroup and its association with other genes	163
A.22	C-MYC Subgroup and its association with other genes	164
A.23	P13k Subgroup and its association with other genes	164
A.24	SHIP2 Subgroup and its association with other genes	164
A.25	PDK1/2 Subgroup and its association with other genes	164
A.26	SIRT1 Subgroup and its association with other genes	165
A.27	TCF7L2 Subgroup and its association with other genes	165
A.28	SOCS3 Subgroup and its association with other genes	165
A.29	IGF1 Subgroup and its association with other genes	166
A.30	CDKN2A Subgroup and its association with other genes	166
A.31	TNFAlpha Subgroup and its association with other genes	166
A.32	SLC30A8 Subgroup and its association with other genes	166
A.33	TNFR1 Subgroup and its association with other genes	166
A.34	IGF2BP2 Subgroup and its association with other genes	167
A.35	CDKAL1 Subgroup and its association with other genes	167
A.36	MTNR1B Subgroup and its association with other genes	167
A.37	PROX1 Subgroup and its association with other genes	167
A.38	HHEX Subgroup and its association with other genes	168
A.39	NOTCH Subgroup and its association with other genes	168
A.40	GLIS3 Subgroup and its association with other genes	168
A.41	LEP Subgroup and its association with other genes	168
A.42	ACDC Subgroup and its association with other genes	169
A.43	NF-KB Subgroup and its association with other genes	169
A.44	C-JUN Subgroup and its association with other genes	169

A.45 GNRHR Subgroup and its association with other genes	169
A.46 GluT4 Subgroup and its association with other genes	169
A.47 EGFR Subgroup and its association with other genes	170
A.48 GnRH Subgroup and its association with other genes	170
A.49 PAA Subgroup and its association with other genes	170
A.50 Gq/11 Subgroup and its association with other genes	170
A.51 PLcBeta Subgroup and its association with other genes	170
A.52 NeuroD Subgroup and its association with other genes	171
A.53 Foxa1/2 Subgroup and its association with other genes	171
A.54 PPARG Subgroup and its association with other genes	171
A.55 Ptf1a Subgroup and its association with other genes	172
A.56 HNF6Subgroup and its association with other genes	172
A.57 HNF1Beta Subgroup and its association with other genes	172
A.58 HNF4Alpha Subgroup and its association with other genes	172
A.59 Ngn3 Subgroup and its association with other genes	173
A.60 HNF1Alpha Subgroup and its association with other genes	173
A.61 Mafa Subgroup and its association with other genes	173
A.62 PDX1 Subgroup and its association with other genes	173
A.63 Hes1 Subgroup and its association with other genes	174
A.64 JAZF1 Subgroup and its association with other genes	174
A.65 CDC123 Subgroup and its association with other genes	174
A.66 CAMk1D Subgroup and its association with other genes	174
A.67 aPKC Subgroup and its association with other genes	175
A.68 ADCYS Subgroup and its association with other genes	175
A.69 IL-6 Subgroup and its association with other genes	175
A.70 IL-6R Subgroup and its association with other genes	175
A.71 ObR Subgroup and its association with other genes	176
A.72 AdipoR Subgroup and its association with other genes	176
A.73 IP3R Subgroup and its association with other genes	176
A.74 PKC Subgroup and its association with other genes	176
A.75 PLA2 Subgroup and its association with other genes	176
A.76 Sox9 Subgroup and its association with other genes	177
A.77 TSH Subgroup and its association with other genes	177
A.78 TSHR Subgroup and its association with other genes	177
A.79 UCP3 Subgroup and its association with other genes	177
A.80 UCP2 Subgroup and its association with other genes	177
A.81 BCL11A Subgroup and its association with other genes	178
A.82 PkG Subgroup and its association with other genes	178
A.83 PkG Subgroup and its association with other genes	178
A.84 BAR Subgroup and its association with other genes	178
A.85 SREBP-1C Subgroup and its association with other genes	178
A.86 AMPK Subgroup and its association with other genes	179
A.87 HSL Subgroup and its association with other genes	179
A.88 GS Subgroup and its association with other genes	179

A.89 AC Subgroup and its association with other genes	180
A.90 Pka Subgroup and its association with other genes	180
A.91 GSK3 Subgroup and its association with other genes	180
A.92 ACC Subgroup and its association with other genes	180
A.93 CAPN10 Subgroup and its association with other genes	181
A.94 WFS1 Subgroup and its association with other genes	181
A.95 BAD Subgroup and its association with other genes	181
A.96 DUSP9 Subgroup and its association with other genes	181
A.97 GrB2 Subgroup and its association with other genes	181
A.98 SMAD3 Subgroup and its association with other genes	182
A.99 G6PC Subgroup and its association with other genes	182
A.100UCP1 Subgroup and its association with other genes	182
A.101Mc4R Subgroup and its association with other genes	182
A.102RXR Subgroup and its association with other genes	182
A.103GK Subgroup and its association with other genes	183
A.104AP-1 Subgroup and its association with other genes	183
A.105GF2 Subgroup and its association with other genes	183
A.106PyK Subgroup and its association with other genes	183
A.107AdPLA Subgroup and its association with other genes	183
A.108ASK1 Subgroup and its association with other genes	184
A.109SHC1 Subgroup and its association with other genes	184
A.110CGI-58 Subgroup and its association with other genes	184
A.111PEPCK Subgroup and its association with other genes	184
A.112GnRH Subgroup and its association with other genes	184
A.113PLIN Subgroup and its association with other genes	185
A.114CDKN2B Subgroup and its association with other genes	185
A.115DGKB Subgroup and its association with other genes	185
A.116JNK1/2 Subgroup and its association with other genes	185
A.117Src Subgroup and its association with other genes	185
A.118AS-160 Subgroup and its association with other genes	186
A.119ATGL Subgroup and its association with other genes	186
A.120GCKR Subgroup and its association with other genes	186
A.121FABP Subgroup and its association with other genes	186
A.122FBP Subgroup and its association with other genes	186
A.123ACC Subgroup and its association with other genes	186
A.124PDE-3B Subgroup and its association with other genes	187
A.125Transitive association of INS with other gene variants	187
A.126Transitive association of INSR with other gene variants	187
A.127Transitive association of IRS1 with other gene variants	188
A.128Transitive association of Akt with other gene variants	188
A.129Transitive association of mTOR with other gene variants	188
A.130Transitive association of THADA with other gene variants	189
A.131Transitive association of P300 with other gene variants	189
A.132Transitive association of ADAMTs9 with other gene variants	189

A.133	Transitive association of P53 with other gene variants	189
A.134	Transitive association of IGFBP3 with other gene variants	190
A.135	Transitive association of FN1 with other gene variants	190
A.136	Transitive association of FST with other gene variants	190
A.137	Transitive association of Act with other gene variants	190
A.138	Transitive association of TGFB with other gene variants	191
A.139	Transitive association of TGFBR with other gene variants	191
A.140	Transitive association of TGFBR1/2 with other gene variants	191
A.141	Transitive association of TRAF2 with other gene variants	191
A.142	Transitive association of SIRT1 with other gene variants	191
A.143	Transitive association of GCK with other gene variants	192
A.144	Transitive association of STAT1 with other gene variants	192
A.145	Transitive association of SLC30A8 with other gene variants	192
A.146	Transitive association of IGF2BP2 with other gene variants	193
A.147	Transitive association of TCF7L2 with other gene variants	193
A.148	Transitive association of CDKN2A with other gene variants	194
A.149	Transitive association of GLIS3 with other gene variants	194
A.150	Transitive association of NOTCH with other gene variants	194
A.151	Transitive association of P53 with other gene variants	194
A.152	Transitive association of HNF1Beta with other gene variants	195
A.153	Transitive association of HNF6 with other gene variants	195
A.154	Transitive association of NeuroD with other gene variants	195
A.155	Transitive association of Hes1 with other gene variants	196
A.156	Transitive association of Sox9 with other gene variants	196
A.157	Transitive association of Ngn3 with other gene variants	196
A.158	Transitive association of HNF1Alpha with other gene variants	197
A.159	Transitive association of HNF4Alpha with other gene variants	197
A.160	Transitive association of PDX1 with other gene variants	198
A.161	Transitive association of Ptf1a with other gene variants	198
A.162	Transitive association of MafA with other gene variants	198
A.163	Transitive association of CDKAL1 with other gene variants	199
A.164	Transitive association of HHEX with other gene variants	199
A.165	Transitive association of PROX1 with other gene variants	199
A.166	Transitive association of IL-6 with other gene variants	200
A.167	Transitive association of LEP with other gene variants	200
A.168	Transitive association of IL-6R with other gene variants	200
A.169	Transitive association of ObR with other gene variants	200
A.170	Transitive association of SOCS3 with other gene variants	201
A.171	Transitive association of ACDC with other gene variants	201
A.172	Transitive association of AdipoR with other gene variants	201
A.173	Transitive association of NF-KB with other gene variants	201
A.174	Transitive association of GnRHR with other gene variants	202
A.175	Transitive association of GnRH with other gene variants	202
A.176	Transitive association of PAA with other gene variants	202

A.177	Transitive association of Gq/11 with other gene variants	202
A.178	Transitive association of PLA2 with other gene variants	202
A.179	Transitive association of PLcBeta with other gene variants	203
A.180	Transitive association of Pkc with other gene variants	203
A.181	Transitive association of IP3R with other gene variants	203
A.182	Transitive association of TNFR1 with other gene variants	203
A.183	Transitive association of PPARG with other gene variants	204
A.184	Transitive association of AMPK with other gene variants	204
A.185	Transitive association of GSK-3 with other gene variants	204
A.186	Transitive association of P13k with other gene variants	205
A.187	Transitive association of PDK1/2 with other gene variants	205
A.188	Transitive association of aPKC with other gene variants	205
A.189	Transitive association of SHIP2 with other gene variants	206
A.190	Transitive association of PIP3 with other gene variants	206
A.191	Transitive association of JAZF1 with other gene variants	206
A.192	Transitive association of CAMK1D with other gene variants	206
A.193	Transitive association of CDC123 with other gene variants	206
A.194	Transitive association of ADCYS with other gene variants	207
A.195	Transitive association of PKa with other gene variants	207
A.196	Transitive association of UCP3 with other gene variants	207
A.197	Transitive association of TSH with other gene variants	207
A.198	Transitive association of TSHR with other gene variants	207
A.199	Transitive association of GS with other gene variants	208
A.200	Transitive association of AC with other gene variants	208
A.201	Transitive association of BAR with other gene variants	208
A.202	Transitive association of BCL11A with other gene variants	208
A.203	Transitive association of PKG with other gene variants	208
A.204	List of nodes with their CC values	209
A.205	List of nodes with their BC values	211
A.206	List of nodes with their BC values	212
A.207	The list of proteins with both high BC and large degree and their functions at the threshold of 0.01	213
A.208	List of nodes with their Degree values	215

Abbreviations

ADCY5	Adenylate cyclase 5
ADAMTS9	ADAM metallopeptidase with thrombospondin type 1 motif, 9
BC	Betweenness centrality
BCL11A	B-cell CLL/lymphoma 11A
CC	Closeness centrality
CDC	Center for disease control and prevention
CHCHD9	Putative coiled-coil-helix-coiledcoil-helix domain-containing protein
CENTD2/ ARAP1	ANK repeat and PH domaincontaining protein 1
CDKN2A/2B	Cyclin dependant kinase inhibitor 2A/2B
CDKAL1	CDK5 regulatory subunit associated protein 1-like1
CDC123/CAMK1D	Cell division cycle 123 homolog (<i>S. cerevisiae</i>)
DGKB	Diacylglycerol kinase beta1
DUSP9	dual specificity phosphatase 9
FTO	Fat mass and obesity associated protein
GI	Genetic Interaction
GO	Gene Ontology
GSIS	Glucose stimulated insulin secretion
GWAS	Genome Wide association studies
GDM	Gestational Diabetes Mellitus
GCK	Glucokinase
GCKR	Glucokinase regulatory protein

HPA	Human Protein Atlas
HNF1Beta	Hepatocyte nuclear factor 1-beta
HHEX/IDE	Hematopoietically expressed homeobox -insulin-degrading enzyme
HMGA2	High mobility group protein HMGI-C
HNF1A	Hepatocyte nuclear factor 1-alpha
IDDM	Insulin dependent diabetes
IGF2BP2	Insulin like growth factor 2 binding protein 2
IRS1	Insulin receptor substrate 1
JAZF1	juxta-posed with another zinc finger gene 1
KCNJ11/ABCC8	Potassium inwardly-rectifying channel, subfamily J, member 11
KLF14	Krueppel-like factor 14
KCNQ1	Potassium voltage-gated channel subfamily KQT member 1
LADA	Latent autoimmune diabetes of adults
MTNR1B	Melatonin receptor type 1B
NCDs	Non-communicable diseases
NOTCH2	Neurogenic locus notch homolog protein 2
PMIE	PubMed Info Extraction
PPI	Protein Protein Interaction
PPARG	Peroxisome proliferative activated receptor gamma gene
PROX1	prospero homeobox 1
PRC1	Protein regulator of cytokinesis 1
SLC30A8	Solute carrier family 30, member 8
SCC	Strongly connected components
TCF7L2	Transcription factor 7 like 2
TP53INP1	Tumor protein p53-inducible nuclear protein 1
TSPAN8/LGR5	tetraspanin8
THADA	Thyroid Adenoma Associated

WHO	World Health Organization
WFS1	Wolfram Syndrome1
ZFAND6	AN1-type zinc finger protein 6
ZBED3	Zinc finger BED domain-containing protein 3

Chapter 1

Introduction

1.1 Background

Biological databases extend quickly with the exponential development of biological data since the time of 1990s with the beginning of Human Genome Project [1]. The text or data could be represented in various structures; it could be well organized as if there should arise an occurrence of databases, it could be semi organized as in a website page with Meta data that depicts or gives information about other data, or it could be absolutely unstructured as in a simple text records. In any case, this biological data is not information. The vast majority believe that the expressions "information" and "data" are exchangeable and mean a similar thing. However, data and information, both terms are discrete entities. In simple words, Data is something that if we don't put it in proper setting, it implies nothing useful about anything. Nonetheless, information will be data organized in a way that enables it to be used by people in some noteworthy way. Contrasted with circumstances before 2003, the key issue today has swung to techniques for information mining. Information mining or content mining is procedure of recognizing the knowledge from vast data index [2].

1.1.1 Association mining

An association can be characterized as a bi-directional ramification between two objects as $A \rightarrow C$, where 'A' and 'C' are the two objects of intrigue. A is an antecedent and C is a consequent, or vice-versa. The objects are words or expressions with regards to content mining. In the biomedical domain the diverse terms are related with each other by words or expressions like "strongly related", "weakly related", "not related" and numerous different word phrases like these. This research not only spotlight on finding the directionality of the relationship among objects of intrigue rather yet in addition the nature of affiliation implies how two objects have association with each other. The extending rate of knowledge in gene disease affiliations can barely coordinate with the development of biological data. One of the significant objectives of the post-genome period is having the understanding of what role genetics play in maladies. Finding distinctive patterns and relationship among objects of interest for differed or single domain from colossal accumulation of data can enhance the speed and exactness of decision making [2]. Association rules mining that is among the acknowledged procedures used to mine data into effective information, can be generated in two steps. Firstly, an algorithm has been applied to entire collected dataset to find the rules that occur more repeatedly. Each rule has an antecedent term and a consequent term. An antecedent represents the term for which we are interested to find the association with all other specific terms. Whereas consequent is the term which shows that this has some type of association with the antecedent term or which occurs more frequently with antecedent in the collected dataset.

Once the frequent association rules have been extorted then the conception of support and confidence is generally used to verify the potency of those frequent item sets. However there are many intended measures for determining the potency of generated rules. The measures include support, confidence, coverage, prevalence, recall, accuracy, specificity, lift, leverage, added value, relative risk, jaccard, certainty factor, odds ratio, Laplace correction, cosine, Yules Q, Yules Y, Piatetsky-Shapiro etc. The main measures to achieve the goal of assessing the strength of the

rules representing the association between two biological objects are confidence, support and lift. The support value of the generated rule showing association can be calculated as the proportion of documents that contain both antecedent and consequent terms and can be premeditated by the subsequent likelihood.

Support: $(A \rightarrow C) = P(A \wedge C)$

An association rule having low support may represent that this rule happens merely by chance and is more likely to be not significant from a scientific viewpoint in case of genetic terms/genes. This makes the perception that the terms/genes that represent such low support rules may not be contributing much to endorse a specific disease. Support is often used to eradicate insignificant rules and to exploit the significant rules.

Confidence value is the ratio of support value to the amount of documents that contain the antecedent term and can be embodied by the subsequent conditional likelihood:

Confidence: $(A \rightarrow C) = P(C|A)$

Confidence value of a generated rule shows the reliability of the implicated rule. Reliability and the confidence value are directly proportional to each other and vice versa. The higher the confidence value of a rule, the more probability the consequent term has to be present in association with an antecedent term. The value of lift for the generated rule can be calculated as the ratio of the documents that contains consequent term given that antecedent term has occurred to the documents that contains consequent term in all transactions of the collected dataset. In other words, the lift is the ratio of one confidence to another confidence, by the perception that consequent transactions and antecedent terms are sovereign to each other.

A lift value more than one shows that the relation between two terms is more significant than that in which the antecedent and consequent terms are sovereign.

1.1.2 Genetic interaction networks

Genetic association (GI) is a consistent interaction between two genes that influences the phenotype (the perceptible attributes) of an organism. Information of genetic interactions can help with distinguishing pathways, perceive gene functionality and discover potential drug targets. Present day research instruments empower performing substantial scale mapping of GIs. Synthetic lethal (or sick) interactions are exceptionally compelling on the grounds that they can empower the identification of genes whose products buffer each other and involve in the related crucial biological process. In addition, it might help with concluding the functionality of obscure genes and the association between various biological pathways. Next, the relation between genetic interactions and an assortment of different characteristics of a gene or protein sets help to investigate functional attributes, sequence similarity and protein interactions. The Gene Ontology (GO) venture supplies a predictable depiction of gene products. It depicts gene products regarding their related biological processes, cell segments and molecular functions [3]. Just 1% of the genes that shows interactions with other genes produce proteins that are the part of same complex. However, it was uncovered that genes that share a noteworthy number of neighbors in the GI network completely tend to create physically associating proteins. This implies for two genes, a and b, and the set of genes represented by $g(a)$ and $g(b)$ that are interrelated with a and b, respectively, the more prominent the number of genes that are in $g(a)$ and $g(b)$, the higher the likelihood that a and b create physically collaborating proteins. It is sensible that such expectation of protein cooperation, by observing the evident genetic association, relies upon measure of known genetic interactions.

1.1.3 Connected components in association networks

Many real life applications are based on graphs. They are elementary to various tribulations in computational and general science areas. Dilemma that crop up on graph is the act due to complicated graph algorithms and size [4], [5]. A directed

graph G can be such that it has vertices and edges that have defined directions denoted as V and E and forms the pair (V, E) and $E \subseteq V \times V$ is a set of directed graph. If $(v, u) \in E$, then u is known as instantaneous successor of v , and v is called instantaneous predecessor of u . Graphs having no directions may be pragmatic as a particular sort of graphs with definite directions and in that case directed edges have no importance $(v, u) \in \mathcal{E} \leftrightarrow (u, v) \in \mathcal{E}$ [5], [6]. A directed graph is called unequivocally associated if there exist a way between v to u and u to v . In the event that the entire graph has a similar property, at that point the graph is unequivocally associated [5], [7]. In the current research we focused to imply the developing techniques on the most crucial metabolic syndromes that are Diabetes Mellitus and AS. These two syndromes were selected after going through their statistics and continuous elevation in prevalence.

1.1.4 Metabolic syndrome

Metabolism is the progression that the body uses to get influence vitality from the nourishment someone eats. Chemicals in our digestive system smash the nourishment into sugars and acids, our body's fuel. A metabolic disorder can happen when a substance response to the human body and brings alteration in the usual metabolic process. Metabolic disorder can be hereditary single gene anomaly and many of them are autosomal recessive [8]. Metabolic syndrome (MetS) is a collection of medical conditions that incorporates hyperglycemia and/or insulin resistance [9], [10]. Due to the increase in deprived nutritional practices, MetS has turned out to be an escalating public health and economic saddle [11], [12], [13], [14], [15], [16]. MetS is allied with at least five-fold augmented threat in developing diabetes mellitus. Using data-mining methods in the research related to metabolic disorders is one of the best ways to employ bulky dimensions of available diabetes-related data for extorting the information.

1.1.4.1 Diabetes Mellitus

Diabetes Mellitus, often alluded just as diabetes, is a disorder caused by metabolism Syndrome. Genetic as well as environmental factors are playing crucial role in bringing about anomalous high glucose levels (hyperglycemia). It is currently a main source of morbidity and mortality all through the world. Diabetes is related with high rates of hospitalization, visual impairment, renal disappointment and non-traumatic amputation.. Diabetes occurs because of a decreased release of insulin (in type 1) by beta cells of pancreas or resistant effects as happen in type II DM and Gestational Diabetes. "Type 1 diabetes" has generally supersedes a few prior terminologies that includes adolescence diabetes and insulin dependent diabetes. Similarly, "type 2 diabetes" has supersedes a few prior expressions that includes adult onset diabetes, obesity related diabetes, and non-insulin-dependent diabetes. As among other types of diabetes (gestational diabetes, insulin-resistant type 1 diabetes (or "double diabetes"), type 2 diabetes), "type 3 diabetes" have been characterized by many different sources. This has advanced to involve infused insulin, and inactive immune system diabetes of adults (LADA) which is also known as "type 1.5" diabetes. Individuals that have the family history with Type II DM also have the susceptibility to be suffer with Maturity onset diabetes of the young (MODY) which is monogenic. In type 1 diabetes, our immune system erroneously demolishes the beta-cells, which are the cells in pancreas that release insulin. Our body regards these beta-cells as remote trespassers and decimates them. The obliteration can occur over few weeks, months, or years. At the point when enough beta cells are demolished, pancreas quits making insulin, or makes so little insulin that is expected to take insulin to live. If there is an occurrence of Type 2 diabetes then the body does not make use of insulin legitimately. In this case, we can say that the beta cells are resistive to insulin production. At initial stage, the beta-cells make additional insulin to compensate for it. But, late on, our pancreas are not able to hold and release sufficient insulin to maintain our blood glucose at ordinary levels. Some individuals with type 2 diabetes can deal with their diabetes with adhering to a good quality diet and exercise. The pathogenesis of the both types of diabetes, T1DM and T2DM is described by a

diminishment in the aptitude of the pancreatic β -cell to produce and secrete insulin, and in addition a reduction in β -cell mass by apoptosis [17], which prompts toxic glucose development in the blood and vitality consumption in peripheral tissues, for example, muscle, liver, and fat. In type 1, an immune system reaction prompts the outright inadequacy of insulin, though type 2 is characterized by a non-immune system relative insufficiency of insulin combined with defective insulin signaling. Of the diagnosed instances of diabetes, around 95% are type 2. It is anticipated that the influenced individuals with type 2 diabetes on the world will achieve well more than 300 million constantly till 2025 [18]. Diabetes by and large, hence, is a huge issue. This is a dynamic illness which prompts β -cell failure and at last β cell failure, as these cells will never again ready to discharge enough insulin because of deformities in glucose-stimulated insulin secretion (GSIS) and a decrease in β cell mass [17]. In type 2 diabetes, β cell loss is basically because of a decrease in β cell number because of an expansion in β cell apoptosis, as found in animal models and type 2 diabetic patients [19]. Since β cell dysfunction and apoptosis are principle obsessive segments of this malady, there has been a concentration in research to comprehend these basic components. Gestational diabetes (GDM) usually occurs during pregnancy. For most ladies, blood glucose levels will come back to normal after giving birth. GDM should have been tested consistently as there is significantly higher hazard for developing Type I diabetes sometime later in life [20]. Maturity onset diabetes of the young (MODY) is another most common type of diabetes that is due to mutations in an autosomal dominant gene [21] distracting the beta cells for insulin production. To discern MODY from the more common type of diabetes especially type 1 diabetes and type 2 diabetes, it is also termed as "monogenic diabetes [22], [23]. This type of diabetes entails more intricate permutations of factors that involve various genes and ecological aspects. Twelve different types of MODY have been identified up till now caused by twelve different genes. Among those MODY 2 and MODY 3 are the most common forms [24]. MODY should not be mystified with suppressed autoimmune diabetes of adults, a form of type 1 Diabetes Mellitus that shows slower progression to insulin dependence than child-onset type 1 DM, and that happens later on in life. The monetary

effect of Diabetes Mellitus is high and it is a noteworthy supporter of the heightening social insurance cost overall [25], [26]. Diabetes is a standout amongst the most widely recognized non-communicable infections all around. Predominance rates of Diabetes Mellitus change extensively among various populaces and ethnic gatherings overviewed [27]. Reliably high pervasiveness rates are currently being accounted for from a few developing nations [28]. The World Health Organization (W.H.O.) has evaluated that the worldwide number of individuals with diabetes will be more than twofold over the following 25 years and the developing world would bear an inexorably bigger encumber of infection in this period [29]. South Asia specifically is viewed as one of the zones of the most noteworthy increment in anticipated numbers. Diabetes is a serious, chronic illness that is on the ascent. Not any more, an ailment of prevalently rich countries, the predominance of diabetes is consistently expanding all around, most notably on the worlds middle wage nations. It is an essential public medical issue, one of four priority non communicable diseases (NCDs) directed for activity by world pioneers. As indicated by World Health Organization, the individuals with diabetes have mounted to 422 million till 2014 from 108 million in 1980 [30]. The overall pervasiveness of diabetes at global level among people that have age greater than 18 years has mounted from 4.7% in 1980 to 8.5% in 2014 [30]. In 2015, round about 1.6 million deaths were due to diabetes. Another 2.2 million deaths occurred due to the fact of having high blood glucose in 2012. Partial deaths inferable from high blood glucose ensue before coming to the age of 70 years. WHO expands that diabetes will be the seventh driving rationale for fatality in 2030 [30]. As indicated by the report, Type II diabetes represents 90% to 95% of all the diabetes cases. The insights demonstrate that, in 2017, an expected 8.8% of the grown-up populace worldwide had diabetes. This figure is anticipated to ascend to 9.9% by the year 2045. As indicated by the Centers for Disease Control and Prevention (CDC), around 1 out of 10 American grown-ups have diabetes. In the event that patterns proceed with, that figure is relied upon to twofold or triple by 2050. In 2012, 13.4 million ladies (11.2 percent) had diabetes, as indicated by the National Diabetes Report. Around 15.5 million men (13.6 percent) had it. The global prevalence (age-standardized) of diabetes

shows that this metabolic disorder has proliferated since 1980, from 4.7% to 8.5%. This mirrors an expansion in related hazard factors, for example, overweight or obesity. Over the last few decades, prevalence of diabetes has grown more rapidly in those nations that have low-and middle income as compared to those that have high-income. Diabetes caused 1.5 million deaths in 2012. Higher-than-ideal blood glucose caused an extra 2.2 million passings, by mounting the dangers of cardiovascular and different sicknesses. 43% of these 3.7 million fatalities ensue before coming to the age of 70 years [25], [26]. On 13 November 2016 in celebration of 25th World Diabetes Day, President of Pakistan Mamnoon Hussain while tending to Diabetes Awareness Walk and Blue-lightening ceremony communicated his perspectives that Pakistan may have eighth most astounding populace of diabetic patients by 2040. Such high predominance of diabetes calls for proceeded with endeavors to enhance open mindfulness in regards to the malady. Right now, more than seven million individuals are living with diabetes in Pakistan while another seven million are very nearly to diabetes if no preventive measures are taken. It is assessed that the number of diabetic individuals in Pakistan may achieve 14.4m by 2040. National Diabetes Fact Sheet, [18] published in 2011 shows that 25.8 million individuals, or 8.3% of the total U.S. populace is suffering from diabetes. This amount adds up the cost of diabetes in the U.S. for 2007 was 174 billion dollar. The overall depiction of diabetes is same around the world; having anticipated 285 million individuals inclined to diabetes in 2010 that represents 6.6% of the world's grown-up populace. Health services consumptions for diabetes are required to be \$490 billion for 2030, representing 11.6% of the cumulative social insurance use on the world [31]. Researchers and scientist have been working for many decades and provided important insights about the global epidemic of metabolic disorders. Type II diabetes is one of those that occur due to complex interactions between genes and environment factors. To find out what role the genetics is playing in type 2 diabetes, researchers are doing epidemiological studies, studies of candidate genes, and genetic linkage in families and have provided imperative imminent into several monogenic diabetes types, but still to understand the genetics of common

type II diabetes left overs a main confront. Over the last few years, many interesting articles have been presented in well known journal that are representing high through put genome wide association studies. These articles along with revealing several novel genetic loci that are associated with diabetes provide researchers to do molecular investigation for new targets, and also encouraging them to rethink the extent of genetic heterogeneity and probably the part of genetics itself in the pathogenesis of diabetes mellitus type II [32]. Genetic elements are acknowledged to be vital components in the cause of T2D. Racial discrepancy of T2D represents sturdy evidence for the genetic foundation of this sickness. However, the role that genetics play in the occurrence of diabetes is poorly obscure [33], [34], [35], [36]. Over the period of last seven years, advancements in genotyping expertise have assists swift progression in large scale genetic research. The development in the genetic studies of more common types of Type II diabetes has been sluggish primarily. Recently, a number of genes had been reproducibly related to T2D danger in more than one genome-wide association studies (GWAS) each creating a modest involvement to the overall hazard. All known alleles associated with type 2 diabetes risk are common and have a low penetrance in the preferred population. The role(s) of many of them still requires to be assured, and for most people, the biological and molecular mechanisms are far from being simply understood. The causes of type 2 diabetes are complicated [37]. Studies have acknowledged as a minimum one hundred fifty DNA versions which are related to the threat of causing type II diabetes and most of them are found to be present both in diabetic and non diabetic people. Everybody has a few variants that amplify possibility of having diabetes and others that decrease the risk of having this disease. It is the aggregate of those modifications that allows determining how much a person has the probability of person developing the ailment. The preponderance of genetic changes related to type 2 diabetes are notion to act by means of deviously varying the amount, timing, and region of gene expression. Those modifications in expression affect genes concerned in lots of facets of type II diabetes, which include the development and characteristic of pancreatic beta cells to the release and processing of insulin, and sensitivity of these cells to the effects of insulin.

Somehow, for the various alterations which have been related to type II diabetes, the mechanism with the aid of which these disparities contribute to disease risk is mysterious [38]. Genetic variations possibly act together with fitness and way of life elements to steer a man or woman's usual chance of type II diabetes. All of these elements are associated, without delay or circuitously, to the body's ability to release and respond to insulin. Health situations that predispose to the disease consist of overweight or obesity problems, insulin resistance, pre-diabetes (higher-than-normal blood sugar tiers that don't reach the cutoff for diabetes), and a form of diabetes known as gestational diabetes that happens at some point of being pregnant. Lifestyle factors along with smoking, a poor diet plan, and physical inactivity additionally boom the threat of type 2 diabetes [38]. It could be intricate to split genetic hazard from environmental risk. The latter is regularly prompted by way of the circle of relatives participants. For instance, mother and father with healthful eating conduct are probably to skip them directly to the following generation. However, genetics performs a huge element in determining weight.

1.1.4.2 Ankylosing Spondylitis

Ankylosing spondylitis is a Greek word got from, ankylos: twisted and spondylos: spinal vertebra. It is an interminable provocative ailment of hub skeleton that causes back agony and dynamic firmness and also aggravation of organs. Almost all age fellows, particularly kids yet the crest time of onset is 2030 years suffer from AS. However, the main recognized period of onset of side effects is in the second and third decade of life [39]. Ankylosing spondylitis regularly begins at a youthful age (from the teenagers to the third decade of life) and is more continuous in males (around 23 times more basic than in females). AS can likewise present as adolescent ankylosing spondylitis, more ordinarily in young men than young women. It's an immune system issue that may be provoked by specific types of bacterial or viral pollution. These germs or outside substances invigorate the resistant reaction that continues constantly even if the contamination is cured. As a result of defense system attacks the individuals own specific tissues [40].

Although the genetic background of AS is yet unknown, but research shows that genetic factors play a significant role in the cause of this disease. HLA- B27, marker present on the surface of leukocytes, seems to be involved because it is found more than 90 % of the individuals suffer with AS [41]. HLA- B27 allows the body's defense system to distinguish between its own or foreigner tissues that implies numerous people experiencing AS will have a ancestral foundation of the circumstances. Numerous individuals with AS won't be equipped for remembering any other person in their family with this illness, then again, since just around 20 % of those with HLA-B27 will encounter provocative conditions, for example, AS. On the basis of all these facts, it is thought that protein protein interactions (PPIs) governed the proteins that are encoded by these defenseless genes are critical in AS variation [42].

1.2 Applications

A significant amount of research work has been done towards text mining in recent years yet it is a challenging momentum issue. The act of discovering affiliations or relationship among different objects of intrigue is known as a form of information extraction. One basic approach is to retrieve relationship from a database of text documents is just perused the pertinent documents and construct the associations among various objects of intrigue manually. This approach would be unreasonable as far as time and exertion for a situation when we have vast volume of data particularly in biological domain where we have substantial corpus of information about genes, proteins, diseases and so on. To encourage the automatic extraction of relationship among biological objects that would be of extraordinary enthusiasm to biomedical specialists, different association revelation algorithms and techniques have been proposed. Biological frameworks include different entities at molecular level, for example, genes, proteins and other modules like that along with associations between those entities. There is a need to have the exact knowledge of the profusion outline of all these modules for clear understanding of a particular phenotype, the working of a cell or tissue, etiology of malady, or cell association. The

mechanistic perceptive of the pragmatic phenotype can be obtained by analyzing the biomedical data through which one can elucidate imperative features of the interactions among different molecular modules. Biological networks confine the multifaceted relations between genes, proteins, RNA molecules, metabolites and hereditary variants in the individuals cells and are usually interchangeably notorious to graphs. Graphs are depictions of complex system components in which the molecular entities are expressed as nodes that are linked by edges [43]. These graphs or networks present an abstract and perceptive structure to model diverse components of numerous omics information from the genome, transcriptome, proteome, and metabolome. The expedient illustration of the genetic components in graphs escort to the field of system biology—a discipline that crams holistic relations amid different organic components by coalescing graph theory, systems biology, and statistical analysis [3]. Furthermore, the quantitative tools of system biology put forward the impending to comprehend cellular organization and confine the impact of perturbations on these intricate intracellular systems. Network Medicine is an annex of network biology with a set of alert aims allied to disease biology, as well as consideration of disease etiology, classifying prospective biomarkers, and manipulative curative interventions, including drug targets, dosage, and synergism innovation [20]. The guarantee of system medication research is to build up an increasingly worldwide comprehension of how irritations engender in the framework by recognizing the pathways, sub-kinds of sickness states, and key segments in the systems that can be focused in clinical mediations. In addition, systems are the focal point of the "new science" in the biomedical information insurgency and interpretation to customized prescription [44].

1.3 Ongoing and near future trends of data mining

Information mining is helpful in different orders, which incorporates database management system (DBMS), Statistics, Artificial Intelligence (AI), and Machine

Learning (ML). The era of information mining applications was considered in the year 1980 basically by research driven tools concentrated on single tasks. In starting days, information mining algorithms work best for numerical information gathered from a solitary information base, and different information mining strategies have developed for level records, conventional and social databases where the information is put away in illicit portrayal [40]. Later on, with the confluence of Statistics and Machine Learning procedures, different calculations developed to mine the non numerical information and social databases. The field of information mining has been enormously impacted by the improvement of fourth era programming dialects and different related processing systems. In the beginning of information mining the vast majority of the calculations utilized just measurable methods. Later on they advanced with different figuring strategies like AI, ML and Pattern Reorganization/recognition [45]. Different information mining procedures and algorithms that have been implemented to mine the substantial volumes of heterogeneous information put away in the information stockrooms. Information mining has been becoming tremendous accomplishment due to its extensive appliance undertakings and logical encroachment indulgent. Various information mining methods have been efficiently executed in many areas particularly medicinal services, media transmission, extortion identification and so on [46]. The consistently intensifying intricacies in different fields and modifications in innovation have presented new obscurities to information mining; these obscurities integrate distinguishing information groups, information from divergent parts, progresses in calculation and systems supervision assets, inquire about and logical fields, regularly emerging industrial confronts and so on. Progressions in information mining with different reconciliations and implications of approaches and trials have formed the current information mining claims to deal with the sundry difficulties; the momentum models of information mining applications are mining the heterogeneous information, explore and logical processing patterns and business patterns [47].

There are different information mining techniques with various sentence structures, thus it is to be institutionalized for making helpful of the clients. Information mining applications needs to amass more in institutionalization of association dialects and adaptable client connections. The present procedures and algorithms of information preprocessing stage are not up to the check contrasted and its importance in discovering the novel examples of information [48]. In future there is an incredible need of information mining applications with proficient information preprocessing procedures. Information mining will enter in all fields of human life; the accessible information mining strategies are confined to mine the customary types of information just, and in future there is a possibility for information mining procedures for complex information objects to articles and transient information. The advancement of World Wide Web and its utilization develops, it will keep on generating perpetually substance, structure, and use information and the estimation of Web mining will continue expanding [49]. Research should be paying attention to edifice the correct pact of Web measurements, and their inference methods, extricating practice sculpts from information, considering how unique elements of the system show affect different Web measurements of interests, by what means the procedure models amend that are made-changing upgrades to the client, generating Web mining systems to increase different parts of Web administrations, strategies to distinguish known false and interlude recognition. Information mining has pulled in the examination in different logical processing applications, because of its productive investigation of information, finding significant new connections, examples and patterns with the assistance of different devices and procedures [50]. More research must be done in mining of logical information specifically approaches for mining galactic, organic, substance, and liquid dynamical information examination. The universal utilization of implanted frameworks in detecting situations plays major looming improvements in logical figuring will require another class of strategies fit for dynamic information investigation in flawed, appropriated structure. The examination in information mining requires more consideration in biological and ecological data investigation to use our natural environment and assets. Huge information mining research must be

done in molecular science issues [51].

1.4 Problem in focus

Researchers are making their efforts to best level in order to reveal the processes that underlie diabetic and Ankylosing spondylitis biological pathways in the clinical contexts . Identification of relationship among various biological terms, for example, genes, proteins, disease, medications and synthetic concoctions, and so forth, is a critical issue for biological analysts. Such data can be extricated from diverse kinds of biological literature that has been growing rapidly with the advancement in technology. Association rule mining that is extracting interactions or relationships among different biological items is the ever best data mining technique. Association rule mining strategies or algorithms allow to extract useful patterns or information from the massive anthology of data that started to appear on internet since the beginning of its creation in the form of text. The web, databases and many other online sources began to be seen as a huge repository of online information as a result of recent technological advancements in every field especially in medical and genomic data. Usually in data mining the data is structured and all the accessible tools are equipped to attain the acquaintance straightforwardly. But in text mining, the information is unstructured and even very vague. A manual analysis of the text documents is not potential and not very effective. The exploitation of automatic tools for scrutinizing the hefty quantity of credentials is exceedingly suggested. The execution of complex natural procedures requires the exact connection and direction of thousands of molecules. Methodical ways deal with about vast quantities of proteins, metabolites, and their alteration have uncovered complex molecular systems. These natural systems are essentially unique in relation to irregular systems and frequently display universal properties as far as their structure and association. Breaking down these systems gives novel bits of knowledge in understanding fundamental components controlling typical cell procedures and infection pathologies. There is an exceptional scenario where our competence to engender biomedical statistics has significantly surpassed our

capability to mine and scrutinize the data successfully and analyze the genetic of mainly these two issues. In this situation, effectual genetic data mining techniques will be fundamental in the development of better perceptive of inherent diseased systems to discern novel drugs and emergent conversant clinical conclusions making and sustaining the assistance for diabetic and Ankylosing spondylitis individuals. Conversely, to interpret the immense quantity of biomedical data into valuable imminent for scientific and healthcare appliances, there are data mining obscurities that have to be surmount such as managing boisterous and curtailed data, for example, disgracefully strident gene expression data and protein protein interactions data, with high false positive and false negative rates, handing out compute-intensive errands, for example, large-scale graph indexing, searching, and mining, integrating various data sources, for example, concerning genomic and proteomic statistics with clinical databases and exploiting biomedical data with ethical and solitude security. All these concerns pretense innovative confronts for information miners to provide effective in the data-intensive post-genomic era.

1.5 Research problem

Making sense of which genes cause which issue is an essential yet troublesome issue [44]. It has an arrangement of employments that differ from the screening of DNA and before time distinguishing proof of an illness, to investigation of gene succession and prescription headway [52]. Ordinarily, this system incorporates both distinguishing genes that are known to have job in the turmoil (through content investigation) and doing prelude tests or screening through linkage or alliance thinks about, copy number examinations, verbalization profiling so as to choose a game plan of promising probability for preliminary endorsement. Regardless, it is resource heightened both with respect to time hypothesis and money related expense. For the most part, to recognize the sickness and quality connection physically is coordinated in two phases. The essential stage is to limit the whole genome into a broadly huge genes that has a high probability of causing a disease.

Unmistakable courses subsist to deal with this stage, for instance, linkage investigation, sequencing of genome and affiliation discoveries [53], [54], [55]. By then, in the subsequent stage, pros likely evaluate the noteworthy qualities to attest whether they are truly candidates causing specific disease. This incorporates *in vitro* contemplates for every competitor quality. In this manner, a basic progress in this field has been the improvement of computational systems that can help the experts with tending to the essential time of this methodology through thusly sorting out a lot of possibility for last test endorsement to extend the yield of the subsequent stage. Biological information mining is ending up progressively by incorporating the whole channel of biological and medical disclosure progressions. Currently there occurs a consistently squeezing requirement for information mining specialists to work together with the scholars and clinical researchers. Information mining scientists have now the chance to add to the advancement of the life and clinical sciences by making innovative computerized procedures for finding helpful learning from huge level genuine biomedical information. Indeed, there are copious open doors for information mining analysts to traverse from the computational area into the biomedical space to add to the important logical interest with the scholars and human services researchers. A definitive accomplishment of the biological information mining for social insurance applications will in this way rely upon parallel changes both in the biological and clinical test strategies from the scholars and clinicians to give biological datasets for information mining network, and in the propelled registering procedures from the researchers, mathematicians, and analysts to give productive and viable approaches to extract the information for learning revelation. Content mining approaches to manage gene ailment association extraction have shown an advancement from clear systems that depend solely on co-occasion experiences [56], [57] to complex structures utilizing ordinary lingo dealing with techniques and machine learning computations [58], [59]. Well understood instruments for discovering gene disease affiliations join DAVID [60], GSEA [61], GOToolBox [62], rcNet [63] and various others. Regardless, a great part of the time, since the present clarifications of infirmity causative genes is far from consummation [64], and an arrangement of quality may simply contain

a short once-over of insufficiently elucidated genes, open procedures much of the time disregard to reveal the connection between gene sets and affliction phenotypes [65]. There is a need to make use of such data mining techniques and developed a system that can not only extract data from online databases automatically on giving the specific query to the system but also find the relevant useful information from that collected data related to T2DM and AS because of the facts that have been discussed in above section. Automatic extraction of such information from printed information can altogether improve natural research efficiency by keeping scientists in the know regarding the cutting edge in their examination area, by assisting them envision biological networks, and by producing novel speculations concerning innovative connections some of which can be great contender for encourage original research and approval.

1.6 Research objectives

Research objective 1: To reveal the genes associated with T2DM and AS that share a noteworthy number of neighbors in the GI network completely tend to create physically associating proteins. This implies for two genes, a and b, and the set of genes represented by $g(a)$ and $g(b)$ that are interrelated with a and b, respectively, the more prominent the number of genes that are in $g(a)$ and $g(b)$, the higher the likelihood that a and b create physically collaborating proteins.

Research objective 2: To give the better comprehension of hereditary genes involved in T2DM and AS at molecular level.

Research objective 3: To demonstrate the vulnerability to T2DM and AS framework, with other important known genes and how these genes influence the ordinary pathway in the organism's body to cause the particular diseases.

1.7 Research philosophy

Fast advancement in mechanized information securing methods has provoked enormous amount of information. In excess of 80 percent of the current information is made possible out of amorphous or semi-organized data collection. The revelation of proper patterns to scrutinize the contented reports from massive quantity of information is a lead matter. Content mining is a practice of removing captivating and non-trivial frameworks from colossal evaluation of content records. There subsist varied systems and devices to excavate the content and discover beneficial data for upcoming conjecture and fundamental guidance process. The fortitude of right and suitable content mining procedure gets better the speed and abatements the time and exertion mandatory for the extraction of important data.

Science is experiencing two quickly evolving marvels: one is the expanding abilities of the PCs and programming instruments from terabytes to petabytes and past, and the other is the progression in high-throughput sub-atomic science delivering heaps of information identified with genomes, transcriptomes, proteomes, metabolomes, interactomes, etc. Science has turned into an information serious science and as an outcome science and software engineering have turned out to be correlative to one another crossed over by different parts of science, for example, measurements, arithmetic and material science. The blend of flexible learning has caused the coming of huge information science, arrange science, and other new parts of science. System science for example encourages the framework level comprehension of the cell or cell segments and subprocesses. It is regularly additionally alluded to as frameworks science. The motivation behind this field is to comprehend living beings or cells in general at different levels and components. biological science is presently confronting the difficulties of investigating huge genomic information and enormous natural systems. Protein-protein connections (PPIs) assume enter jobs in life forms, for example, signal transduction, translation controls, and insusceptible reaction, and so forth. Recognizable proof of PPIs empowers better comprehension of the useful systems inside a cell. Normal test strategies for distinguishing PPIs are tedious and costly. Recent advancements in computational

methodologies for deriving PPIs from protein successions in view of coevolution hypothesis maintain a strategic distance from these issues. In the coevolution hypothesis demonstrate, cooperated proteins may indicate coevolutionary changes and have comparative phylogenetic trees. The current coevolution strategies rely upon different succession arrangements (MSA); be that as it may, the MSA-based coevolution techniques regularly deliver high false positive connections.

1.8 Research hypothesis

We are progressively intrigued by applying enormous information driven strategies to accomplish a superior comprehension of essential science and infection forms.

Numerous vast articulation profiling thinks about and protein protein cooperation (PPI) screens right now in progress are relied upon to acquire generous new data on the known but then unfamiliar cell forms. Sorting out these outcomes will be instrumental in making a brought together perspective of biomolecular pathways of living cells. The guarantee of powerful combination of these heterogeneous datasets depends on the presumption that there is a connection between communication of two proteins in vivo and comparable articulation examples of the qualities encoding those proteins.

In this research, we generally hypothesize the possible interaction between the common known gene variants of T2DM ad the known protein coding genes of human chromosome 1 to chromosome 22. The aim is to give a in depth view of gene involvement in the two major health issues of metabolic syndrome that are T2DM ad AS. This strategy could be applied to any genetic issue to extract the interaction at genetic level, protein level and disease level. Further the topological analysis of the networks and extraction of connected components are assumed to be found by applying the applied strategy which will be helpful to give better cure or treatments of any disease.

1.9 Research methodology

The methodology that has been adopted to carry out this research work is divided into four phases from text data collection to extraction of useful patterns from that available text and then from biological network analysis to insights of biological pathways.

1.9.1 Data Retrieval through PubMed Info Extraction (PMIE)

The typical process flow of utilities in PMIE is given below:

- Get keywords from the user. Keywords may be a disease or gene/protein name.
- Connects the relevant keyword to the respective database.
- Extract records against the particular keyword from the specified database
- Generates and returns the output in the database containing titles and links of the records from the database
- Using links of articles to extract more detailed information about each article such as abstract of the paper, authors name, authors country, publication date and journal name.

1.9.2 Association Rules Generation

- Data Collection
- Identification of Hypothetical Association terms
- Pre-Processing of data
- Segmentation
- Sentence splitting Technique

- Text Documents representation
- Mining association rules and association graph generation
- Transitive association rule mining
- Nature of relationship between proteins

1.9.3 Extraction of connected components within the network for pathways crosstalk

- Algorithm for extracting strongly connected components in a directed network
- Identification of cell boundaries of each component
- Analysis of functional pathway for a disease of interest
- Analysis of components pathways
- Selection of the components involved in specific cells
- Identification of disrupted pathway

1.9.4 Topological analysis of normal and disrupted pathways

- Scanning Protein-Protein interactions
- Construction of PPIs network
- Extracting giant component from the extended network
- Topological analysis of Protein Interaction Network
- Shrinking a network
- Identification of structural holes

- Scale-Freeness topology of the network
- Detection of Hub Nodes

The main objectives of this research and the major achievements by following the given strategy are: In order to reveal the genes associated with T2DM and AS that share a noteworthy number of neighbors in the GI network completely tend to create physically associating proteins, we implied that for two genes, a and b , and the set of genes represented by $g(a)$ and $g(b)$ that are interrelated with a and b , respectively, the more prominent the number of genes that are in $g(a)$ and $g(b)$, the higher the likelihood that a and b create physically collaborating proteins. To achieve this goal we used association mining techniques of data mining. The step by step methodology is given above in the same section. The achievement obtained by applying the given strategy is gene gene interaction network for T2DM ad AS briefly explaining these two at molecular level in the form of network.

The second research objective is to give the better comprehension of hereditary genes involved in T2DM and AS at molecular level by finding the strongly connected components in the obtained network using strongly connected component algorithm of Tarjen. The strategy used for this purpose gives the 7 major strongly connected components and the cross talk between them that shows the close relation between different genes involved in different biological pathways for multiple diseases. This achievement may help to find the association of one disease with another or how a person suffering from a disease has the percentage chances to attain another disease y having the mutations in a gene that is responsible for the both diseases.

The third objective is to demonstrate the vulnerability to T2DM and AS framework, with other important known genes and how these genes influence the ordinary pathway in the organism's body to cause the particular diseases. This was done by doing the topological analysis of generated network and analyzing the topological parameters of the network. This analysis gives hub proteins diameter, bC , CC and many other such type of information that is important to understand a molecular pathway.

1.10 Summary

We introduced here the basic terminologies regarding data mining; association rules mining, genetic interactions, metabolism, metabolic disorders with their genetic facts, objectives of our research and the major steps to be followed to conduct this research work. This PhD thesis is organized into 5 chapters; where chapter 1 is the introductory chapter describing the background details of the topic chosen. Chapter 2 describes the review of literature mainly focusing the methods and technologies relevant to the field. Chapter 3 describes the detailed methodologies which have been opted to carry out this research work. Chapter 4 describes the results and their discussion and the last chapter 5 concludes the studies followed by the references and appendices section.

Chapter 2

Literature Review

Mining data is one of the area that gains a lot of practical significance and is making continuous progress at a brisk pace with innovative techniques, methodologies and conclusions in diverse relevances allied to medication, computer science, bioinformatics and stock market prophecy, weather forecasting along with text, audio and video dispensation. Data turns out to be the main anxiety in data mining. Data mining practices have been significantly subjective by the advancement of fourth generation computing languages and a variety of interrelated computing methods. In early ages of data mining, many of its algorithms are engaged only on statistical techniques. With the passage of time they get evolved with numerous computing techniques like Artificial Intelligence, Machine Learning and Pattern Reorganization. Various data mining techniques including induction, compression and approximation and algorithms have been proposed to excavate the different dimensions of varied data published in the data warehouses. The current research is to make use of gigantic and fast escalating volume of scientific prose to extract important source of knowledge that will help the researchers and scientists for going through the course of any genetic disease. This research consists of four main steps: Association extraction among genes, identification of shared functionality, if any, between and among the genes, connected components in the resulted network and topological analysis of the network and the components for the particular disease. Text mining has a elevated prospective for the knowledge

disclosure for the reason of being most common form of reporting significant research work, storing it in relevant databases, and conveying data as text. In order to accomplish our tasks in the current research, different methods and techniques have been applied. Identification of relations amongst diverse biological modules, for example, genes, proteins, diseases, drugs and chemicals, etc. is an imperative predicament for researchers particularly in biological field. Although this type of information can be extorted from various sorts of biological data, a considerable resource of such acquaintance is the biological literature which is progressively more and being made accessible as electronic databases. Automated extraction of such relationships from textual data can drastically improve biological research efficiency by keeping researchers up-to-date in their research sphere, by assisting them envisage biological pathways, and by engendering innovative hypotheses regarding new connections with the hope that there may be some good candidates for advanced biological research and substantiation. In this chapter we have presented an overview of the current research being carried out using the data mining techniques, role of connected components in large networks and topological parameters for the analysis of biological networks in order to diagnosis and prognosis of various diseases. We have highlighted critical issues and summarizing the approaches in a set of learned lessons. The goal of this chapter is to identify and evaluate the most commonly used data mining algorithms on biomedical databases for the extraction of useful information.

2.1 Motivation of current research

The concept of big data has been around for years. Big data analysis approaches can play an imperative part in health research. This can be a helpful asset for diabetes researchers because it can reveal veiled acquaintance from a massive sum of diabetes-related data. The growing availability of genome-wide expression data for T2DM has enabled a big data approach to identify molecular markers by finding robust statistical associations between genes and identifying key regulators in a large network using graph theory. There are many computational approaches for

understanding the hidden relationships among gene variants is very important and T2DM is not an exception either. These hidden patterns can also be identified using well know big data approach [1]. The area of data mining has been significantly subjective by the advancement of fourth generation computational languages and a variety of interrelated computing methods -[16]. In cells, thousands of various sorts of proteins perform as enzymes-catalysts to proceed biochemical reactions. Several proteins take part in precise roles in particular cellular compartments, whereas others shift from one compartment to another to perform specific role [44]. Usually two or more proteins bind together and form a complex to carry out their biological functions. Biological and biomedical scientists are progressively more concerned in applying high-throughput proteomics practices to accomplish an improved perceptive of crucial molecular biology and disease course of action [2].

2.2 Theory and background information

2.2.1 Association studies using data mining techniques

A lot of data mining techniques up till now have been developed and are in process of development and they are being applied to several data mining projects including association, classification, clustering, prediction, sequential patterns and decision tree. In this research, it is supposed to find the association among Genome wide Association studies (GWAS) known gene variants of disease of interest and the association of these genes with other known protein coding genes of humans chromosome 1 to chromosome 22. This chapter will mainly focus on the association part of data mining techniques. The most well known data mining technique is assoicaiton mining between different set of items in which frequent patterns are retrieved from a bulk of dataset on the basis of correlation between items in the same dataset. This technique can also be termed as relation method. Association rule mining normally includes two phase search; one is generation of frequent item sets and second is to find the strength of those generated rules by applying strength

measuring parameters. However it emerged much later than machine learning but got the better influence from the research domain of databases.

2.2.2 Connected components in biological networks

An undirected graph of having vertices and edges is called a connected graph if any two vertices of that graph have an edge between them. Assume an undirected graph with nodes a, b, c and d as given below in the Figure 1.1. Each vertex in an undirected graph acts both as a source and a destination node. There is an edge that connects node a and b. Similarly there is an edge a and c, c and d and then d and a. Possible connected components in this graph are (a, b), (a, c), (c, d), (a, d), (b, d), (a, c, d) and (a, b, d).

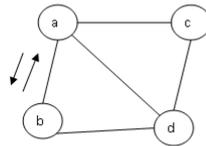


FIGURE 2.1: Connected graph in an undirected network.

A connected component in an undirected graph is defined as a subgraph that has two connected vertices via an edge and no additional vertex is connected to those two connected vertices in that given graph. Figure 1.2 shows the two examples of

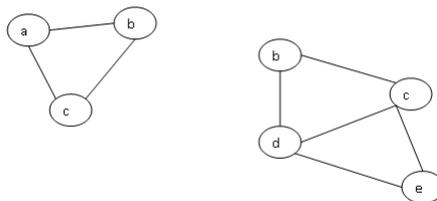


FIGURE 2.2: Connected components in an undirected network.

connected components in an undirected graph in which there is a path between any two vertices. A directed graph is supposed to be strongly connected if all

the vertices in that graph have a path or edge between them. Figure 1.3 shows a directed graph of vertices 0, 1, 2 and 3 in which each vertex has a path between them. A strongly connected component (SCC) of a directed graph is a maximal

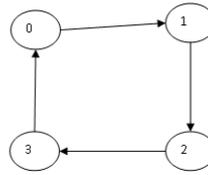


FIGURE 2.3: Connected graph in a directed network.

strongly connected sub-graphs. Figure 1.4 shows a directed network having vertices 0 to 4. There are 3 possible strongly connected subgraphs in this directed network. One SCC is 1-0-2, second SCC is 3 and third SCC is 4. There exist nu-

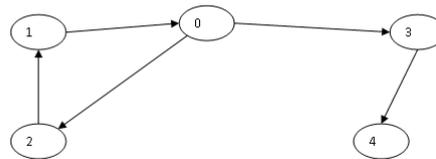


FIGURE 2.4: Strongly connected component is a directed graph.

merous algorithms that compute strongly connected components within a directed or undirected network in linear time, based on the concept of depth first search (DFS). The most common one are Kosarajus algorithm and Tarjans algorithm. Both these algorithm use the DFS strategy but the difference is that Kosarajus set of rules make use of two loops of depth first search. First loop utilize the original graph to find whether each node of the transpose of input graph has been gone through or not by applying the recursive function. This whole process in return finds a solitary new firmly associated segment or component [66]. In constrast to Kosarajus algorithm, in 1972 Tarjans strongly linked additives set of rules posted by Robert Tarjan execute only one scan of depth first search. It preserves a stack of nodes that have been explored by the search but not yet relegates to a component, and figure out "few numbers" of each node that indicates an index number

of the uppermost precursor accessible in one pace from a descendant of the vertex, which it exploits to terminate when a set of nodes should be popped off the stack into a innovative component. Similar to Kosarajus and Tarhans algorithm, there exist path based strong components algorithms that use a depth first search. They use the same concept like Tarjan's algorithm, but work with two stacks [67]. One stack tag on the nodes that are not assigned any component so far and the other maintains trail of the present path in the depth first search tree. Although Kosaraju's proposed algorithm is theoretically unsophisticated, Tarjan's and the path-based algorithm entail only one depth first search instead of two.

2.2.3 Topological analysis of genetic interaction networks

To assess the characteristics of nodes involve in a network, there exist several parameters among which Degree and betweenness centrality are particularly two key parameters in the system hypothesis [68], [69]. Degree of a node is the most elementary attribute that is characterized by how many neighboring nodes are connected to a node. Similarly betweenness centrality can be measured by calculating the overall influence of a node over the stream of information between each pair vertices in a network. However this is assumed that that information largely go through the shortest paths between them. The node which has large value of BC has astonishing impact at what streams in the system. Closeness centrality (CC) is characterized by the normal's converse length of the most limited ways to/from the various nodes in the network, which lets us know the topological focus of the system. Diameter (D): is the longest among all shortest paths. Numerous in silico techniques have been accomplished for human gene prioritization and a number of review articles describe how these approaches work out, their differences with existing techniques in the same area [70], [71], [72], [73], [74]. These techniques show fluctuation in their inputs, outputs and their prioritization approach.

2.3 Contemporary research work

2.3.1 Data mining techniques

Rakesh Agrawal, in 1994, [75] proposed the first association mining technique with the aim of finding correlations in the objects of a market basket database. That algorithm proposed by Rakesh Agrawal in 1994 was named as Apriori. Although, association rule mining was first introduced as a market basket analysis tool, it has since become one of the most important paraphernalia for performing unendorsed exploratory data analysis over a broad range of research and commercial areas, including biology and bioinformatics.

Diti Gupta *et al.*, in 2013, [76] discussed how significant association rules can be extracted from a collected dataset. They characterized the different parameter to determine the strength of association rules. The most important ones are support value and confidence value. They also discussed how to select the threshold values for the extracted association rules. An association rule can be characterized as $X \rightarrow Y$, along with support and confidence value (S, C). X and Y are item set; S represent the support of the generated rules, calculated as the number of the transactions that contain all items in X and all items in Y and C is the certainty, it is characterized as the proportion of S with the rate of transactions containing X. Support and certainty are measures of the intriguing quality of the extracted rules. They have computed the support an incentive for supporting the value of the items introduces in the informational collection. Higher support esteem shows the viability for the endeavor. Negative affiliation rules of form $X \rightarrow \tilde{Y}$ implies support $(X \cup \tilde{Y})$ greater and equals to minimum support $(XUY) = \text{support}(X) - \text{support}(XU\tilde{Y})$. For most transactions, the $\text{support}(X) < 2 * \text{minimum support value}$ implies XUY is occasional itemsets. To discover negative affiliation rules, prompts find rare itemsets first. The support tally demonstrates the recurrence of the patterns in the rule; it is the level of exchanges that contain both X and Y. Certainty is the quality of ramifications of a rule; it is the level of exchanges that contain Y on the off chance that they contain X.

Karthikeyan *et al.*, in 2014, [77] suggested the concept that to measure the strength of association rules, there exist many parameters but the two most significant and commonly used measures are support and confidence. As the database contains massive collection of data to any specific field so the researchers mainly apprehension about only the repeatedly item sets. The users can set a threshold value termed as minimal support and minimal confidence as mentioned in the previous paper to slump the rules that are not much functional or important.

Li Xiaohui *et al.*, in 2012, [78] presented out an improved version of Apriori algorithm by overcoming the defects of old version on the basis of the analysis in terms of its efficiency. Their proposed model proved to be better in terms of time complexity and minimize the input output operation of the mining processing by reducing the times of probing in the catalog of articles. It is proved in the experimental outcome that the enhanced version is many times proficient than the conventional version.

Temkin and Gilder, in 2003, [79] utilized a full parser with a lexical analyzer and a context free grammar to extort relationship between two proteins [80].

Yakushiji *et al.*, in 2005, [81], anticipated a protein-protein association mining system based on head-driven expression structure language. Despite the fact that the pattern generation is convoluted, the execution is not adequate. Also, reliance sentence structure is utilized every now and again in this space.

Erkan *et al.*, in 2007, [82] projected a semi-supervised classification for mining protein interaction sentences via reliance parsing.

Fundel *et al.*, in 2007, [83] characterized few rules on the basis of dependency parse tree for relationship extraction. The issue of the frameworks that utilize dependency parse is that they cannot extravagance non-local dependencies, and thus rules obtained from the development are incomplete. Differently, in our effort, sentence filtering construct system is applied on themes and phrase structure parsing for relation withdrawal. The extorted sentences conceivably restrain information that how mutations in specific genes are responsible for specific disease. Phrase

structure grammars depend on the constituency relation, rather than the dependency relation linked with dependency grammars. Phrase structure parsing is full parsing, which considers the full sentences to be parsed. Furthermore, numerous researches utilized an preliminary list of seed genes to construct a disease-specific gene-interaction network system, and thus they are influenced in favor of the seed genes, therefore the consequences also rely on the pickup seed genes [78],[84], [83], [85], [86], [87].

Ji Hoon *et al.*, in 2012, [88] adopted a text mining technique to find the useful and informative association rules that represents the relationship between diseases and herbals used in Oriental medications. Along with enlightening the functional rules generation technique, they also performed network analysis on the resulted association network in order to find the degree, betweenness centrality, closeness centrality and other network parameters. The deduced rules are filtered using the strength measuring parameters frequently used in many association rule mining techniques.

Sajid *et al.*, in 2014, [89] proposed an algorithm for finding positive and negative affiliation rules among frequent and rare itemsets. They recognized relationship among medicines, side effects, and research center outcomes utilizing best in class information mining innovation. They had proposed a novel technique that catches the negative relationship among frequent itemsets and furthermore separates the positive relationship among the rare itemsets. In any case, their examination does not have the idea of nature of relationship among itemsets.

Atif *et al.*, in 2014, [90] extorted affiliation principles to limit the impacts of Dengue by utilizing content mining strategies. Utilizing distinctive information sources, they played out some preprocessing strategies, for example, change, filtration, stemming and ordering of the records and after that connected information mining procedures to build up a framework that would not just recognizes land spreading examples of the infections however it would likewise proposes proactively next geological area where infection has most likelihood to append with the goal that legislature can take cure measures.

Changqin Q. and Fuji R. in 2014, [91] analyze the gene disease extorted network by utilizing information sorting, parsing and network analysis techniques. Their proposed models work in two phases; firstly extracting the potential sentences that have information related to specific gene or disease of interest for which one is interested to construct an association network. For this purpose they used maximum entropy classifier with topic features. For parsing, they exploited Probabilistic ContextFree Grammars on the extorted sentences for gene-disease association extraction. Network analysis parameters were used to find the significance of each gene involved in the association network. To evaluate the proposed model breast cancer was used as testing disease and the reported genes for this disease. The 31 top ranked genes and diseases had been found to show the relevance with breast cancer through NCBI database with the accurarcy of 83.9%.

Much exertion is at present spent on extricating gene disease affiliations [78], [92]. Biomedical connection extraction procedures essentially incorporate two branches: cooperation database based techniques and content mining strategies. Communication database construct strategies depend in light of the accessibility of interaction databases, for example, OMIM, MINT [93], IntAct [56], BIND [57], which anticipate associations between entities utilizing succession, auxiliary, or transformative data [58]. In spite of the fact that biomedical databases are expanding day by day and to curate these databases manually requires significant exertion and time, still there is a vast collection of manually extracted associations from these databases.

Content mining ways to deal with gene disease association extraction have demonstrated a development from straightforward frameworks that depend exclusively on co-event insights [59], [60] to complex frameworks using normal dialect handling procedures and machine learning calculations [61], [62]. Well understood instruments for finding genedisease association incorporate DAVID [63], GSEA [64], GOToolBox [65], rcNet [94] and numerous others. In any case, much of the time, since the current explanations of ailment causative genes is a long way from

completion [95], and a set of gene may just contain a short rundown of inadequately clarified genes, accessible methodologies frequently neglect to uncover the relationship between gene sets and sickness phenotypes [96].

Network based methodologies [97], [84] are presented by evaluating how genes associate in protein systems and mutation in any point of the protein network is responsible to cause a specific disease. Recently, syntactic investigation has been well thought-out for such relation mining, and distinctive parsing language structures have been connected.

2.3.2 Related techniques on extracting strongly connected components in a graph

Depth First Search algorithm has been used to traverse tree or graph data structures and is the building block to find Strongly Connected Component. DFS selects some arbitrary node as the root and investigates beyond what many would consider the probability for each branch prior to backtracking. Strongly associated parts can be figured utilizing diverse methodologies presented by Tarjan, Gabow and Kosaraju and so forth. Tarjan's and Gabow strategies require just a single DFS, yet Kosaraju's calculation requires two DFS [7]. It impels from the foundation of the graph, investigate its first tyke, investigate the child of next vertex until reach to the objective vertex or reach to definite vertex having no further tyke. By then, backtracking is used to give back the last vertex which isn't yet completely investigated. Changing the post-visit and pre-visit, DFS is utilized as a solution of numerous imperative issues and it takes $O(|V|+|E|)$ steps [7].

Tarjan, in 1972, [6], [7], [98], [99] presented strongly connected components algorithm that acknowledge directed graph as an input and returns all possible components in that graph by exploring all nodes of the graph utilizing DFS strategy. The algorithm embarks on from a random start node and mutely disregards the nodes that have been visited already. The nodes put in a stack in the way in which they investigated and maintain record number (node.index). Also every

node holds a lowlink which is constantly subordinate than the current node file. The current node is the root hub of firmly associated segments if $\text{node.lowlink} = \text{node.index}$. A subgraph is returned on the off chance that it is enduring that it is an unequivocally associated segment.

Kosaraju's in 1978, utilizes two scans of DFS on the transpose graph of the input graph, and each iterative scan finds a solitary a firmly associated components.

Gabow, in 2000, [6], [7], [99], [100] proposed strongly associated subgraphs strategy that resembles with Tarjan's strategy for the same task. It acknowledges a graph that is directed as an input and output containing accumulation of all conceivable firmly associated segments. It likewise utilizes DFS to investigate each vertex of the input graph. Gabow strategy keeps up two stacks, one stack have a rundown of nodes not yet figured as firmly associated parts and second one contains an arrangement of nodes not have a place with various emphatically associated segments. A counter is utilized to check number of nodes that had been gone through, which is utilized to process preorder of the nodes.

Swati Dhingra *et al.*, in 2016, [101] proposed a strongly connected components detection algorithm (SCCD) but these components contains low number of nodes. This proposed algorithm has a strong impact on social networking service to construct social networks involved in the formation of communities. These social networking services can identify how these communities evolve and let us know what community a persons belongs to in order to get relatively better ads to target their essentials.

Surender Baswana *et al.*, in 2017, [102] concentrated on the issue of keeping up the strongly associated components of a graph within the sight of disappointments. The issue of keeping up the SCCs of a cgraph was examined in the decremental demonstration. Specifically, they demonstrated that given a directed graph $G = (V, E)$ with $n = |V|$ and $m = |E|$, and a whole number esteem $k \geq 1$, there is a strategy that calculates in $O(2kn \log^2 n)$ time for any set F of size at most k the emphatically associated segments of the graph $G \setminus F$. The running time of their model is relatively ideal since the ideal opportunity for yielding the SCCs of $G \setminus F$

is minimum $\omega(n)$. They utilized an information structure that is registered in a preprocessing stage in polynomial time and is of size $O(2kn^2)$. In their model the objective is to keep up the SCCs of a graph whose edges are being eliminated by an adversary. The primary parameters in this model are the most pessimistic scenario update time per an edge deletion and the aggregate update from the first edge deletion until the last.

Geldenhuis and Valmari, in 2005, [7] modified Tarjans algorithm for the purpose of LTL model checking. The idea of the algorithm is that the last found accepting state of the current search path, is kept track of. Whenever a back-edge is found, which points to a previously visited state from the current search path (which updates the lowlink value), the algorithm terminates with an accepting cycle.

Dijkstra, in 1982, [98] proposed a different variation to Tarjans algorithm. Instead of keeping track of lowlink values, this algorithm maintains a stack of possible root candidates. On finding a back-edge, the algorithm pops vertices from the stack until the root of the cycle is found. At backtracking, the current flag of reachable states are set to false so that these states do not interfere with a future search. This algorithm also runs in linear time, $O(|V| + |E|)$ with $O(|V|)$.

Couvreur, in 1999, [103] designed a variant on Dijkstras algorithm for the purpose of finding accepting cycles. The main difference with a Tarjan-based accepting cycle algorithm is that here, information on partial SCCs is propagated when finding a back-edge.

2.3.3 Topological analysis of genetic networks

There have been various fruitful investigations that have discovered vital target genes and markers by developing and breaking down the protein interaction systems related with infections [104]. .

Lee *et al.*, in 2011, [105] developed PPI systems of genes that expressed abnormally for schizophrenia, bipolar illness and significant discouragement, and recognized

a few ailment markers, for example, strawberry indent homolog 2 (*Drosophila*) (SBNO2) for schizophrenia, SEC24 homolog C, COPII coat complex segment (SEC24C) for bipolar confusion and serrate, RNA effector particle (SRRT) for major depression.

Ran *et al.*, in April 2013, [106] built and analyze PPI systems for essential hypertension (EH), and proposed that pulse variations in association with EH are organized by an incorporated PPI system with the protein encoded by the nitric oxide synthase 3 (endothelial cell) (NOS3).

Rakshit *et al.*, in 2014, [107] built PPI systems on the basis of gene articulation profiles of Parkinson's malady (PD) and distinguished 37 biomarkers that can be utilized as potential remedial focuses for PD applications advancements.

Wenhai Xie *et al.*, in 2014, [108] built a protein protein collaboration (PPI) system for Spermatogonial sperm cells (SSC) self-restoration in view of connections between 23 genes for SSC self-reestablishment, which were acquired from a content mining framework, and the interfacing accomplices of the 23 key genes, which were differentially communicated in SSCs. The SSC self-reestablishment PPI network comprised of 246 nodes associated by 844 edges. Topological examinations of the PPI system were led to distinguish genes fundamental for support of SSC self-reestablishment. The subnetwork of the SSC self-recharging system proposed that the 23 enter qualities associated with SSC self-reestablishment were associated together through other 94 genes. Bunching of the entire system and subnetwork of SSC self-reestablishment uncovered a few densely associated areas, inferring huge molecular interaction modules fundamental for SSC self-restoration.

Chen *et al.*, in 2016, [109] mined the differentially expressed proteins (DEPs) affirmed more than 2 times in the proteomics composing of Prostate Cancer (PCa). The DEPs were seen as seed proteins to build up a widened PPI system, which involved the seed proteins, and in addition their direct PPI neighbors and the joint efforts between these proteins. Topological examinations were performed to choose the basic framework biomarkers. The relationship of these biomarkers with the start of PCa was inquired about. The backbone network created of key nodes and

the subnetwork of the briefest routes, and also the thickly related territory, were in like manner inquired about. Along these, the revelations of this examination may give understanding into the potential concentrations for making novel treatment systems for PCa.

Soudabeh Sabetian and Mohd Shahir Shamsir, in 2016, [110] exhibited the principal protein-protein interaction system identified with azoospermia and investigate the perplexing impacts of the related genes efficiently. They made a novel system comprising of 209 protein and 737 collaborations, recognized three proteins as hubs and a bottleneck protein inside the system and furthermore distinguished new hopeful genes which may assume a part in azoospermia. The gene ontology examination proposes a hereditary connection amongst azoospermia and liver sickness, colorectal, pancreatic, incessant myeloid leukemia and prostate malignancy.

Safaei *et al.*, in 2006 [111] evaluated the biological characteristics of 13 identified proteins of patients with cirrhotic liver disease and their results pointed out that regulation of lipid metabolism and cell survival are important biological processes involved in cirrhosis disease.

2.3.4 Limitations and bottlenecks of existing approaches

In bioinformatics there is a need to extricate data from a plenty of writings. For instance, how proteins are associated with each other is vital information that is utilized as a part of the improvement of molecular pathways. A few systems that have been talked about in this section have been executed as of late to attempt and achieve this undertaking. A significant number of these strategies are computationally concentrated and their applications to on-line investigation of a substantial arrangement of recovered reports will require critical holding up time with respect to clients. Besides, not very many of them endeavor to aid the content examination with area learning, accessible from legitimate associations, specialists, or clients themselves. The existing tools used for the purpose of association extraction between genes are just limited to the relationship between genes. A Thesaurus

based text analysis approach and tool to discover the existence and the functional nature of relationships between genes relating to a problem domain of interest also need to be focus for better understanding of molecular pathways. Similarly the cutting edge technology of system biology to get insight from large networks to pathways and the topological analysis of these pathways are also missing in the above mentioned technologies.

2.4 Current research vs. existing work

A bunch of examinations are available in the exposition that are identified in finding intriguing relations with respect to a particular expected variable in a given data set. This part of information mining has become much pondering among the specialists in modern era. Keeping the bottleneck points mentioned in previous section, we progressed in that direction. Along with data mining techniques proposed by [75],[76], [77], [78],[79], [83], [82], [85], [88], [89], [90] and many others that have been discussed in above section, several other techniques have also been reviewed [112], [113], [114], [115], [116], [117]. The existing tools used for the purpose of association extraction between genes are just limited to the relationship between genes. A Thesaurus based text analysis approach and tool to discover the existence and the functional nature of relationships between genes relating to a problem domain of interest also need to be focus for better understanding of molecular pathways. Similarly there exists numerous algorithms that compute strongly connected components within a directed or undirected network in linear time, based on the concept of depth first search (DFS) [7], [98], [99], [100], [101], [102], [105], [106], [118], [119], [120],[121]. The techniques applied so far mainly focused in computer science issues and there is still need for the examination in information mining requires more consideration in biological and ecological data investigation for better understanding of molecular connectivity ad pathways analysis and their crosstalk. To assess the characteristics of nodes involve in a network, there exist several parameters [104],[105], [106], [107], [108], [109] among which Degree and betweenness centrality are particularly two key parameters in the

system hypothesis. The cutting edge technology of system biology to get insight from large networks to pathways and then topological analysis of these pathways is missing in the above mentioned technologies. We in our experiments tried to address those questions using different approaches including association extraction, transitive associations in network, nature of relationship between two genes, topological analysis of association network and other graph theory approaches, which the existing methodologies were lacking. In the current research we have merged the techniques that exists independently for finding association between two genes upto analysis for the resulted association network topology. The most commonly used association mining rules such as confidence, support, lift, conviction has been used for the task of discovering associations among genes. The lexicon that contains all the promising relationship terms showing the associations among gene terms have been applied to all extorted sentences from the abstracts to determine the nature of relationship between two genes. Given a gene/protein interaction network, finding strongly connected components has been done by examining the elementary constituents individually, by using models and methods from the graph theory universe. In this research, we have approached the topological analysis of the gene/gene interaction network and the components that have been found to be strongly connected in the obtained network. For the topological analysis, Cytoscape and Pajek tool has been used to measure the basic characteristics of nodes in the association network. Meanwhile, a simple web based interface PubMedInfo Crawler (PMIC) has been developed to provide data extraction utilities for commonly used database PubMed. Although these technologies are not new to anyone and many previous researchers used these techniques but the pipeline followed for this research in unique to selected disease as far as our knowledge is concerned.

2.5 Summary

In this chapter we discussed different studies related to our area of interest and have looked at their limitations. We also concluded on how to address the some left over questions in the field. After thorough analysis of competing algorithms, it is

very intricate to name a solo mining method to be most suitable for the prognosis of diseases. At times some of them perform better than others, but there are cases when a combination of the best properties of some of the aforementioned algorithms together results more effective which we have done in the current research. Genetic interactions and their properties along with graph theory approaches for pathways analysis have been focused in this study.

Chapter 3

Methodology details

3.1 Introduction

Genetic interactions and their topological properties along with molecular pathways are retrieved so far via various computational and wetlab approaches as discussed in chapter 2. We based our studies on Type II diabetes mellitus and identified the association patterns among known gene variants for T2DM and the protein coding genes of human chromosome 1 to chromosome 22, their transitive (indirect) relationships and what type of relation exists between them using data mining association techniques. The current research start from the collection of dataset related to T2DM from PubMed. For this purpose a screen scrapping algorithm was developed which was implemented in php. After collecting data, tf-idf algorithm was applied to evaluate the frequent of the gene variants supposed to be susceptible for T2DM. The strength of relationship between those variants was checked using commonly used association mining parameters named as confidence, support and lift value. After generating an association network, we proposed a simple computational approach to extract the connected components from the large biological network using appropriate technique of graph theory using MATLAB code. Along with this, we applied bioinformatics tools for the topological

analysis of obtained patterns. Figure 3.1 represents the architecture model of current research.

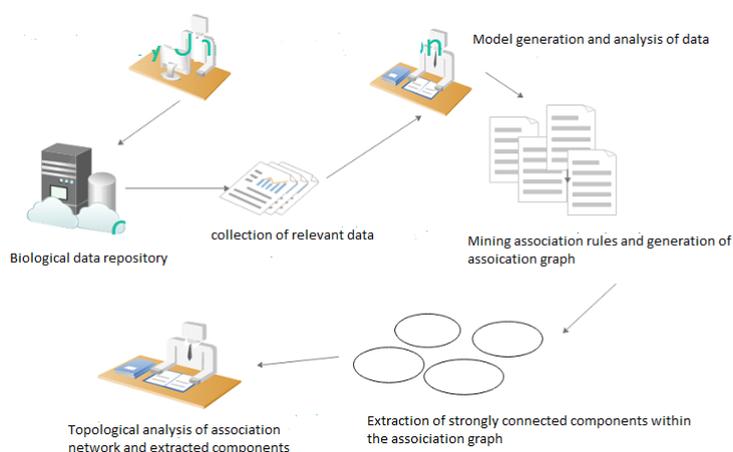


FIGURE 3.1: Architecture Model of current research.

”Graphical model of the proposed research from the text data collection from biological database PubMed to the generation of association rules from the collected data and then extraction of strongly connected components to the topological analysis of generated network”.

3.2 Tools and equipment

3.2.1 Hardware specifications

The system used in the research was Hp with 4.00 GB RAM and Intel(R) core(TM) i3-3217U CPU. The processor specifications were 1.80GHz.

3.2.2 Software(s), tool(s) used

3.2.2.1 Windows platform

Windows 8 with 64 bits Operating system, x64-based processor has been used for experimental work.

3.2.2.2 Language used in project

Language used In This project: HTML, PHP, CSS Need an Apache server for local system run: In windows (xampp, Wamp, local-server) Framework Used: Bootstrap MATLAB: MATLAB version R2015a has been used for the purpose of finding connected component within the generated association network

3.2.2.3 Biological database

- PubMed: The biomedical domain was chosen for performing the experiments on the proposed model. The data was from the PUBMED database of biomedical articles, the one extensively used in biomedical sciences [122].
- GWAS Catalog: The Catalog was founded by the NHGRI in 2008, in response to the rapid increase in the number of published genome-wide association studies (GWAS). These studies provide an unprecedented opportunity to investigate the impact of common variants on complex disease. The GWAS Catalog gives a steady, accessible, visualisable and unreservedly accessible database of distributed SNP-trait affiliations, which can be effectively coordinated with different assets, and is gotten to by researchers, clinicians and different clients around the world [123].
- Human Atlas Protein: The Human Protein Atlas (HPA) is a Swedish-based program began in 2003 with the intend to guide of all the human proteins in cells, tissues and organs utilizing mixture of different omics advances, including antibody based imaging, mass spectrometry based proteomics, transcriptomics and system biology. All of the information in the learning asset is open access to permit researchers both in the scholarly community and industry to uninhibitedly get to the information for investigation of the human proteome [124].
- Polysearch Engine: PolySearch is a web based mining tool that mine genes, proteins, drugs, SNPs related to a specific disease by using biological information sources like PubMed OMIM, Drugbank etc [125].

- OMIM: Online Mendelian Inheritance in Man is a constantly refreshed index of human genes and hereditary issue and characteristics, with a specific spotlight on the gene-phenotype relationship [126].

DataTypesCaptured : Genes, genetic disorders, phenotypic traits

Description : Catalog of all known human genes and genetic phenotypes

3.2.3 Bioinformatics tools

- Pajek:Pajek (Slovene word for Spider) is a program, for Windows (32 bit), for investigation of vast systems. It is openly accessible, for noncommercial use, at its landing page: "http://vlado.fmf.uni-lj.si/bar/systems/pajek". Pajek adaptation 5.02a has been utilized for the topological examination of hereditary system.
- Cytoscape: Cytoscape; Network information integration, investigation and representation instrument (version 3.6.1) has been utilized for the examination of hereditary segments. Cytoscape is an open source bioinformatics programming stage for picturing molecular interaction systems and incorporating with gene articulation profiles and other state information. Extra highlights are accessible as modules.

The methodology details of each section is discussed below:

3.3 Data collection from biological database

The typical process flow of utilities in PMIE is given below, represented graphically in Figure 3.2.

- Get keywords from the user. Keywords may be a disease or gene/protein name.
- Connects the relevant keyword to the respective database.

- Extract records against the particular keyword from the specified database
- Generates and returns the output in the database containing titles and links of the records from the database
- Using links of articles to extract more detailed information about each article such as abstract of the paper, authors name, authors country, publication date and journal name.

3.3.1 System requirements

Minimum 32bit Operating system Operating system: Windows xp, 7, 8, 8.1, 10, Macintosh, Linux, Unix etc, Need an Apache server for local system run: In windows (xampp, Wamp, local-server) for Macintosh (Mamp ,local-server) for all other operating system (Lamp). Language used In This project: HTML, PHP, CSS Framework Used: Bootstrap Output: Output was stored in the local disk Drive → Xampp → htdocs → crawler → output

3.3.2 Getting keywords

The PMIE provides the opportunity to extract articles from PubMed by using disease or gene/protein name and through year wise. To proficiently distinguish and extricate the most recently added data from the PubMed makes the extraction process much more productive, because it must extricate a considerably littler volume of information of the given year.

3.3.3 Connecting to PubMed

After getting disease or gene name from the user the request was sent to eu-tils.ncbi.nlm.nih.gov , to fetch results from this server and then generate html file to display output. The output contains the links to articles and the URLs against entered disease or gene name.

3.3.4 Extracting articles details

By using the keywords as query terms in PUBMED, we have saved all the papers links that are returned to us as a result of these queries. From each paper, the following information was extracted; Pubmed id of each document, title, abstract, authors names and year of publication. All this work was done by screen scrapping methodology, a code was written in PHP platform to extract all the required information from the PUBMED database against each query term. The retrieved information was saved in different formats like SQL, CSV or excels, PDF etc to carry out further processing.

3.4 Association rules mining

The conceptual model of proposed methodology for the second phase of this research study is represented in Figure 3.3.

3.4.1 Data collection

Broad endeavors were made to distinguish articles utilizing techniques on disease of interest. The biomedical domain was chosen for performing the experiments on the proposed model. The data was from the PUBMED database of biomedical articles, the one extensively used in biomedical sciences. To be more accurate in finding all related articles, the search were performed in PUBMED based on all the possible queries of particular disease. We used the automatic screen scrapping methodology (PubMedInfo crawler) to extract information related to articles obtained in return to queries. The abstracts of all the articles obtained in return to these queries were acquired from PubMed. The obtained data was transformed into .xml format.

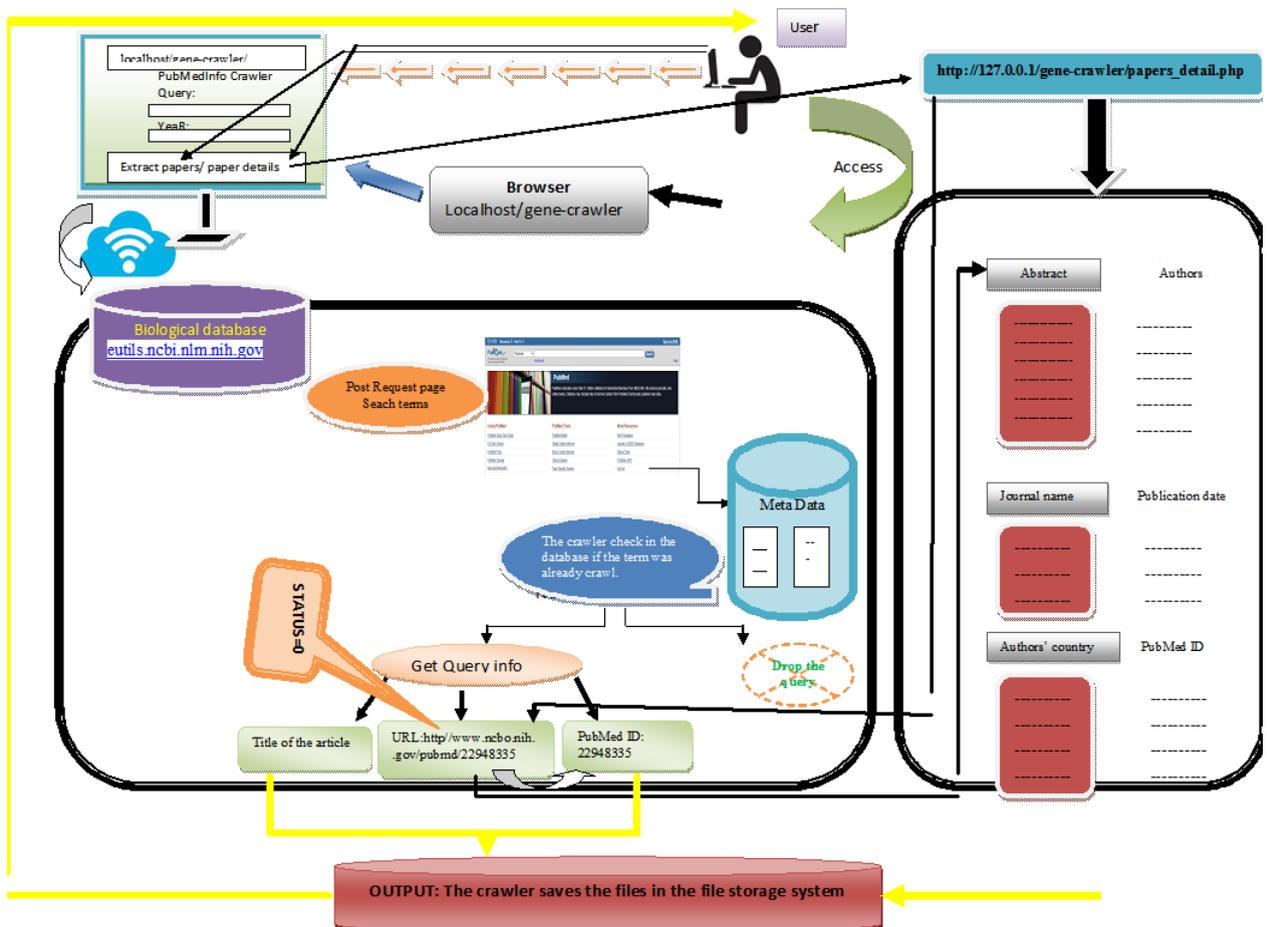


FIGURE 3.2: Architecture Model of PubMed Information Extractor

”Typical flow process of *PubMedInfoExtractor*. The system works by getting keywords (disease, gene/protein name) from the user along with year for which user is interested to extract data for. After sending request to respective database, the PMIE extract details, such as abstract, title, journal name, author’s affiliation etc are stored in the database at the back end, which can be used later for further processing.”

3.4.2 Identification of hypothetical association terms

In this research, the association among GWAS known gene variants of disease of interest and the association of these genes with other known protein coding genes of humans chromosome 1 to chromosome 22 have been retrieved from the biological literature.. A detailed literature search was conducted to get the gene names of that protein coding genes. Over a moderately brief time, GWAS have altered the scene of hereditary qualities. In spite of the fact that the larger part of variations

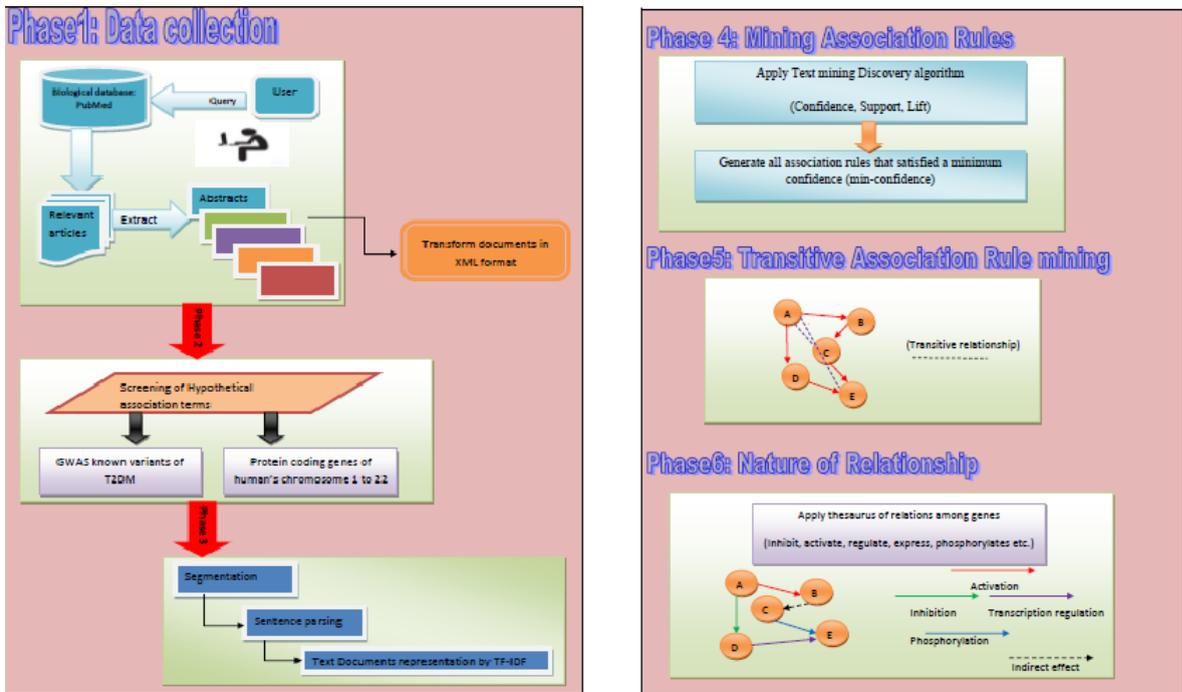


FIGURE 3.3: Conceptual model of mining association rules

found so far have been basic with little impacts, we have attempted to reveal numerous new loci, prompting energizing bits of knowledge into the hereditary architecture of specific malady.

3.4.3 Processing phase: segmentation

Segmentation technique was applied for selecting relevant articles. This was done by querying a gene variant identified by GWAS for disease of interest. Those articles that contain gene variant against disease and the hypothetical gene terms were splitted from the entire dataset of papers obtained from PubMed. The extracted articles potentially contain information about GWAS gene variant and other humans protein coding gene from chromosome 1 to 22.

3.4.4 Processing phase: sentence splitting technique

A splitta technique was applied for segmenting articles into sentences which include high accuracy sentence boundary detection. The sentences of each segmented article are splitted by ”.”. We assume that each splitted sentence should contain at least one target gene and hypothetical gene term. As mentioned above, we built the gene name dictionary from OMIM database and GWAS. All gene names are considered as keywords.

3.4.5 Processing phase: text documents representation

Data about the topic of a document is given by the quantity of occurrences of a term. The most normally utilized approach is the term frequency inverse document frequency (tf-idf) model. In this model, having a gathering of archives, the significance of a word in specific report increments with the quantity of time that this word shows up in this record yet is balanced with the circumstances that show up in each of the documents of the collection.

$$\text{Tf-idf}(w) = \text{tf} * \log(N/\text{df}(w))$$

Tf = frequency of term in a document

idf = inverse document frequency: number of documents that contain specific term

N = Number of all documents

In other words, terms that appear many times in a document but also appear many times in the collection have a lower score. This is very interesting to penalize common words that appear very often in all documents and which are not part of the main topic.

3.4.6 Mining association rules

In general, association rules are extracted from a given dataset in two stages. First, a set of frequent rules is generated that shows association between two objects or terms. Second, the strength of the rules, obtained from the first stage is evaluated using support, confidence values of those generated rules and minsup and minconfidence threshold value. An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item that is found in the data, the term for which appropriate consequents are required to be found from the dataset. On the other hand, a consequent is an item that is found in combination with the antecedent. However there are many objective interestingness measures for determining association rules. The measures include support, confidence, coverage, prevalence, recall, accuracy, specificity, lift, leverage, added value, relative risk, jaccard, certainty factor, odds ratio, Laplace correction, cosine, Yule's Q, Yule's Y, Piatetsky-Shapiro etc. Still the most common and important measures are support, confidence and lift. Support is an important measure because a rule that has very low support may occur simply by chance. A low support is also likely to be uninteresting from a scientific perspective in case of biological objects because it may not be playing important role to promote disease that seldom occur with that term. For these reasons, support is often used to eliminate uninteresting rules. Support also has a desirable property that can be exploited for the efficient discovery of association rules. Confidence, on the other hand, measures the reliability of the inference made by a rule. For a given rule ($A \rightarrow B$), the higher the confidence, the most likely it is for B to be present in association with "A". Confidence also provides an estimate of the conditional probability of "B" given "A".

3.4.7 Transitive association rule mining

Association rules show the ways in which things can stand with regard to one another or to themselves. Relationship between two things can be symmetric or

asymmetric. Relation R is transitive if in an association network the relation R of (a, b) and relation R of (b, c) imply relation of (a, c) . In symbols, R is transitive if and only if $((R(a, b) \text{ and } R(b, c)) \rightarrow R(a, c)) / =$. In other words, two nodes A and C in a graph G^* can be associated transitively if there exist any edge between these two nodes in a graph G . In biology or real life examples, transitive relations may follow the following pattern: (a) if in a pathway there is a gene A that has regulatory affect on a gene B and on regulation of gene B , inhibition of another gene C occurs then according to transitive closure property, gene A has transitive relation with gene C .

3.4.8 Nature of relationship

Once association rules and association network has been generated between genes/proteins using those rules, the next very important step to make that network more significant is to find out what is the nature of relationship between two genes/proteins in that PPI network. This will assist to expand prospective pathways and pace up the process of discovering genetic interactions. To accomplish this task we acquired thesaurus of terms from the literature that have all possible relationships between genes. After having the thesaurus, we applied it to the parsed sentences, which contain co-occurring genes. If a sentence that has co-occurrences of genes matches a relationship in the thesaurus, it is assigned a score. The more score for a relationship over all sentences would be more likely to indicate that relationship between the two genes/proteins. A score of as little as one could be significant because a relationship may be mentioned in one abstract. If there occurs no functional relationship between two genes/proteins, we cross checked through those sentences where the terms co-occurred to see if a function might have been omitted from the generated thesaurus containing the relationships. The list of potential relations used in this research work is shown in table 3.1.

TABLE 3.1: Relationship thesaurus used in the current research

Acetylates, acetylation; Activates, activation, activator; Binds, binding, complexes, Represses repression; Inhibits, inhibition, inhibitor; Phosphorylate, phosphorylation, Catalyses, catalytic, catalyst; Regulates, regulation; Induces, induction; Cleaves, Cleavage, Creates, creations; Suppresses, suppression; Releases, releasing; Hydrolyzes, hydrolysis, hydroxylation; Adhesion; Transports, exports; Donates, donation; Suppresses, suppression

3.5 Extraction of strongly connected components within a network

The methodology adopted to extract the components that strongly connected is represented graphically in Figure 3.4.

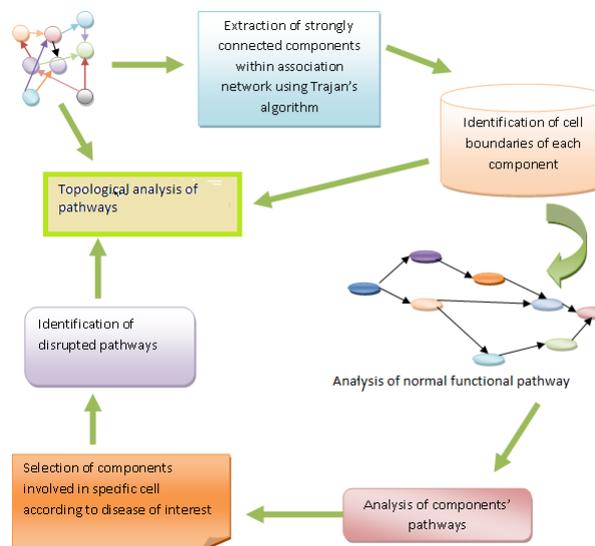


FIGURE 3.4: Flow diagram for extraction of strongly connected components
 ”Connected components are extracted by using Tarjan’s algorithm in MATLAB. Human Protein Atlas is used for the identification of cell boundaries and sub-cellular locations. Functional pathway of T2DM was analyzed for determining the disrupted pathway along with pathway crosstalk from the network generated by association rules.”

3.5.1 Extraction of connected components within association graph

Algorithm for strongly connected components in a directed graph 3.2.

TABLE 3.2: Algorithms for extracting strongly connected components

```

Connected components (G)
for each vertex  $V \in G.V$ 
  make set of (V)
for each edge  $(u,v) \in G.E$ 
  if Find-set(u)  $\neq$  Find set (v)
  then
    union (u,v)
  same component (u,v)
  then
    if Find-set(u)  $==$  Find set (v)
    then
      return true
    else
      return false

```

Given a directed graph showing the association between two gene and among genes, we used MATLAB syntax of $[S, C] = \text{graphconncomp}(G)$ that finds the strongly connected components within the graph represented by matrix G using Tarjans algorithm. Input is the text file containing two columns that shows association between two vertices. An emphatically associated part in a graph is a maximal gathering of nodes that are commonly reachable without disregarding the edge bearings. Input G is a N -by- N scanty lattice that represents a graph. Non-zero sections in grid G show the nearness of an edge. The quantity of parts found is returned in S and C is a vector showing to which segment every node of a graph

belongs to. The steps for finding strongly connected components for directed graph are as follows:

```
DG = sparse (Filename (:, 1), Filename (:, 2), true, No. of nodes, No of nodes)
```

```
DG = sparse (examplediabetes (:, 1), examplediabetes (:, 2), true, 181, 181) //  
create a directed graph with 181 nodes.
```

```
h = view (biograph(DG)); // viewing a directed graph using the above mentioned  
file.
```

```
[S, C] = graphconncomp(DG) // Find the number of strongly connected compo-  
nents in the directed graph and determine to which component each of the nodes  
belongs.
```

```
colors = jet(S);
```

```
for i = 1: numel(h.nodes)
```

```
h.nodes(i).color = colors(c(i), :);
```

```
end // Color the nodes for each component with a different color.
```

3.5.2 Identification of cell boundaries of components in a network

To systematically explore and visualize large and complex networks, one solution is to such an interaction network into components of densely interacting nodes and implies functional modules. The protein expression of each protein of extracted components has been identified from Human Atlas Protein Database. Each protein shows protein expression in different cells of human organs and has different sub-cellular locations. Similarly some proteins are secreted into the cells, some are membrane bounded and some are intracellular. On the basis of the information from Human Atlas Protein Database, the cell boundaries of each protein of resulted components have been identified.

3.5.3 Analysis of functional pathway of genes

Single gene scrutiny is inadequate to portray the intricate perturbation conscientious for a disease. A deeper perceptive of the system of meticulous ailment can be attained by spotlighting the deregulation of gene sets or pathways rather than individual genes. In this innovative perception, the focal point moves to recognize biological processes, cellular functions and pathway agitated in the phenotypic circumstance by exploring the genes in a specified pathway taking into account the potential relations among them and the association of their expression with the phenotypical state of affairs. We analyzed the functional pathway for the genes that are involved in the disease of interest. The pathway was gone through in detail to identify how functional relationship among particular genes for disease of interest is disrupted to cause a specific disease.

3.5.4 Analysis of components pathways

In order to identify the crosstalk between pathways, we analyzed different pathways in which genes of strongly connected components are involved. Pathways were derived from the KEGG database. We followed KEGG database for the purpose. KEGG is an accumulation of databases managing genomes, natural pathways, sicknesses, medications and concoction substances.

3.5.5 Selection of the components involved in specific cells

In the previous step, the protein expression of each gene in a component in different cells of human organs has been revealed. We have also identified the sub-cellular locations of each gene. From the resulted components we have selected only those components that show highly expressed genes in a particular cell according to disease of interest.

3.5.6 Identification of disrupted pathways

Biological pathways are response chains where concoction items turn into the substrate for the subsequent stage. All substrates are synthetically changed in responses that have a place with either pathway. We have tried to explore potential functional biomarkers by utilizing gene pathway crosstalk. The genes in components show the nature of association between the genes, whether one gene inhibit or activate the expression of its associated gene. From this information the selected components whose genes are secreted or membrane bounded to specific cells, we have identified how two pathways are connected with each other and disrupt the normal functioning pathway of a particular disease. We have also revealed how genes in the selected cell specific components involve feedback at multiple levels and affect the normal functioning pathway of a disease.

3.6 Topological analysis of biological networks

3.6.1 Extraction of genes associated with disease of interest from the literature

We sought applicant genes connected with disease of interest by PolySearch content mining framework, which can create a rundown of ideas significant to the users queries by scrutinizing numerous data sources. We utilized PolySearch framework to look the genes connected with disease of interest. The query type is Disease-Gene/Protein Association and the question query keyword relevant to particular disease. PolySearch framework gives back 1876 literary works by its default parameters. On the resulted gene names of specific threshold against disease of interest, we applied the first two sections of the current research; Data collection against genes from PubMed and Association rules mining for those gene.

3.6.2 Scanning protein-protein interactions

The competitor genes recorded were changed over to be the seed proteins. PPIs were acquired from STRING database, a pre-processed database for the investigation of proteinprotein interactions. The most up to date adaptation of STRING, 9.0, spreads pretty nearly 2.5 million proteins from 630 distinct creatures.

3.6.3 Construction of PPIs network and extracting the giant component from the extended network

An extended network is developed that comprises of the seed proteins as well as their direct PPI neighbors and the interactions between these proteins. The system was built utilizing Pajek, an exceptionally adaptable project for the investigation, operation and representation of expansive systems. In this study, the amplified system incorporates a giant segment and some little separate segments got from seeds proteins. This study expected to investigate the mechanism of disease of interest at the molecular level and the nodes with huge BC esteem must be in the titan arrange clearly on the grounds that little separate parts comprise of little number of hubs, so just the giant component and its parameters identified with the system hypothesis had been broken down or handled. We make more wise system design on the premise of Fruchterman Reingold vitality. The Fruchterman-Reingold Algorithm is a power coordinated format calculation that considers a power between any two nodes. The fundamental thought is to minimize the framework's vitality by moving the nodes and changing the strengths between them.

Net \rightarrow partition \rightarrow degree \rightarrow output

The giant network is partitioned on distinctive premise like degree, domain, depth, center and so forth. These segments will give us the clusters of nodes on closeness premise. These segments will group the nodes on the premise of same degree or

nodes having same domains or depth and so forth by assigning different colours to different clusters.

Net \rightarrow vector \rightarrow summing up lines

A partition just gatherings nodes together, yet a vector can relegate to every node a specific numerical value that can be shown in the graph utilizing the node' sizes. Pajek can figure a few vectors about the network. For example, we can ask what number of aggregate nodes every node is an essential for by requesting that it aggregate up what number of arrows lead out from every node. The outcome demonstrates the most obliged nodes in a network.

3.6.4 Topological analysis of protein interaction network

In this study, properties of nodes and estimations used to describe network were figured by Pajek programming.

3.6.5 Shrinking a network

It is fascinating to get a superior perspective of the system being considered here. Note that every group in the system was named with a subjective node from that segment. The proteins with high BC ought to be the intensely utilized convergences, these proteins and the connections between them make up a spine system. The basic purpose of high BC was set at 5% of the aggregate node set of the system. These high BC hubs and the connections between them were extricated from the titan system to make a spine system. BC was initially acquainted with measure the nodes' centrality in a system. By definition, the greater part of the briefest ways in a system experiences the nodes with high BC. These hubs capacity as bottleneck controls the correspondence among different nodes in the system.

3.6.6 Identification of structural holes

Structural holes are another topological element of systems which are critical in distinguishing the part of nodes in the connections among sub clusters of graphs. Structural holes isolate non-excess wellsprings of data, sources that are more added substance than covering. The expression "Structural holes" alludes to some essential parts of positional point of preference/disservice of nodes that outcome from how they are implanted in neighborhoods. These standards shape supporting for the hypothesis of structural gaps, which thinks about the routes in which nodes, fill the "holes" between gatherings that are not generally cooperating. It is finished by evacuating circles, uprooting various lines, from system unique and looks at the position of every node in their neighborhood for the vicinity of basic gaps. Various measures (most proposed by Burt) that depict different parts of the point of preference or hindrance of the nodes are additionally registered.

3.6.7 Scale-freeness topology of the networks

The idea of scale free system has developed as a capable binding together world-view in the investigation of complex frameworks in science and in physical and social studies. Metabolic, protein, and gene collaboration systems have been accounted for to show scale free behaviour based on the analysis of distribution of the number of associations of the system nodes. Natural systems have been portrayed by topological components which build up their scale-freeness property. Protein collaboration systems and co-expression arranges additionally show a scale free geometry, where the nodes are not consistently populated with neighbors. Every one of the node of these systems don't take after the tenet of having a normal number of connections per node. The greater part of the nodes has few accomplices, while a couple of hubs likewise called "center points" interface with numerous accomplices. Power law procedure is utilized for evaluating the parameters and accepting the system models with their scale-freeness property. The systems are without scale i.e., they are unevenly populated with center points and less dense

nodes. Biological networks are observed to be extremely touchy to the evacuation of center point proteins. It has been watched that the cancellation of center proteins in yeast protein-protein cooperation system applies an expanded deadly impact.

3.6.8 Detection of hub nodes

In the context of a network, a center is a node with a huge degree, significance it has associations with numerous different nodes. Biological networks are observed to be extremely touchy to the evacuation of center proteins. It has been watched that the cancellation of center point proteins in yeast protein-protein association system applies an expanded deadly impact.

3.7 Summary

This chapter is based on methods to achieve the objectives of the research. The methodology and the materials that have been adopted to carry out the current work are divided into four different sections, each describing the details of approached utilized in detail. Our approaches efficiently identify the association rules/biological interactions and would be useful for further downstream analysis. PMIC has been developed with the target of giving a platform to information extraction and mining helpful data from the extracted information from the biological database PubMed. It adds to the abilities of the current apparatuses and servers for information extraction and analysis. Further in this chapter, details of the tools and techniques have been discussed which have been used while conducting out our research work. Tools have been categorized into different sections, including hardware and software specifications. Details of each tool have been given in the respective sections.

Chapter 4

Results and analysis

In this chapter, we explain step by step the results obtained as a result of methodologies followed in chapter 4 described above. It is mainly comprised of different sections of research methodology and the summary concluding all the findings. The results of each phase of preprocessing methodology were evaluated and validated using KEGG pathways and from literature source. The biological network was generated for the gene variants and gene coding proteins of T2DM using association rule mining techniques. Topological analysis was done for the network generated by Pajek for Ankylosing Spondylitis. To generate gene-gene association network for T2DM and to extract the connected components from that generated network, following steps had been employed.

4.1 Data collection from biological database for T2DM

The analysis of the proposed computational approach for collecting text data from PubMed for further processing has been explained step by step.

4.1.1 Getting Keywords related to T2DM

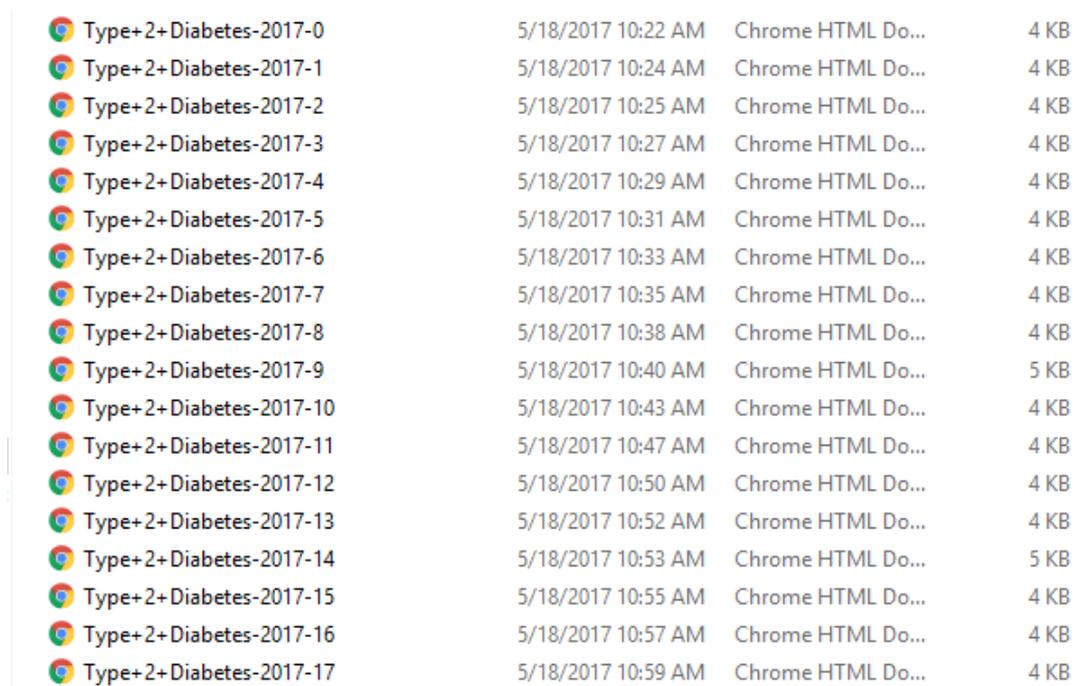
As the common diseases with high incidence, Type 2 diabetes mellitus and Ankylosing spondylitis gains much attention among researchers and has a rather large literature accumulation. We initially start our work by construction and analysis of PPI network for Ankylosing Spondylitis, one of the most common types of Arthritis. For this purpose, we used the already existing tool Pajek. But onwards we switch our research to T2DM. This disease is not involved in a particular organ or cell. Cell and organ specificity was must in this research for the purpose of extracting connected components and pathways cross talk, otherwise the current research could have multiple directions which is beyond the scope. We used Type 2 diabetes as testing disease for system evaluation. We experimentally tested our application on Type II Diabetes Mellitus data for the last five years (2013 to 2017). The tables below show the different possible query keywords for Type II diabetes mellitus that were used for testing the proposed application. In order to be more accurate in our results we used all possible terms for Type II Diabetes for the collection of data. Here is a list of possible queries used to work on current model.

- Query 1: Type II Diabetes
- Query 2: T2D
- Query 3: T2DM
- Query 4: type 2 diabetes
- Query 5: diabetes type 2
- Query 6: type 2 diabetes mellitus
- Query 7: diabetes mellitus type 2

4.1.2 Connecting to PubMed

After getting above mentioned query terms from the user the request was sent to `entrez.ncbi.nlm.nih.gov`, to fetch results from this server and then generate html

file to display output. The output against each query was returned in the form of .html files that contains titles of articles and the URLs against the specific query term. Figure 4.1 and figure 4.2 show the screenshot of the results against the query Type 2 Diabetes for the year 2017 from PubMed.



Type+2+Diabetes-2017-0	5/18/2017 10:22 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-1	5/18/2017 10:24 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-2	5/18/2017 10:25 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-3	5/18/2017 10:27 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-4	5/18/2017 10:29 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-5	5/18/2017 10:31 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-6	5/18/2017 10:33 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-7	5/18/2017 10:35 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-8	5/18/2017 10:38 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-9	5/18/2017 10:40 AM	Chrome HTML Do...	5 KB
Type+2+Diabetes-2017-10	5/18/2017 10:43 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-11	5/18/2017 10:47 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-12	5/18/2017 10:50 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-13	5/18/2017 10:52 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-14	5/18/2017 10:53 AM	Chrome HTML Do...	5 KB
Type+2+Diabetes-2017-15	5/18/2017 10:55 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-16	5/18/2017 10:57 AM	Chrome HTML Do...	4 KB
Type+2+Diabetes-2017-17	5/18/2017 10:59 AM	Chrome HTML Do...	4 KB

FIGURE 4.1: Screenshot of the Output of the PubMed Info Extractor (PMIE) against Type 2 Diabetes for the year 2017

The .html files in Figure 4.1 contain the articles that are in PubMed related to Type 2 Diabetes. On clicking each html file, we get the titles and URLs of articles against Type 2 Diabetes. Each html file contains maximum 20 articles for the specific query. Type-2-Diabetes-2017-0 contains the first 20 results for Type 2 diabetes. Type-2-Diabetes-2017-1 contains next 20 articles found for the query Type 2 Diabetes for the year 2017. Figure 4.2 shows the first 20 results of html file Type-2-Diabetes-2017-0.

In Figure 4.2, the address bar shows first 20 articles against the query Type 2 Diabetes for the year 2017. File name Type+2+Diabetes-2017-0.html means this is the first page of the PubMed results against the specific query. Similarly in the

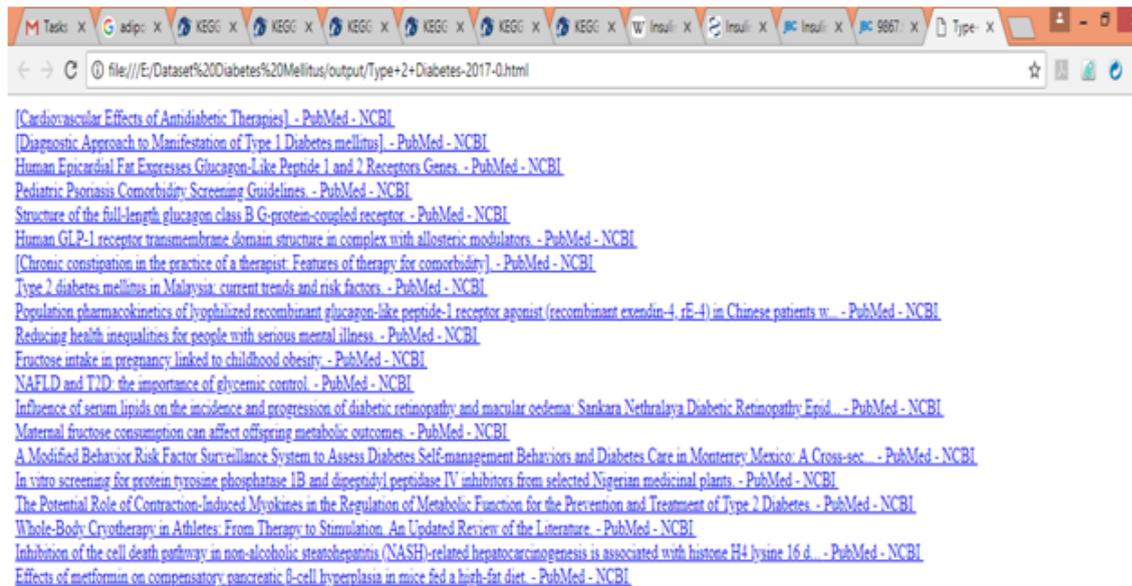


FIGURE 4.2: Screenshot of the Titles of the output of figure 5.1 with URLs

address bar %20diabetes%20Mellitus represents that each .html page of Figure 4.1 contains 20 articles per page.

4.1.3 Extracting articles details

By using the keywords as query terms in PUBMED, we have saved all the papers links that are returned to us as a result of these queries as shown in Figure 4.2. From each paper, the following information was extracted; Pubmed id of each document, title, abstract, authors names and year of publication. All this work was done by screen scrapping methodology, a code was written in PhP platform to extract all the required information from the PUBMED database against each query term. The retrieved information was saved in different formats like SQL, CSV or excels, PDF etc to carry out further processing. Figure 4.3 shows the screenshot of what type of information we get as a result of further processing on the above mentioned steps. The highlighted portion shows the authors affiliation (Department of Molecular and human Genetics, Journal Name: Carcinogenesis and PubMed ID: 28535186). Using the screen scrapping code, for each article link shown in Figure 4.2, the abstract, Authors names of an article, authors affiliation,

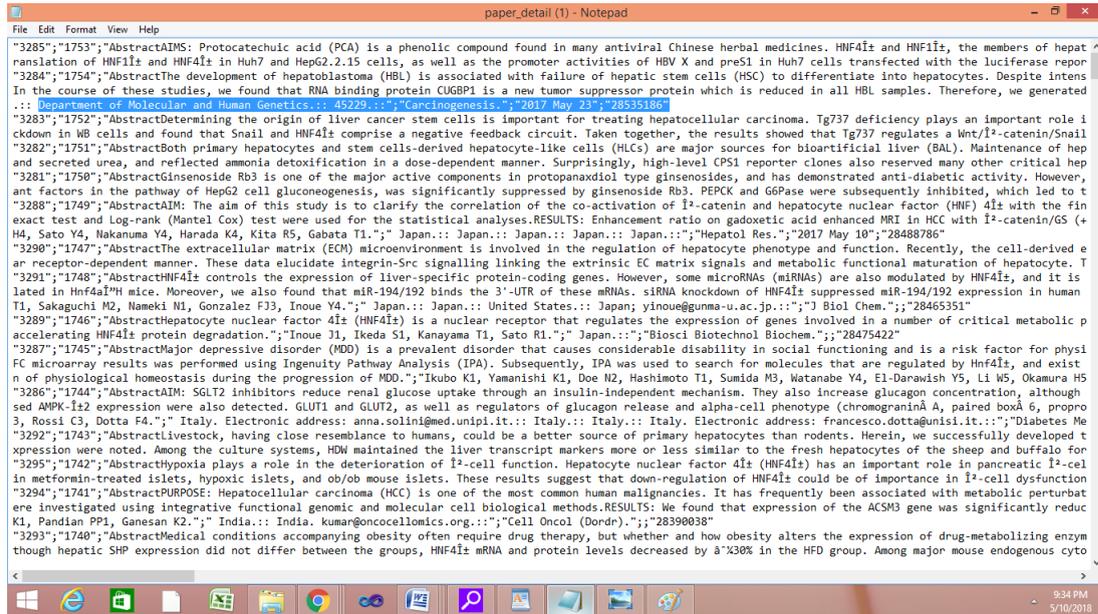


FIGURE 4.3: Screenshot of the paper details against specific query

Journal name and PubMed ID has been extracted. We evaluated our tool by using Heuristic expert based evaluation methods. The articles and their detailed information related to different possible query terms against Type II diabetes mellitus were extracted. The results of each phase of preprocessing methodology were evaluated and validated. Figure 4.4 shows the graphical representation of the data articles obtained from PubMed from 2013 to 2017 against different query terms of Type II Diabetes. This figure shows the rise and fall of research work in the area of Diabetes Mellitus. The evaluation was done by the real user that

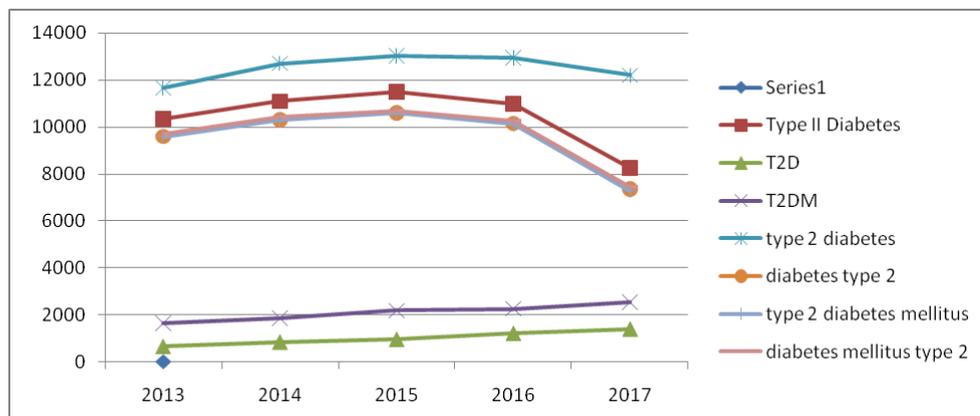


FIGURE 4.4: Graphical representations of the data against PubMed queries

belongs to department of Computer science, department of Bioinformatics and Biosciences of Capital University of Science and technology. Table 4.1 shows the evaluation results along with tools accuracy against different query keywords and their results of the crawler. The accuracy was obtained through the ratio of results in PubMed to the results through crawler. Percentage accuracy is obtained in the same way.

The number of articles in PubMed against the query Type II Diabetes for the year 2013 was found to be 10353. The number of articles we got through PubMedInfo Crawler was 10,965 for the year 2013 against the same query term. The %age accuracy of the results calculated through ($\text{Accuracy} = \text{Results in PubMed} / \text{Results through crawler} * 100$) was approximately 94%. Similarly the number of articles in PubMed against the query Type II Diabetes for the year 2014 was found to be "11088". The number of articles we got through PubMedInfo Crawler was 10,562 for the year 2014 against the same query term. The %age accuracy of the results calculated through ($\text{Accuracy} = \text{Results in PubMed} / \text{Results through crawler} * 100$) was approximately 95%. In the same way we have checked the accuracy of our results for the years 2015, 2016 and 2017 against the query term "Type II Diabetes". The average accuracy calculated against this query term for years 2013 to 2017 was found to be approximately 94%. Likewise the average accuracy calculated against the queries "T2D", "T2DM", "Type 2 diabetes", "Diabetes type 2", "Type 2 diabetes mellitus" and "Diabetes mellitus type 2" for years 2013 to 2017 was found to be approximately 97%, 96%, 95%, 96%, 95%, 96%. The overall accuracy of the crawler was found to be 96% for the number of articles against query terms. The graphical representation of how much accuracy obtained in the extraction of papers details like abstract, authors names, authors country, journal name, publication date is shown in Figures 4.5, 4.6, 4.7, 4.8, 4.9. From those URLs, we got through PubMedInfo Crawler, the abstracts, Authors names, Authors affiliation, and Journal names got the accuracy 97% for the year 2013 against the query term "Type II Diabetes". The publication dates were extracted with the accuracy of 86% for the year 2013 against the same query term. The accuracy obtained for the abstracts, Authors names, Authors' affiliation, and Journal

TABLE 4.1: Last five years results evaluation by different users against possible T2DM queries

Query	Type II Diabetes				
Year	2013	2014	2015	2016	2017
Results in PubMed	10353	11088	11492	10982	8257
Results through crawler	10,965	10,562	10,789	10,167	7863
Accuracy	94%	95%	94%	93%	95%
Query	T2D				
Year	2013	2014	2015	2016	2017
Results in PubMed	651	842	951	1206	1395
Results through crawler	651	856	912	1149	1337
Accuracy	100%	98%	96%	95%	96%
Query	T2DM				
Year	2013	2014	2015	2016	2017
Results in PubMed	1636	1854	2183	2257	2549
Results through crawler	1596	1791	2147	2196	2456
Accuracy	97%	96%	98%	97%	96%
Query	Type 2 diabetes				
Year	2013	2014	2015	2016	2017
Results in PubMed	11641	12694	13022	12917	12186
Results through crawler	10967	12145	12323	12175	11586
Accuracy	94%	96%	95%	94%	95%
Query	Diabetes type 2				
Year	2013	2014	2015	2016	2017
Results in PubMed	9612	10323	10622	10172	7359
Results through crawler	9087	9865	10289	9693	7159
Accuracy	94%	96%	97%	95%	97%
Query	Type 2 diabetes mellitus				
Year	2013	2014	2015	2016	2017
Results in PubMed	9581	10281	10592	10124	7284
Results through crawler	8869	10112	10163	9599	6839
Accuracy	93%	98%	96%	95%	94%

Query	Diabetes mellitus type 2				
Year	2013	2014	2015	2016	2017
Results in PubMed	9674	10394	10666	10234	7438
Results through crawler	9359	9973	10259	9993	6934
Accuracy	97%	96%	95%	98%	93%

names against the query term T2D was 92% and the publication dates were extracted with the accuracy of approximately 88% against the query term T2D for the year 2013. In the same way the accuracy of extracted detailed information was evaluated manually against all other query terms for the years 2013, 2014, 2015, 2016 and 2017. The overall accuracy for abstracts, authors, author's country and journal name was found to be approximately 98% whereas the publication dates got relatively less accuracy as compared to other information of about 83%. The reason for this was that many articles do not have exact date including day, month and year, they just have year of publication.

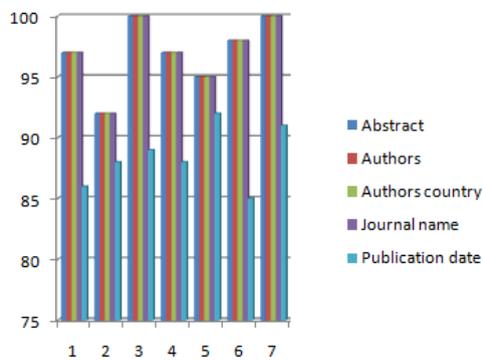


FIGURE 4.5: Paper details accuracy results against the query terms for year 2013

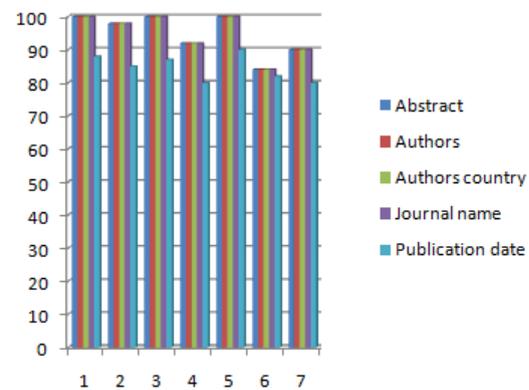


FIGURE 4.6: Paper details accuracy results against the query terms for year 2014

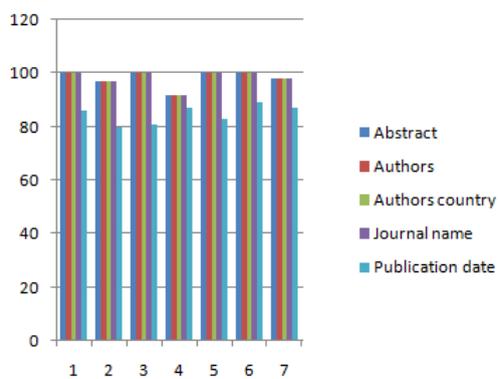


FIGURE 4.7: Paper details accuracy results against the query terms for year 2015

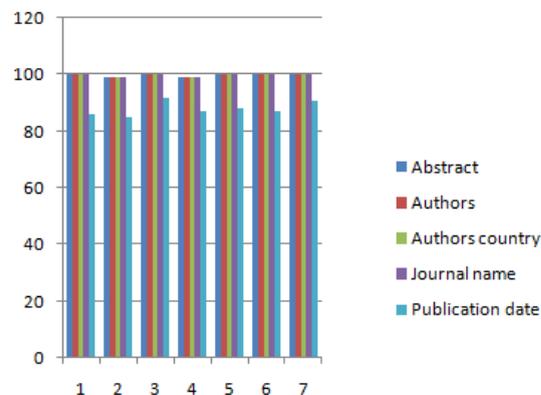


FIGURE 4.8: Paper details accuracy results against the query terms for year 2016

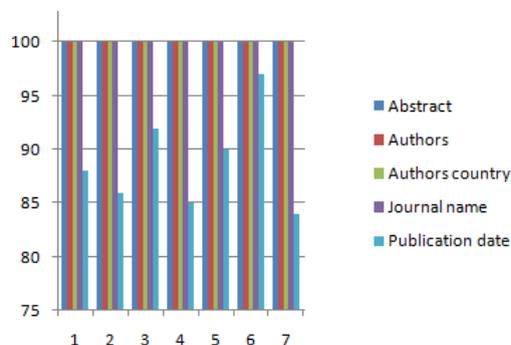


FIGURE 4.9: Paper details accuracy results against the query terms for year 2017

PMIC has been developed with the target of giving a platform to information extraction and mining helpful data from the extracted information from the biological database PubMed. It adds to the abilities of the current apparatuses and servers for information extraction and analysis.

4.2 Association rules mining techniques to generate T2DM network

4.2.1 Phase 1: Data collection

The course of action towards the disclosure of spurring relations among various Type 2 Diabetes genes and their outcomes are as per the following: The data resource for the proposed system assessment is PubMed database (<http://www.ncbi.nlm.nih.gov/pubmed>). PubMed comprises of around 20 million archives for writings about biomedicines from MEDLINE, journals of life sciences, and online books. These reports may include relationship to full content substance from PubMed Central and distributor sites. We have to recover all the reports or documents significant to the queries. The abstracts of the considerable number of papers were downloaded from PubMed database by sending the accompanying queries Table 4.2:

TABLE 4.2: Articles retrieved from PubMed using PubMed Info Extractor against specific queries of Type II Diabetes Mellitus disease.

Queries	No. of Relevant articles
Type 2 diabetes mellitus	1, 07361
type 2 diabetes	1, 29564
T2DM	12875

Around 2, 49,800 papers were found in PubMed that were relevant to these query terms. By using the keywords as query terms in PUBMED, we had saved all the papers that are returned to us as a result of these queries. From each paper, the following information was extracted; PubMed id of each article, title, abstract, authors names and year of publication. All this work was done by screen scraping methodology, a code was written in php platform to extract all the required information from the PUBMED database against each query term. The retrieved information was saved in .XML file format to carry out further processing.

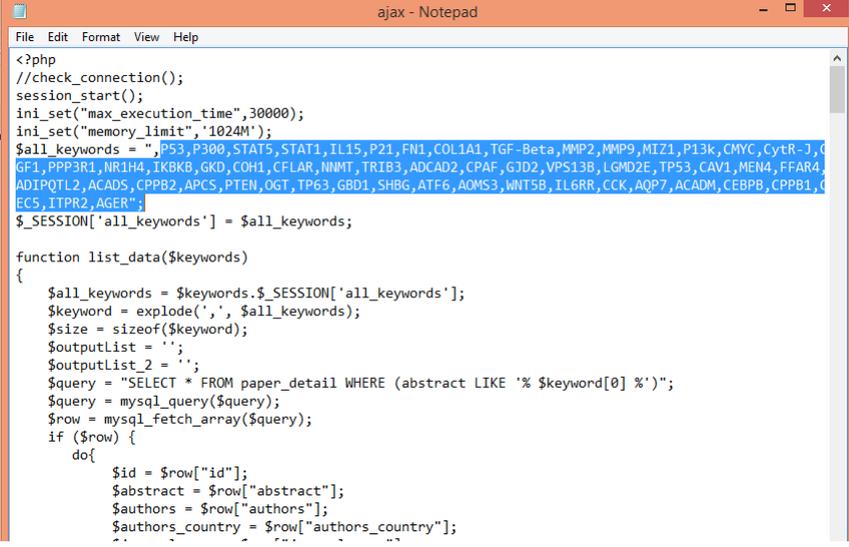
4.2.2 Phase 2: Identification of hypothetical association terms for T2DM

In order to screen the gene variants of T2DM, GWAS database has been used. GWAS central encloses 69,986,326 relations between 2,974,967 distinctive SNPs and 829 inimitable MeSH disease/Phenotype descriptions. To date, numerous mutations have been shown to affect type II diabetes risks. The contribution of each gene is generally small. GWAS have identified nearly 52 common risk variants that show association with T2D. These variants and their details are given in Table A.1 (Appendix) [127–157]. Scientists utilize distinctive ways to deal with genome annotation, their prophecy of the amount of genes on every chromosome change. Among different undertakings, collaborative consensus coding sequence project (CCDS) takes an amazingly traditionalist technique. So CCDS's gene number approximation corresponds to a lower bound on the aggregate number of human protein-coding genes. Some of the notable genes of human chromosome 1 to human chromosome 22 were taken from CCDS's gene list [158].

4.2.3 Phase 3: Processing phase

4.2.3.1 Segmentation

The obtained 249800 articles from PubMed against different query terms were then subdivided/segmented. The connection was established with the obtained results of phase 1 that were saved in .XML format. The segmentation was done by using the principle that the query given to the model was gene names screened in phase 2, one by one. The tools paired that query word with all the other gene variants and protein coding genes of chromosome 1 to chromosome 22. The gene names were also mentioned at the back end in the code file. The path is as follows: Local drive → Xampp → htdocs → associationrulegenerationmodel → function → ajax.php (Figure) 4.10.



```

<?php
//check_connection();
session_start();
ini_set("max_execution_time",30000);
ini_set("memory_limit","1024M");
$all_keywords = "P53,P300,STAT5,STAT1,IL15,P21,FH1,COL1A1,TGF-Beta,MMP2,MMP9,MIZ1,P13k,CMYC,CytR-J,
GF1,PPP3R1,NR1H4,IKBKB,GKD,COH1,CFLAR,NM1,TRIB3,ADCAD2,CPAF,GJD2,VPS13B,LGMD2E,TP53,CAV1,MEN4,FFAR4,
ADIPQTL2,ACADS,CPPB2,APCS,PTEN,OGT,TP63,GBD1,SHBG,ATF6,AOM3,WNT5B,IL6RR,CCK,AQP7,ACADM,CEBPB,CPPB1,
ECS,ITPR2,AGER";
$_SESSION['all_keywords'] = $all_keywords;

function list_data($keywords)
{
    $all_keywords = $keywords.$_SESSION['all_keywords'];
    $keyword = explode(",",$all_keywords);
    $size = sizeof($keyword);
    $outputList = '';
    $outputList_2 = '';
    $query = "SELECT * FROM paper_detail WHERE (abstract LIKE '% $keyword[0] %)";
    $query = mysql_query($query);
    $row = mysql_fetch_array($query);
    if ($row) {
        do{
            $id = $row["id"];
            $abstract = $row["abstract"];
            $authors = $row["authors"];
            $authors_country = $row["authors_country"];
        }
    }
}

```

FIGURE 4.10: Screenshot of Gene variants and notable protein coding genes for association rules

As a result of segmentation process, all those PubMed articles related to the above mentioned queries were subdivided and articles that contained query word, gene name paired with all other genes variants and known protein coding genes on the basis of given query word were fragmented. In the same way segmentation process is applied on all genes variants of T2DM.

4.2.3.2 Sentence splitting technique

The sentences of each segmented article are splitted by ”.”. Each splitted sentence contains at least one target gene of T2DM and hypothetical gene term. If both query gene and its paired gene (gene variant of T2DM or known protein coding gene on chromosome 1 to chromosome 22) occur in the same sentence in an abstract of an article, it is counted as a score of one, showing that there is some type of strong association between paired genes. If both query gene and its paired gene occur in an abstract of an article but not in the same sentence, then it is counted as a score of ”0.5”. This shows that both genes have some association between them but it may not be very strong. The higher score shows the strength of association between two genes. The sentence splitting technique is applied to all the other gene variant of T2DM one by one and the scores are calculated.

4.2.3.3 Text documents representation

In data recovery, tf idf, short for term frequency inverse document frequency, is a numerical measurement that is planned to reflect how vital a word is to a report in a gathering or corpus. Usually utilized as a weighting factor in data recovery, content mining, and user modeling. The tf-idf of each query gene variant is calculated to show how much it is important for causing T2DM or how much relation it has with T2DM.

4.2.4 Mining association rules

The next phase towards association mining is generation of rules. Association mining rules were used to generate the rules from the dataset with no support and confidence threshold. As the rule having minor support and confidence value is important to show relationship between two genes. Tables 4.3, 4.4, 4.5 in results and analysis section and in Appendix section from Tables table A.2–A.124 show the association rules along with confidence and support value of each rule. Each table has antecedent(X) and consequent(Y) terms along with confidence, support and lift values for each antecedent and consequent. In an association rule($X \rightarrow Y$), antecedent is the input variable that we can control, and the consequent is the variable we are trying to predict. Antecedent is the specific keyword for which we are interested to find association with other terms in the data set. Consequent is the term that comes in association with specific input keyword in the given/available dataset. Antecedent and consequent coupled to form a rule. The support and confidence is one of the essential tasks in generating strong association rules from the frequent item sets. An earlier most research, the interesting rules were considered based on a two basic measure such as support and confidence. Support, confidence and lift values couple can be used for choosing the best rules. These terms are interdisciplinary terms. The expected confidence of a rule is defined as the product of the support values of the rule (XUY) divided by the support of the rule (X). The lift is a value between 0 and infinity: A lift

value greater than 1 indicates that the antecedent and consequent terms appear more often together than expected; this means that such an association rule has more significance. A lift smaller than 1 indicates that the antecedent and consequents terms appear less often together than expected, this means that such an association rule has less significance. A lift value near 1 indicates that the antecedent and consequent appear almost as often together as expected; this means that the occurrence of such association between antecedent and consequent has no specific impact. On the basis of lift values for an extracted association rule, we have highlighted only the significant association rules in the given tables. In table

TABLE 4.3: INS Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	INS	→	INSR	0.85	0.67	1.23
2	INS	→	CDKN2A	0.543	0.75	1.61
3	INS	→	Foxa1/2	0.628	0.69	1.42
4	INS	→	JNK1/2	0.56	0.62	2.09

4.3, antecedent is INS. It means INS is the term for which we are trying to predict all possible rules or associations within the available dataset. Consequents are INSR, CDKN2A, Foxa1/2 and JNK1/2. These four are the terms or genes that have some association with INS. Each rule has different lift value. As mentioned above, significant rules are those, which have value greater than 1. The greater the lift value from 1, the more significant association or relationship is between two terms. The association rule $INS \rightarrow JNK1/2$ is found to be significant among all other association rules for INS gene. In table 4.4, antecedent is INSR. INSR is the term for which we are trying to predict all possible rules or associations within the available dataset. Consequents are IRS1, HNF1Alpha, INS and PL1N and AdPLA. This means they have some association with INSR. The association rules $INSR \rightarrow PL1N$ and $INSR \rightarrow AdPLA$ are found to be significant among all other association rules for INSR gene. In table 4.5, antecedent is IRS1/2. IRS1/2 is the term for which we are trying to predict all possible rules or associations within the

TABLE 4.4: INSR Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	INSR	→	IRS1	0.94	0.73	1.30
2	INSR	→	HNF1Alpha	0.844	0.77	1.47
3	INSR	→	INS	0.86	0.79	1.75
4	INSR	→	PL1N	0.53	0.68	1.86
5	INSR	→	AdPLA	0.53	0.64	2.05

TABLE 4.5: IRS1/2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	IRS1/2	→	P13k	1.01	0.78	1.39
2	IRS1/2	→	SHC1	0.94	0.74	2.34
3	IRS1/2	→	IGFBP3	1.04	0.54	1.26
4	IRS1/2	→	Akt	0.98	0.49	1.15

available dataset. Consequents are P13K, SHC1 and IGFBP3. This means they have some association with INSR. The association rule $IRS1 \rightarrow SHC1$ is found to be somehow more significant among all other association rules for INSR gene. Similarly, $Akt \rightarrow AS160$, $Akt \rightarrow PDE-3B$, $Akt \rightarrow FST$, $Akt \rightarrow BAD$, $THADA \rightarrow CAMK1D$, $THADA-STAT1 \rightarrow BAD$, $MIZ1 \rightarrow UCP1$, $SHIP2 \rightarrow PIP3$, $PDK1/2 \rightarrow PDE-3B$, $CDKN2A \rightarrow LEP$, $SLC30A8 \rightarrow ObR$, $CDKAL1 \rightarrow WFS1$, $MTNR1B \rightarrow PL1N$, $NOTCH \rightarrow CAMK1D$, $GL1S3 \rightarrow Ngn3$, $ACDC \rightarrow EGFR$, $C-JUN \rightarrow DUSP9$, $GLuT4 \rightarrow ATGL$, $Foxa1/2 \rightarrow PyK$, $PDX1 \rightarrow TSH$, $aPKC \rightarrow SREBP-1C$, $ObR \rightarrow AMPK$, $BCL11A \rightarrow Sox9$, $BCL11A \rightarrow JNK1/2$, $AMPK \rightarrow ACC$, $AMPK \rightarrow FAS$, $BAD \rightarrow WFS1$, $RXR \rightarrow AS160$, $GK \rightarrow RXR$, $SHC1 \rightarrow UCP3$, $AS160 \rightarrow FN1$ are some of the associations that are found to be more significant in this research. The rest of the tables showing association rules are shown in AppendixA tables [A.2](#) to [A.124](#) .

Figure [4.11](#) shows the graphical layout of the above mentioned retrieved association rules. We represent all the association in the graphical format using Pathvisio,

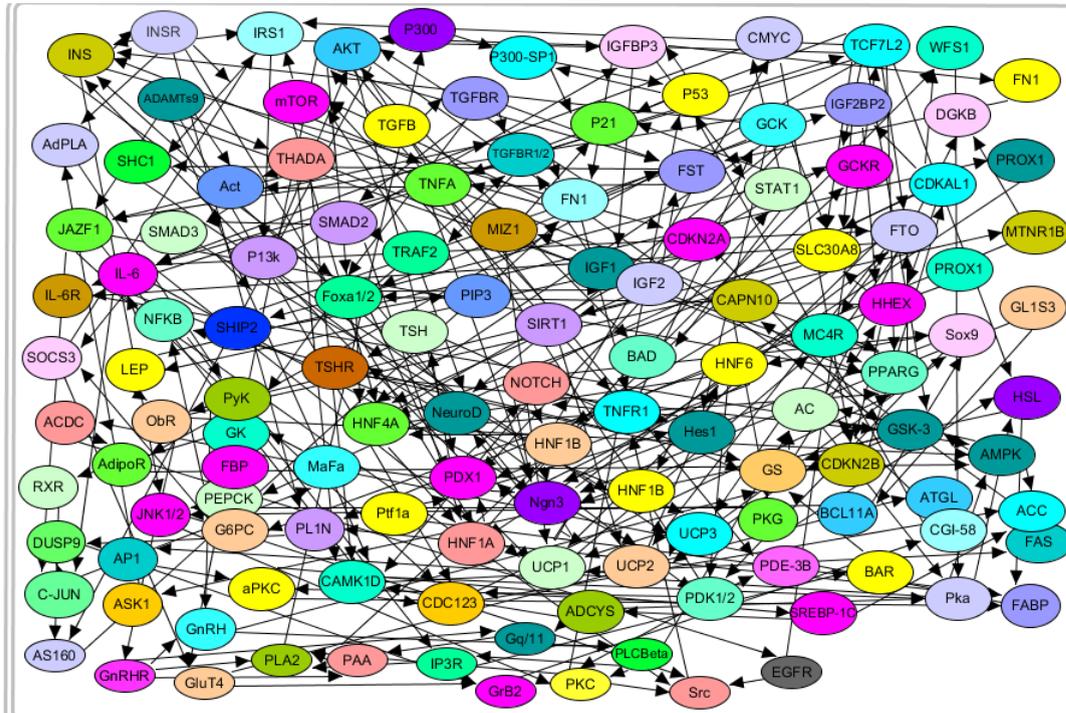


FIGURE 4.11: Graphical representation of association rules showing association between genes

version 3.3.0, without applying any threshold value. Some might be keen on affiliations that are self-evident (relationship with higher potency) and some might be more inspired by slightest clear or anomalies (relationship with less potency). Yet, others might be keen on relationship with moderate strength. Nodes color represents nothing just to make the layout attractive. These associations have been verified by Gephi and KEGG database. Georgios A. Pavlopoulos et al., in 2017, listed a list of accessible system representation devices and, from a user encounter perspective recognized four competitor tools appropriate for bigger scale network examination, perception, and investigation. Among several existing tools that they tested, Cytoscape [159], Tulip [160], Gephi [161], and Pajek [162] standalone applications were found to be the main four contender for perception, control, investigation, and examination of huge systems. Figure 4.12 shows the network generated by Gephi for the GWAS gene variants of T2DM with their consequents.

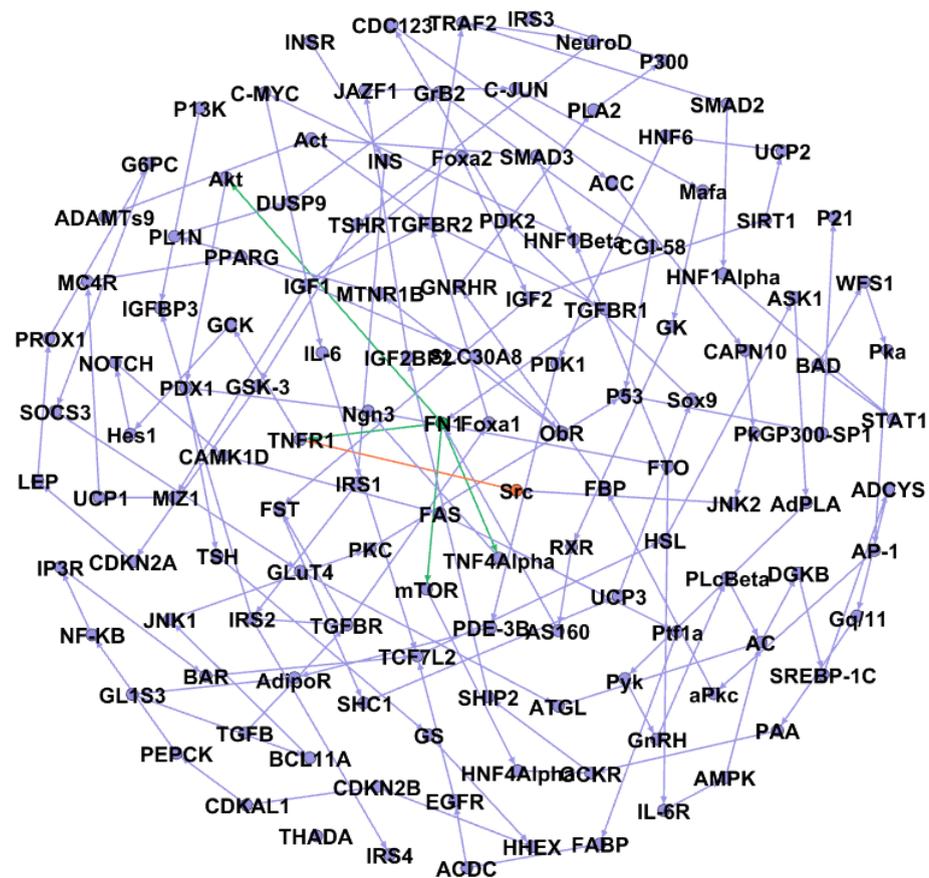


FIGURE 4.12: Association among GWAS known gene variants of T2DM and their consequents by using Gephi [161]

4.2.4.1 Biological interpretation of T2DM network:

The interactions among different genes in our extracted association network and their nature of relationship strongly correlate with pathways available in KEGG database. These pathways include [MAPK signaling pathway](#), [TGFB signaling pathway](#), [P53 signaling pathway](#), [Insulin signaling pathway](#), [Age-Rage signaling pathway](#), [mTOR signaling pathway](#), [Type II diabetes mellitus pathway](#), [MODY pathway](#), [Cell cycle pathway](#), [Protein processing in endoplasmic reticulum](#), [Insulin secretion](#), [Apoptosis pathway](#), [Neuroactive ligand receptor interaction](#), [Glycolysis](#), [Lipid homeostasis](#), [Regulation of lipolysis in adipocytes](#), [cGMP-PkG signaling pathway](#), [cAMP signaling pathway](#), [Glycerolipid metabolism](#), [Non alcoholic fatty liver disease](#), [TNF signaling pathway](#), [Adipo cytokine signaling pathway](#), [GnRH](#)

signaling pathway, Calcium signaling pathway, Cytokine-Cytokine receptor interaction. Insulin (INS) is involved in the growth and development of tissues and control of glucose homeostasis by stimulating the glucose transport into muscles and adipose cells. INS is secreted by the beta cells of the pancreas. Active INS molecule consists of two chains of alpha and beta that are held together by disulfide bond. Normal INS secretion occurs through activation of specific insulin receptor tyrosine kinases INSR, INS attaches to the alpha subunit of the receptor by the hindrance of tyrosine auto phosphorylation by the beta subunit [164]. Triggered INSR phosphorylates insulin receptor substrate IRS (1-4). There are four elements of IRS family, IRS1, IRS2, IRS3 and IRS4. Among these IRS1 and IRS2 assume essential part in glucose transport. Tyrosine enacted IRS proteins proceed as requisite sites for flagging molecules containing SH-2 (Src homolog 2) sphere, for example, P13K [165]. P13K is the fundamental flagging negotiator of metabolic and mitogenic proceedings of insulin and forms the signaling complex to mediate downstream signaling. P13K is composed of p85 subunit that binds to IRS protein and a p110 catalytic subunit. On binding to IRS1/2 proteins P13k increases the catalytic activity of p110 and allows the phosphorylation PtdIns P2 to generate P3 [166]. Tyrosine phosphorylated IRS1/2 recruits p85 and p110 at plasma membrane and produce second messenger PIP3. PIP2 and PIP3 are phospholipid components of cell membrane; they are enriched at plasma membrane and act as substrate for many important signaling proteins [167]. PIP3 is the key lipid signaling intermediary enduring dephosphorylation by two lipid phosphatases, SHIP2 (SH2-containing inositol 5-phosphatase-2) and PTEN (phosphatase and tensin homolog). PIP3 initiates PDK1, PKB/Akt and PKC to plasma layer by means of PH sphere [166]. Commencement of these kinases brings about insulin responses, for example, GluT4 translocation to the membrane, glycogen production of GSK-3 and AS160 and lipogenesis by up directing the synthesis of unsaturated fat synthase gene [168]. PKB mediated phosphorylation inactivates GSK3 that relieve inhibition of GS. GS turns out to be stimulated and endorse glycogen synthesis. PKB mediated phosphorylation regulates GluT4 at plasma

membrane through inhibitory phosphorylation of AS160 and results in glucose uptake [169]. aPKC proceeds in correspondence as a substituent for PKB. Activated PKB/Akt is responsible for antiapoptotic effect of insulin by phosphorylation of BAD. Side by side, activated IRD1/2 recruits Grb2, associates with SOCS3 and activates ERK1/2 MAPK pathway. IRS1/2 binds with GrB2 that on binding with SOCS3 activates SMAD7 that results in the inhibition of TRAF2. TRAF2 has indirect effect with mTOR that involves in the phosphorylation of IRS1/2 [170]. P13K maintains the insulin sensitivity in the liver; reduce activity of P13K leads to insulin resistance that refers also to type II diabetes mellitus. P85 subunit of P13K regulates the activation of P13K enzyme by binding to the binding sites on the IRS. Increased phosphorylation of IRS is involved in the insulin resistance by reducing their ability to attract P13K. Diabetic factors including TNF(s) induce insulin resistance through inhibition of IRS1 function. Interaction with SOCS3, regulation of the expression, degradation and many other molecular mechanisms are stimulated by the diabetic factors. Several kinases like mTOR, PKC, JNK are also involved in this process [166]. Activated tumor necrosis factor (TNF), is accumulated homotrimer and on coupling with its receptor TNFR1/2 ensuing in the trimerization of TNFR1 or TNFR2. TNFR1 signal provoke the commencement of several genes, mainly prohibited by two distinctive pathways, NFkB pathway and the MAPK cascade [171], [172], [173]. TNFR2 signaling stimulates NF-kappa B pathway as well as PI3K-dependent NF-kappa B pathway and JNK pathway that are important for endurance. NeuroD1 is the main regulator of pancreatic islet and insulin hormone gene transcription in islet beta cells. NeuroD1 activates the transcription of INS genes that binds with INSR and phosphorylates IRS1/2. NeuroD1 is also found to interact with P300 the results implicates P300 in NeuroD1 function to target the INS gene transcription during differentiation and in adult islet beta cells [174]. Inactivation of several genes encoding NeuroD1, PAX4, and PAX6 profoundly influences islet development. P53 acetylation is interceding by the P300 acetyltransferases. Overexpression of P300 efficiently induces specific P53 acetylation. P53 regulates transcription of IGFBP3 that causes the

inhibitor of IGF1/2. SIRT1 that activates IRS1/2 and phosphorylates P13K is inhibited by IGF1/2. IRS proteins modulation during insulin signaling takes place in serine/threonine phosphorylation. IRS1 tyrosine phosphorylation is reduced in skeletal muscles in case of obese and type II diabetes patients [175].

4.2.4.2 Self inhibition of Ngn3:

Regulatory genes marking endocrine precursors including PAX4, PAX6 and NeuroD1 are not expressed in the absence of Neurog3 (ngn3). Discrimination of pancreatic endocrine ancestry is reliant on ngn3 [176]. Ngn3 is a component of basic helix loop helix (BHLH) family. Several regulators that are involved in transcription play important roles in the activation of ngn3. These regulators include Hnf6 that is directly involved in the transcription of ngn3 and members of FoxA family, Foxa1/2. Ngn3 has been shown to have negative feedback as it represses its own expression. During the development of pancreas, ngn3 expression is instigating during the pancreatic budding as untimely as embryonic day. A dramatic upregulation occurs in the expression of ngn3, smearing the start of second transition. Expression is believed to peak till some stage and then swiftly declines to unnoticeable levels in juvenile and adult islets. PAX4, Nkx2 and BHLH are targets for the activation of ngn3 [177]. In pancreas, it is assumed that Notch signaling prevents the differentiation of cell. with the activation of Sox9 gene expression, Notch enacts the proendocrine gene Ngn3. Notch activity obstructed the endocrine differentiation when overexpressed, as it prompts enunciation of Hes1 which repress the Ngn3 expression. By lowering Notch overexpression, just Sox9, yet not Hes1, is kept up, in this way derepressing Ngn3 and starting endocrine delineation [178]. Comprehensively activated Foxa1 and Foxa2 can collaborate to intensify Ngn3 autoregulation in vitro [179].

4.2.5 Transitive association rule mining

The direct associations between genes/proteins have been exposed from the collected set of articles by utilizing the co-occurrence rule of data mining. In order to accomplish the task of finding transitive relations between genes, the most common transitive closure property has been applied to the association mining graph. The prospective transitive associations are basically those relations that definitely are exposed by iterative repositioning and drawing out of the PubMed database. The affiliations that are not discovered unequivocally in the whole PubMed database are termed as transitive and such relations are contender for generating tentative hypothesis. tables [A.125](#) to [A.203](#) shows the transitive association rules for the known GWAS gene variants of T2DM with known protein coding genes of human chromosome. Figure [4.13](#) shows the graph layout of the above mentioned retrieved transitive association rules using Pathvisio version 3.3.0. These associations are not mentioned directly in literature. Association rule $INS \rightarrow INSR$ shows direct relationship of INS gene with INSR. Similarly association rule $INSR \rightarrow IRS1$ shows direct relationship of INSR (insulin receptor) with IRS1. Now according to the definition of transitive association, if gene X has relationship with gene Y and gene Y has relationship with gene Z then X and Z are interconnected indirectly. INS has relationship with INSR and INSR has association with IRS1 then it means INS has indirect effect on IRS1. In the pathway description of insulin receptor signaling interactive pathway, it is cleared that Insulin is the most important hormone controlling vital energy roles that includes glucose and lipid metabolism. Insulin prompts the insulin receptor (INSR), which phosphorylates and commences assorted substrate adaptors like the IRS family of proteins. This relationship proves our results for the nature of relationship between INS, INSR and IRS1.

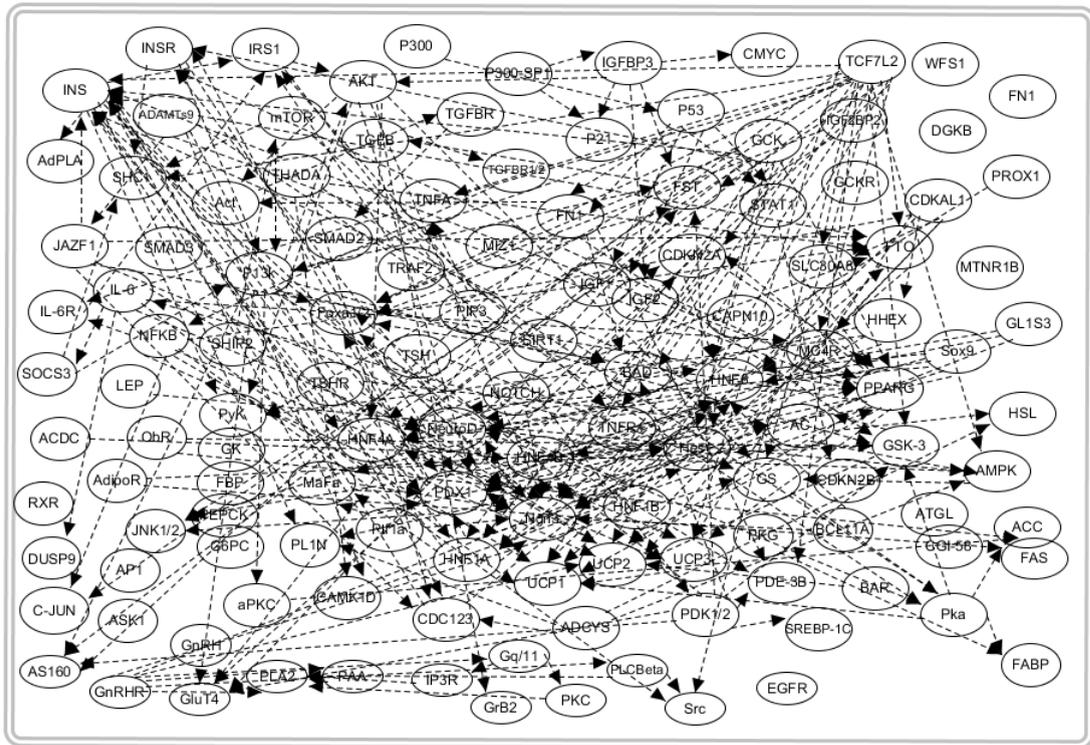


FIGURE 4.13: Transitive Association among gene variants of T2DM

4.2.6 Nature of relationship between genes responsible for T2DM

The last move on in the process is to extort acquaintances between the labeled entities. These associations incorporate for instance, protein A represses protein B. For this situation, a researcher not just has to realize what the relationship is yet additionally the direction in which the relationship happens. Figure 4.14 gives the elaborated graphical view of the above mentioned association graph by showing the nature of relationship between two genes.

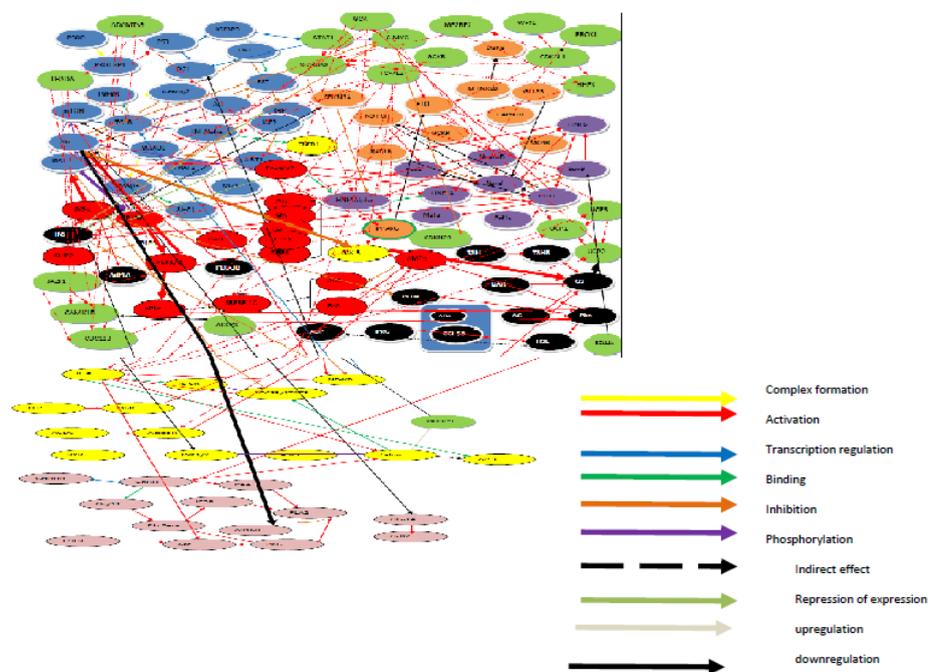


FIGURE 4.14: Association among gene variants of T2DM along with nature of relationship

4.3 Extraction of strongly connected components within a T2DM network

4.3.1 Extraction of connected components within association graph

In a network, there exist many connected components that lead towards many different pathways. In other words, a network is a combination of many pathways. Each pathway has a starting node and it is called connected component of the network, if there exists a path to reach the starting node of the pathway. The nodes that make the path to reach the starting node of the pathway are called main drivers or master regulatory genes that control normal cellular processes and disease pathologies and provide novel insights in understanding basic mechanisms of biological networks. The break point at any point in the connected component can altered the normal functional pathway and leads towards the disease. So such connected components play important role in understanding the disease mechanism.

We applied the Tarjan's algorithm for finding the connected components in our generated gene association network. From the association network generated by the association rule mining approach as mentioned in the result section, seven connected components have been retrieved. These components have been shown below in Figures figs. 4.16 to 4.22. The starting genes for the corresponding components, component 1 to component 7 are P53, HNF1Alpha, HNF1Beta, INSR, INS, IL-6 and GnRH. The path followed by the nodes of the component 1 to reach its starting node P53 is P53→IGFBP3→IGF1→IRS1→SHC1→TRAF2→mTOR→TNFAlpha→P21→P300→P300-SP1→P53. The path followed by the nodes of the component 2 to reach its starting node HNF1Alpha is HNF1Alpha→HNF4Alpha→GCKR→NOTCH→CAMK1D→CDC123→Ngn3→DGKB→SREBP-1C→MTNR1B→HNF4Alpha→HNF1Alpha. The path followed by the nodes of the component 3 to reach its starting node HNF1Beta is HNF1Beta→HNF4→HNF1Alpha→Ngn3→PDX1→Hes1→Mafa→PDX1→Ngn3→NeuroD→Sox9→HNF6→HNF1Beta. The path followed by the nodes of the component 4 to reach its starting node INSR is INSR→INS→Foxa1/2→IRS→P13K→PIP3→PDK1/2→Akt→Foxa1/2→PyK→GK→G6PC→FBP→PEPCK→FAS→Akt→Foxa1/2→IRS→Akt→INS→INSR. The path followed by the nodes of the component 5 to reach its starting node INS is INS→INSR→IRS→FABP→CGI-58→ATGL→AC→Pka→PL1N→AC→BAR→GS→Akt→PDE-3B→AC→BAR→TSH→TSHR→GS→AC→INSR→INS. The path followed by the nodes of the component 6 to reach its starting node IL-6 is IL-6→IL-6R→SOCS3→IRS1/2→P13K→Akt→GSK3→TNFA→TNFR1→NFkB→P13K→Akt→AdipoR→AMPK→GSK3→IL-6. The path followed by the nodes of the component 7 to reach its starting node GnRH is GnRH→Gq/11→PLCBeta→IP3R→PAA→PLA2→GS→GnRHR→GnRH.

P53 is capable to counter in metabolic anxieties. For example, glucose lack state is perceived through AMPK, which triggers p53 and elicits its transcriptional control [180]. P53 controls the statement of IGF-BP3 (insulin-like development factor-restricting protein 3) [181]. Metabolic stressors speak to a colossal hazard towards replication similarity power p53 to close down mitogenic flagging verbalized by the hindrance of IGF-1 [182]. P53 is known to quell SIRT1 articulation

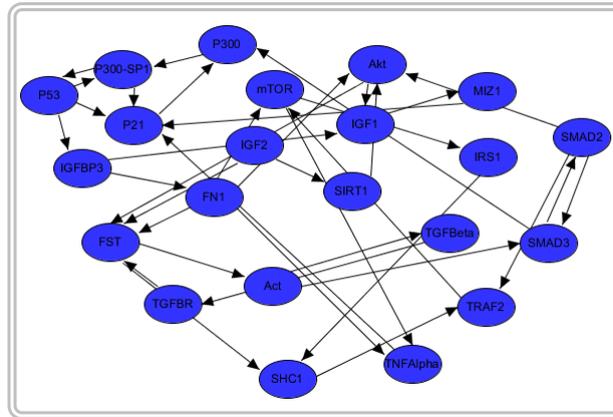


FIGURE 4.15: Extracted component 1 within the association network in which P53 act as a gene regulator and involve in the activation or inhibition of various other genes of different biological pathways.

that initiates insulin receptor substrate IRS1. SIRT1 is restrained by IGF1/2. The quality SHC1 encodes three primary protein isoforms p66SHC, P52SHC and p46SHC. Every one of the three SHC1 proteins share a similar area game plan comprising of a N-terminal phosphotyrosine-authoritative (PTB) space and a C-terminal Src-homology2 (SH2) area [183]. Tyrosine initiated IRS proteins contain the locales for flagging particles containing SH-2 (Src homolog2) space, for example, P13K, Grb2 and SHIP2 to tie with tyrosine deposits of theories proteins [166]. P13k that contains SH-2 area and ties to tyrosine residus of IRS proteins is initiated by TRAF2 flagging. The unthinking objective of mTOR has a primary capacity in the absorption of arranged physiological jolts to control various cell heightening and metabolic pathways [184]. mTOR for the most part assume a critical part as a reactant subunit in two fundamentally related yet practically unmistakable multi-segment kinase edifices, mTOR complex 1 (mTORC1) and mTORC2. Dysregulation of mTOR flagging is aligned with a scope of human infections, including metabolic disarranges and malignancy [185]. TRAF2 by implication impacts mTOR which is associated with the transcriptional control of TNFalpha. TNFAlpha enlistment was noted after P21 transfection. The statement of P21 is normally managed at transcriptional level and is described as an essential effector whose action is repressed by P53. P300 frames complex with P300-SP1 and associated with the actuation of P53 [186].

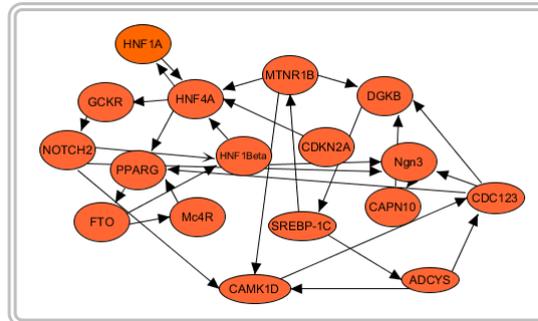


FIGURE 4.16: Extracted component 2 within the association network in which HNF1-Alpha act as a regulator and involve in the activation or inhibition of various other genes of different biological pathways.

Hepatocyte Nuclear Factor 1 (HNF1) and Hepatocyte Nuclear Factor 4 (HNF4) are two liver-enhanced transcription factors coexpressed in particular tissues where they assume a significant part through their inclusion in an unpredictable cross-administrative system. HNF1 down manages HNF4-intervened initiation of interpretation by means of a direct protein protein connection [187]. HNF4A is known to be an essential transcription factor for glucose and lipid homeostasis [188]. HNF4A builds the LGCK quality articulation and its coupling site HRE (HNF reaction component) is recognized in human [188]. Amid the fasting time frame, LGCK interpretation by HNF4A is quelled by FOXO1 which goes about as a corepressor, though the concealment is reestablished by encouraging where FOXO1 is phosphorylated and expelled to cytosol by insulin [188]. Initiation of the Notch flagging upregulated HNF1beta articulation, while it downregulated the outflow of HNF1alpha, HNF4 [189]. Indent relatives assume a part in an assortment of formative procedures by controlling cell destiny choices. The Notch flagging system is a developmentally preserved intercellular flagging pathway which manages associations between physically contiguous cells [190]. CAMK1D and CDC123 are engaged with the phone cycle division. CAMK1D and CDC123 diminishes the insulin discharge by influencing ngn3.

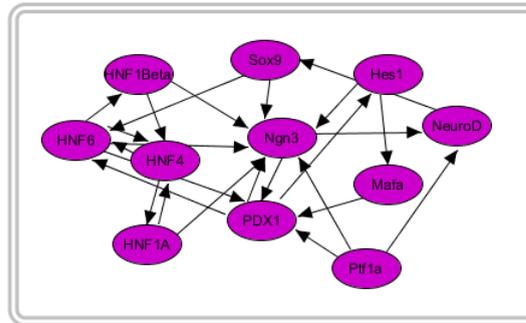


FIGURE 4.17: Extracted component 3 within the association network in which HNF1-Beta act as a regulator and involve in the activation or inhibition of various other genes of MODY, T2DM, Insulin secretion pathways [163].

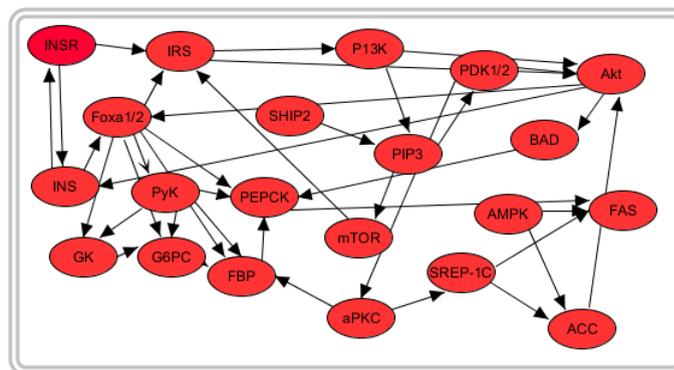


FIGURE 4.18: Extracted component 4 within the association network in which INSR act as a regulator and involve in the activation or inhibition of various other genes of Insulin secretion KEGG pathway, Insulin signaling KEGG pathway, Apoptosis KEGG pathway, Glycolysis/Gluconogenesis KEGG pathway [163].

In component 7, Gonadotropin releasing hormone (GnRH) emission follow up on its receptor GnRHR to direct the creation and arrival of gonadotropins. GnRHR is coupled to Gq/11 proteins to actuate phospholipase C beta PLCBeta which transmits its signs to IP3. IP3 on official with its receptor IP3R invigorates arrival of intracellular calcium. The expanded intracellular Ca (2+) levels and phosphorylation enacts the PLA2 quality. PLA2 quality encodes an individual from the cytosolic phospholipase A2 gathering. The chemical catalyzes the hydrolysis of layer phospholipids to discharge lipid-based cell hormones and is associated with the quality articulation and emission of GnRH. Gq/11 additionally couples with GS bringing about the actuation of quality articulation of gonadotropins through their receptors references: GnRH signaling pathway, Neuroactive ligand receptor

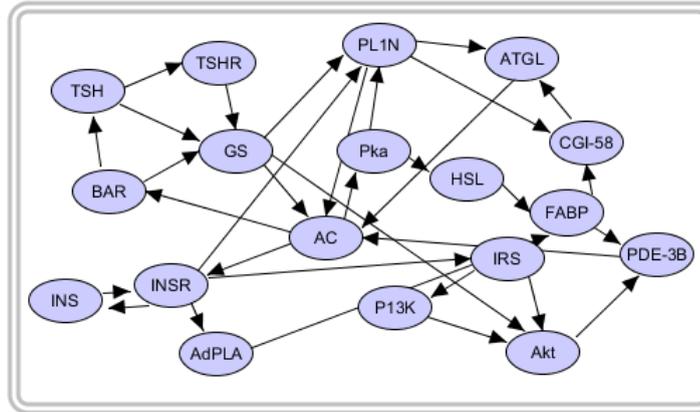


FIGURE 4.19: Extracted component 5 within the association network in which TSH act as a regulator and involve in the activation or inhibition of various other genes for the regulation of lipolysis in adipocytes KEGG Pathway, cGMP-PkG signaling KEGG Pathway, Insulin signaling KEGG pathway [163].

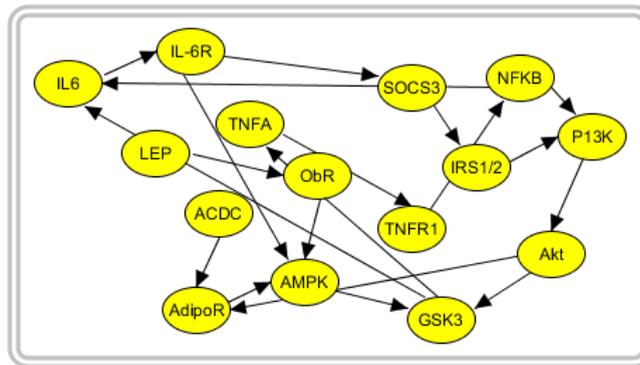


FIGURE 4.20: Extracted component 6 within the association network in which IL-6 act as a regulator and involve in the activation or inhibition of various other genes of Non-alcoholic fatty liver disease KEGG pathway, TNF signaling KEGG pathway, Adipo-cytokine signaling KEGG pathway, PPAR signaling KEGG pathway, P13K-AKT signaling KEGG pathway, Apoptosis KEGG pathway [163].

interaction: KEGG pathway, calcium signaling KEGG pathway, MAPK signaling KEGG pathway, Cytokine-cytokine receptor interaction:KEGG pathway. From the above mentioned identified components we had noticed that some genes were involved in more than one component. Table 4.6 shows the tabular representation of genes and in which components these genes are playing their functional role.

TABLE 4.6: Genes involved in identified components

S. No	Gene	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
1.	P53	✓						
2.	P300	✓						
3.	P21	✓						
4.	FN1	✓						
5.	Act	✓						
6.	TGFBeta	✓						
7.	TGFBR	✓						
8.	mTOR	✓						
9.	IGF1	✓						
10.	IGF2	✓						
11.	SIRT1	✓						
12.	TNFAlpha	✓					✓	
13.	Akt	✓			✓	✓	✓	
14.	MIZ1	✓						
15.	IRS1	✓			✓	✓	✓	
16.	SMAD2	✓						
17.	SMAD3	✓						
18.	TRAF2	✓						
19.	SHC1	✓						
20.	FST	✓						
21.	IGFBP3	✓						
22.	HNF1Alpha		✓	✓				
23.	HNF4Alpha		✓	✓				
24.	HNF1beta		✓	✓				
25.	CDKN2A		✓					
26.	NOTCH2		✓					
27.	FTO		✓					
28.	GCKR		✓					

S. No	Gene	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
29.	MTNR1B		✓					
30.	DGKB		✓					
31.	GLIS3		✓					
32.	CAPN10		✓					
33.	Mc4R		✓					
34.	PPARG		✓				✓	
35.	HNF6			✓				
36.	Ngn3			✓				
37.	HES1			✓				
38.	NEUROD			✓				
39.	Sox9			✓				
40.	PDX1			✓				
41.	MAFA			✓				
42.	Ptf1A			✓				
43.	Foxa1				✓			
44.	BAD				✓			
45.	Pyk				✓			
46.	GK				✓			
47.	C6PC				✓			
48.	FBP				✓			
49.	PEPCK				✓			
50.	ACC				✓	✓		
51.	FAS				✓			
52.	AMPK				✓		✓	
53.	aPKC				✓			
54.	SREBP-1C				✓			
55.	PDK1/2				✓			
56.	SH1P2				✓			
57.	INSR				✓	✓		
58.	TSH					✓		

S. No	Gene	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
59.	TSHR					✓		
60.	GS					✓		
61.	B-AR					✓		
62.	AC					✓		✓
63.	PKA				✓	✓		✓
64.	PKG					✓		
65.	PL1N					✓		
66.	ATGL					✓		
67.	HSL					✓		
68.	FABP					✓		
69.	P13					✓	✓	
70.	AdPLA					✓		
71.	PDE3B					✓		
72.	CGI-58					✓		
73.	IL-6						✓	
74.	IL-6R						✓	
75.	SOCS3						✓	
76.	TNFR1						✓	
77.	NFKB						✓	
78.	LEP						✓	
79.	ObR						✓	
80.	ACDC						✓	
81.	AdipoR						✓	
82.	GSK-3						✓	
83.	RXR						✓	
84.	ASK1						✓	
85.	JNK1						✓	
86.	AP-1						✓	
87.	C-JUN						✓	
88.	GnRH							✓

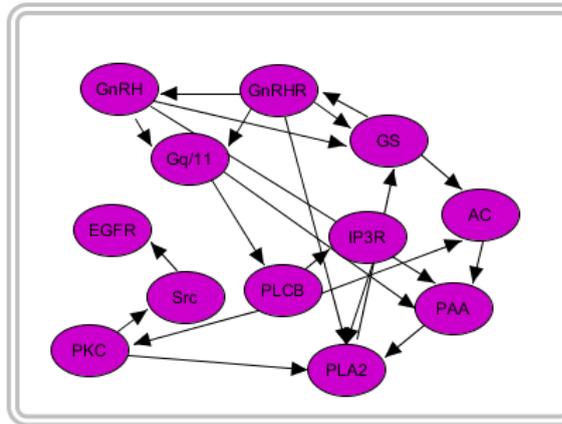


FIGURE 4.21: Extracted component 7 within the association network in which GnRH act as a regulator and involve in the activation or inhibition of various other genes [163].

S. No	Gene	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
89.	GnRHR							✓
90.	GS							✓
91.	Gq/11							✓
92.	PLCBeta							✓
93.	IP3R							✓
94.	EGFR							✓
95.	Src							✓
96.	Pkc							✓
97.	PLA2							✓

4.3.2 Identification of cell boundaries of components genes in a network

There are different types of cells in the human body which make up the tissues and organs. These cells are categorized on the basis of tissue formation and on the basis of their functions. The cells that are involved in tissue formation include bone cells, nerve cells, cartilage cells, epithelial cells, muscle cells, secretory cells, adipose cells, blood cells etc. Similarly cells are categorized on the basis of the functions they perform. Some of the examples of cells are conductive cells,

connective cells, glandular cells, storage cells, supportive cells, sperms, oocytes, stem cells etc. Each organ in a human body contains specific cells that have specific function to perform within that organ. usually three sorts of cells present in pancreas are named as acinar cells, islets of langerhans, pancreatic setllate cells [191]. Pancreas is different from other glands in that it has both endocrine and exocrine gland. Acinar cells carry out exocrine function and produce the enzymes involved in digestion and along with this perform the function of their secretion into intestine. Islets of langerhans, on the other hand, due to endocrine function are involved in the perpetuation of glucose level in blood and growth [192]. Pancreatic setllate cells assist to renovate the offended fractions of pancreas. They also facilitate demolition of tumor cells. Skin has fibroblasts, keratinocytes, langerhans, melanocytes and epidermal cells. Like these, each organ has specific types of cells. Different genes show their protein and RNA expressions in different cells of the organs. INS gene shows its RNA expression in tissue enriched (Pancreas) and has highly selective cytoplasmic expression in pancreatic Islets [193], [164]. Insulin receptor (INSR) shows its RNA expression in brain, pancreas, liver, bone marrow, muscle tissue, lungs etc and have Ubiquitous cytoplasmic expression. In our research we have identified the cell boundaries of each gene that are involved in the extracted connected components within the association network. The details are collected from The Human Protein Atlas <https://www.proteinatlas.org/>. Figure 4.22 shows the screenshot of the retrieved information from The Human Protein Atlas. The file contains genes involved in seven components, genes description, chromosome cytoband, their localization whether secreted, intracellular membrane or membrane bounded genes/proteins, RNA and protein expression of the genes, and the sub-cellular location of the genes as different genes are expressed in different organelles of the cell. Figure 4.23 shows the expression of genes in different cells of human organs.

Gene Name	Gene Description	Chromosome	Localization	RNA Expression	Protein expression	Subcellular location	Gene
1 INS	Insulin	11p15.5	Secreted	Tissue enriched (Pancreas)	Highly selective cytoplasmic expression in pancreatic islets.		INS
2 INSR	Insulin receptor	19 p13.2	Intracellular,Membrane	Expressed in all (Brain, pancreas, liver, Endocrine tissues,	Ubiquitous cytoplasmic expression.	Vesicles	INSR
4 IRS1	Insulin receptor substrate 1	2q36.3	Intracellular	Expressed in all (Brain, pancreas, liver, Endocrine tissues,	Cytoplasmic and/or nuclear expression in most tissues.	Nucleoplasm, cytosol	IRS1
5 IRS2	Insulin receptor substrate 2	13q34	Intracellular	Expressed in all (Brain, pancreas, liver, Endocrine tissues,	General cytoplasmic expression.	Aggresome, cytosol	IRS2
6 P300	E1A binding protein p300	22q13.2	Intracellular	Expressed in all (Brain, pancreas, liver, Endocrine tissues,	Nuclear and cytoplasmic expression at variable levels in most tissues.	Nucleoplasm	P300
7 P53	Tumor protein p53	17 p13.1	Intracellular	Expressed in all (Brain, pancreas, liver, Endocrine tissues,	weak nuclear and/or cytoplasmic expression	Nucleoplasm	P53
8 P21	ODKN1A (Cyclin-dependent kinase inhibitor 1A)	6p21.2	Intracellular	Expressed in all (Brain, pancreas, liver, Endocrine tissues,	Nuclear expression in fractions of cells in most tissues.	Nucleoplasm, nucleolus	P21
9 IGFBP3	Insulin-like growth factor bindin	7p12.3	Intracellular,Secreted	Tissue enriched (Placenta)	Placenta	Nucleoplasm, vesicle	IGFBP3
10 ADAMT9	ADAM metalloproteinase with	3p14.1	Intracellular,Secreted	Tissue enriched (Placenta)	Cytoplasmic expression	Endoplasmic reticulum	ADAMT9
11 THADA	Thyroid adenoma associated	2 p21	Intracellular	Expressed in all (Brain, pancreas, liver, Endocrine tissues,	Cytoplasmic expression	Cytosol	THADA
12 TGFBI	Transforming growth factor, be	19q13.2	Intracellular,Secreted	Expressed in all (Brain, pancreas, liver, Endocrine tissues,	Extracellular deposits and cytoplasmic expression (Bone	Cytosol	TGFBI
13 TGFBI2	Transforming growth factor,	1q41	Secreted	Tissue enriched (Prostate)			TGFBI2
14 TGFBI1	Transforming growth factor 9q22.33		Intracellular,Membrane	Expressed in all (Brain, pancreas, liver, Endocrine tissues,	Cytoplasmic expression	Vesicles, plasma membrane	TGFBI1
15 TGFBI2	Transforming growth factor, be	3p24.1	Membrane	Expressed in all (Brain, pancreas, liver, Endocrine tissues,	bone marrow, muscle tissues, lungs, kidney, etc)	Plasma membrane	TGFBI2

FIGURE 4.22: Details retrieved from The Human Protein Atlas against each components gene [124]

Cell Type	INS	INSR	IRS1	P13K	IRS2	IRS3	IRS4	P300	P53	P300-SP1	P21	IGFBP3	ADAMT9	THADA	TGFBI	TGFBI2	TGFBI1	TGFBI2
1 Human Cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey
2 Glial cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey
3 Neuronal cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey
4 Endothelial cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey
5 Neuropil	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey
6 Cells in granular layer	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey
7 Cells in molecular layer	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey
8 Purkinji cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey
9 Glandular cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey	Grey
10 Hematopoietic cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Pink	Grey	Grey	Grey	Grey	Grey	Grey	Grey
11 Germinal center cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Pink	Grey	Grey	Grey	Grey	Grey	Grey	Grey
12 Non-Germinal center cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Pink	Grey	Grey	Grey	Grey	Grey	Grey	Grey
13 Squamous epithelial cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Pink	Grey	Grey	Grey	Grey	Grey	Grey	Grey
14 Cells in red pulp	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Pink	Grey	Grey	Grey	Grey	Grey	Grey	Grey
15 Cells in white pulp	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Pink	Grey	Grey	Grey	Grey	Grey	Grey	Grey
16 Myocytes	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Pink	Grey	Grey	Grey	Grey	Grey	Grey	Grey
17 Smooth muscle cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Pink	Grey	Grey	Grey	Grey	Grey	Grey	Grey
18 Respiratory epithelial cells	Orange	Orange	Blue	Grey	Green	Grey	Grey	Purple	Red	Grey	Pink	Grey	Grey	Grey	Grey	Grey	Grey	Grey

FIGURE 4.23: Expression of genes in different human cells

4.3.3 Analysis of functional T2DM pathway

The Pancreas that is located in the abdomen are assumed to play a fundamental role in converting the diet we take into energy. This energy is used by bodys cells to perform different functions. It has two primary tasks: an exocrine portion of pancreas aids in assimilation and an endocrine portion direct glucose in blood or in tissues. Endocrine pancreatic cells cover the surface of pancreas. Endocrine part of pancreas has small bundles of cells so also known as islets of langerhans from where several capillaries run through each islet to transmit hormone to other parts of the body. Pancreas releases two antagonistic hormones, glucagon and insulin

in order to control blood sugar. The pancreatic alpha cells release glucagon and beta cells of the pancreas releases insulin. If the blood sugar level is high then insulin is released to make it proficient to pass on sugar from blood into cells of the body where it can be transformed into the energy. Insulin not only permists liver and the muscles to accumulate additional sugar but also keep the liver to produce more sugar. This has outcome of reducing blood glucose level. If blood sugar level is squat then glucagon is released into the blood stream. In response to this, liver cells of the liver liberate stored sugar and translate proteins into sugar. Active insulin molecule consists of alpha and beta subunits, bonded by disulfide bond. Insulin is entailed in development and growth of tissues and manages glucose homeostasis by invigorating glucose transport in muscles and adipose cells.

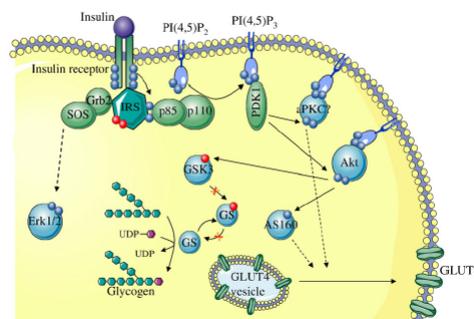


FIGURE 4.24: Schematic depiction of the insulin signal transduction cascade. [194]

Normal insulin secretion occurs through activation of specific insulin receptors (tyrosine kinases). The activity of insulin is started by binding to its related receptor and actuation of the receptor's inborn protein tyrosine kinase action, bringing about the phosphorylation of tyrosine deposits situated in the cytoplasmic face [194]. The initiated receptor, thusly, selects and phosphorylates a board of substrate molecules. Among these, IRS1 and IRS2 seem, by all accounts, to be the connector molecules assuming a noteworthy part in the coupling to the PI3K-PKB and MAPK downstream kinases [195]. Tyrosine phosphorylated IRS1/2 enlist the heterodimeric p85/p110 PI3K at the plasma layer, where it delivers the lipid second messenger PIP3, which thusly enacts a serine/threonine phosphorylation course of PH-domain containing proteins [196]. PIP3 targets incorporate PDK1, the serine/threonine protein kinase B (PKB)/Akt, and the atypical protein

kinases [197], [198]. Mechanistically, PDK1, PKB and aPKCs, which all contain a PH space, are enrolled at the plasma membrane by coupling to PIP3; consequently, PDK1 phosphorylates PKB and aPKCs on a threonine deposit situated in the initiation circle of the reactant area, causing their enactment [199], [200], [201], [202]. Main focuses of actuated PKB are GSK-3 [203] and AS160 [204]. Upon PKB-intervened phosphorylation on Ser9, GSK-3 is inactivated [205]. This inactivation, in parallel to protein phosphatase-1 (PP1) initiation, mitigates the inhibitory phosphorylation of GS, which ends up enacted and advances glycogen synthesis [206]. PKB likewise directs the insulin-stimulated translocation of the glucose transporter GLUT-4 at the plasma layer, bringing about expanded glucose take-up. This pathway includes an inhibitory phosphorylation of the RabGTPase actuating protein AS160. Restraint of AS160 favors the GTP-stacked territory of Rab and mitigates an inhibitory impact towards GLUT-4 translocation from intracellular compartments to the plasma layer [207]. Notwithstanding the part of PKB in controlling GLUT-4 translocation, aPKCs act in parallel to or can even be substitutive for PKB [208]. On a parallel pathway, initiated IRS1/2 select Grb2, which partners to SOS and enacts the Erk1/2 MAPK pathway [209]. The p38 and JNK push actuated kinases whose initiation is principally reliant on stretch signs and provocative cytokines [210], [211] have similarly been appeared to be phosphorylated/enacted in response to insulin; in spite of the fact that the pathway prompting their initiation has not yet been completely illustrated. Generally, adjustments of the actuation status of the proximal insulin flagging compounds (IR, IRS1/2, PI3K), and downstream targets (PDK, PKB and its objectives GSK-3 and AS160, aPKCs, and MAPK-family protein kinases) have been considered in muscle and fat tissue from opposing insulin, obese and diabetic subjects, and the basic insulin obstruction has been ascribed to absconds in at least one stages of the insulin transduction course.

4.3.4 Analysis of components pathways

In modern era, pathway examination has turned into the principal decision for extricating and clarifying the basic science for high-throughput molecular estimations. One of the successful biological advancements to deal with recognizing pathway collaboration is through hereditary screenings, in which manufactured lethality frequently shows cooperation between at least two pathways while genes engaged with two pathways reside independently [193]. Given the intricate idea of biological frameworks, pathways frequently need to work in an organized manner to create apt physiological reactions to internal and external boosts. Luckily, foundation pathway cross-talk network gives a quantifiable portrayal of the molecular systems that describe the perplexing cooperations and the mind boggling interwoven connections that represent cell functions, among those tissues and illness related genes to clarify the molecular procedures amid malady improvement and movement [212]. In networks, 2 pathways are probably going to communicate with or impact each other (cross-talk) if essentially more protein associations are recognized between these 2 pathways than anticipated by chance or if expression of one gene in a pathway effects the expression of other gene in another pathway and both genes are found to be responsible for the development of a disease [213]. Therefore, in the present study, pathway cross-talk analysis was conducted based on the extracted connected components within the association network to identify the key pathways for T2DM, so as to better understand the pathogenesis of T2DM. All the components pathways were analyzed through Kyoto encyclopedia of genes and genomes (KEGG). The KEGG pathway database (<http://www.genome.jp/kegg>) is a collection of graphical diagrams (pathway maps) for the biochemical pathways. In this study, all human pathway data were analyzed from the KEGG pathway database, and a total of 27 pathways for the components genes were obtained. Table 4.8 shows the pathways and the corresponding components containing genes involved in these pathways.

TABLE 4.7: Pathways involved in extracted components

Pathway Name	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
Mapk signaling pathway	✓						✓
TGFBeta signaling pathway	✓						
P53 signaling pathway	✓						
Insulin signaling pathway	✓		✓	✓	✓	✓	
Age-rage signaling pathway	✓						
mTOR signaling pathway	✓						
Type II diabetes pathway	✓	✓	✓			✓	
MODY pathway		✓	✓				
Cell cycle pathway		✓					
Protein processing in endoplasmic reticulum		✓					
Insulin secretion		✓		✓			
Apoptosis pathway		✓		✓		✓	
Neuroactive ligand receptor interaction		✓					✓
Glycolysis/Glucone genesis				✓			
Lipid homeostasis				✓			
Regulation of lipolysis in adipocytes					✓		
cGMP-PkG signaling pathway					✓		

TABLE 4.8: Pathways involved in extracted components

Pathway Name	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7
cAMP signaling pathway			✓		✓		
Glycerolipid metabolism					✓		
Non-alcoholic fatty liver disease						✓	
TNF signaling pathway						✓	
Adipo cytokine signaling pathway						✓	
PPAR signaling pathway						✓	
P13-k-akt signaling pathway						✓	
GnRH signaling pathway							✓
Calcium signaling pathway							✓
Cytokine-cytokine receptor interaction							✓

4.3.5 Selection of the components involved in specific cells

In the above step, we have identified the cells in which components genes show their protein expression. Different proteins are expressed in different cells and different cells are present in different organs of our body. Each organ has specific function to perform. In case of T2DM, the cells of our concern are pancreatic cells and skeletal muscle cells. It is mentioned previously that the endocrine pancreatic cells release alpha and beta cells. Beta cells are responsible for the secretion of insulin. Similarly, skeletal muscles are involved in the large uptake of glucose and they are the important sites for the insulin resistance in T2DM. Due to this reason, we have selected these two particular cells and identified the expression of genes involved in extracted components in pancreatic endocrine cells and skeletal muscle cells. tables [4.9](#) to [4.15](#) show the gene expression of different components in these two specific cells.

TABLE 4.9: Protein expression of component 1's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].

Genes	Pancreatic endocrine cells	Skeletal muscle cells
P53		
P300	Yellow	Yellow
P21	Yellow	Blue
FN1	Blue	Yellow
Act	Blue	Blue
TGFB	Blue	Blue
TGFBR	Blue	Blue
mTOR	Yellow	Blue
IGF1	Blue	Blue
IGF2	Blue	Blue
SIRT1	Yellow	Yellow
TNFAlpha	Blue	Blue
Akt	Yellow	Yellow
MIZ1	Yellow	Yellow
IRS1	Yellow	Yellow
SMAD2	Yellow	Yellow
SMAD3	Yellow	Blue
TRAF2	Yellow	Yellow
SHC1	Yellow	Yellow
FST	Yellow	Yellow
IGFBP3	Blue	Blue

"Yellow color presenting the expression of genes in the corresponding cell and blue color presents that the genes are not showing protein expression in pancreatic or skeletal muscle cells."

TABLE 4.10: Protein expression of component 2's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].

Genes	Pancreatic endocrine cells	Skeletal muscle cells
HNF1Alpha	Yellow	Blue
HNF4Alpha	Yellow	Blue
HNF1Beta	Yellow	Blue
CDKN2A	Yellow	Blue
NOTCH	Yellow	Yellow
FTO	Yellow	Yellow
GCKR	Blue	Yellow
MTNR1B	Blue	Blue
DGKB	Blue	Blue
GLIS3	Blue	Blue
MC4R	Yellow	Yellow
PPARG	Blue	Blue

”Yellow color presenting the expression of genes in the corresponding cell and blue color presents that the genes are not showing protein expression in pancreatic or skeletal muscle cells.”

Yellow color presenting the expression of genes in the corresponding cell and blue color presents that the genes are not showing protein expression in pancreatic or skeletal muscle cells. There may be a possibility that its protein expression is still not mentioned in The Human Protein Atlas as the functional annotation of some genes are not available in the Human Protein Atlas. From this expression information, we analyzed that genes of component 3 and component 6 show their expression in only specific cells. The majority genes of all other components are expressing or not expressing in both pancreatic and skeletal muscle cells but in the genes of component 3 are only showing protein expression in pancreatic cells. Similarly, most of the genes of component 6 are showing their protein expression only in skeletal muscle cell. From this we have selected these two components for further analysis. As two components extracted from a gene-gene association

TABLE 4.11: Protein expression of component 3's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].

Genes	Pancreatic endocrine cells	Skeletal muscle cells
HNF6	Yellow	Blue
ngn3	Blue	Blue
Hes1	Blue	Blue
NeuroD	Yellow	Blue
Sox9	Blue	Blue
HNF4	Yellow	Blue
HNF1Alpha	Yellow	Blue
PDX1	Yellow	Blue
MAFA	Blue	Blue
Ptf1A	Blue	Blue
HNF1Beta	Yellow	Blue

”Yellow color presenting the expression of genes in the corresponding cell and blue color presents that the genes are not showing protein expression in pancreatic or skeletal muscle cells.”

network are expressing them in two different cells of different organs. This analyses further goes to the cross talk between the two pathways of component 3 and component 6. It would help to identify how feedback mechanism of one pathway is involved in the mechanism of other pathway that would help us to identify any disruption that leads towards T2DM.

TABLE 4.12: Protein expression of component 4's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].

Genes	Pancreatic endocrine cells	Skeletal muscle cells
IRS1	Yellow	Yellow
IRS2	Yellow	Yellow
Akt	Yellow	Yellow
Foxa1	Blue	Blue
BAD	Yellow	Blue
Pyk	Blue	Blue
GK	Blue	Blue
G6PC	Yellow	Yellow
FBP	Yellow	Blue
PEPCK	Blue	Blue
ACC	Yellow	Yellow
FAS	Yellow	Blue
AMPK	Yellow	Yellow
aPKC	Blue	Blue
SREBF-1C	Blue	Blue
PDK1	Yellow	Yellow
PDK2	Yellow	Yellow
SHIP2	Yellow	Yellow
INSR	Yellow	Yellow

”Yellow color presenting the expression of genes in the corresponding cell and blue color presents that the genes are not showing protein expression in pancreatic or skeletal muscle cells.”

TABLE 4.13: Protein expression of component 5's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].

Genes	Pancreatic endocrine cells	Skeletal muscle cells
TSH	Blue	Blue
TSHR	Blue	Blue
GS	Yellow	Blue
BAR	Blue	Blue
AC	Yellow	Yellow
PKA	Yellow	Yellow
PKG	Yellow	Yellow
PL1N	Yellow	Blue
ATGL	Yellow	Yellow
HSL	Blue	Blue
FABP	Blue	Blue
INSR	Blue	Blue
IRS1	Yellow	Yellow
IRS2	Yellow	Yellow
P13k	Yellow	Yellow
Akt	Yellow	Yellow
AdplA	Yellow	Blue
PDE-3B	Blue	Blue
CGI-58	Yellow	Blue

”Yellow color presenting the expression of genes in the corresponding cell and blue color presents that the genes are not showing protein expression in pancreatic or skeletal muscle cells.”

TABLE 4.14: Protein expression of component 6's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].

Genes	Pancreatic endocrine cells	Skeletal muscle cells
IL-6	Blue	Yellow
IL-6R	Blue	Yellow
SOCS3	Blue	Blue
TNFAlpha	Blue	Blue
TNFR1	Yellow	Blue
NFKB	Yellow	Yellow
P13k	Blue	Yellow
Akt	Yellow	Yellow
GSK3	Yellow	Yellow
LEP	Blue	Blue
obR	Blue	Blue
ACDC	Blue	Yellow
AdipoR	Yellow	Yellow
AMPK	Yellow	Blue
PPAR-Aplha	Blue	Yellow
RXR	Blue	Yellow
JNK1	Blue	Yellow
JNK2	Blue	Yellow
ASK1	Blue	Yellow
C-Jun	Blue	Yellow

”Yellow color presenting the expression of genes in the corresponding cell and blue color presents that the genes are not showing protein expression in pancreatic or skeletal muscle cells.”

TABLE 4.15: Protein expression of component 7's genes in pancreatic and skeletal muscle cells based on the information obtained from Human Atlas Protein database [124].

Genes	Pancreatic endocrine cells	Skeletal muscle cells
GnRH	Yellow	Blue
GnRHR	Blue	Blue
GS/GLUL	Yellow	Blue
AC	Yellow	Yellow
PKA	Yellow	Yellow
Gq/11	Yellow	Blue
PLCB	Yellow	Yellow
IP3R	Yellow	Blue
PKC	Blue	Blue
PLA2	Blue	Blue
EGFR	Blue	Blue
SRC	Yellow	Blue

”Yellow color presenting the expression of genes in the corresponding cell and blue color presents that the genes are not showing protein expression in pancreatic or skeletal muscle cells.”

4.3.6 Identification of disrupted pathway

In this research, we have identified how two pathway crosstalk to each other in the pathogenesis of T2DM. The genes of both component 3 and component 6 are involved in different KEGG pathways. Genes belong to component 3, HNF1beta, HNF6, HNF4, HNF1Alpha, NeuroD, Hes1, Ngn3, Mafa, PDX1 are involved in Maturity onset diabetes of the young (MODY) pathway and Type II diabetes pathway. Gene sox9 of component 3 is involved in cAMP signaling pathway. Ptf1a gene of the same component is involved in the insulin signaling pathway. Genes of component 6 are involved in insulin signaling pathway, Type II diabetes pathway, apoptosis pathway, Non alcoholic fatty liver disease, TNF signaling pathway,

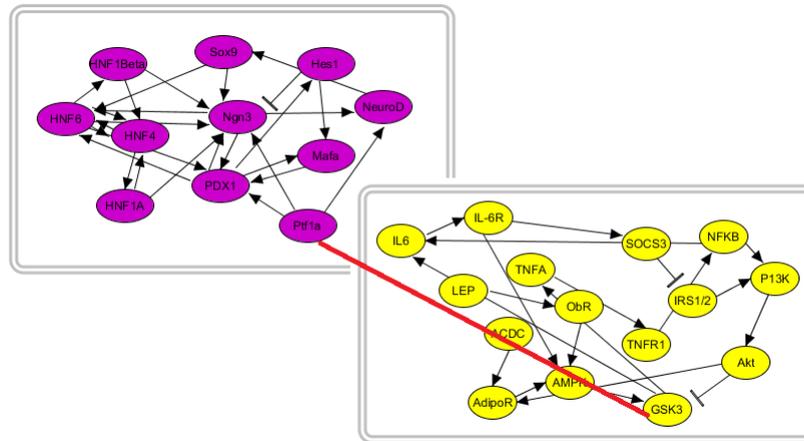


FIGURE 4.25: Pathway crosstalk between component 3 and component 6

PPAR signaling pathway, Adipo cytokine signaling pathway and P13k-Akt signaling pathway. In the analysis of normal functional pathway of T2DM, we have INS activates IRS1/2 and involves in the increased release of glucagon that transfer glucose into the blood in case of low blood glucose concentration. If a person has T2DM, it means pancreas is not releasing enough insulin to allow the fat cells to take the glucose away from the blood. If phosphorylation of IRS1/2 that occurs due to the activated insulin receptor, is prohibited by any feedback mechanism then synthesis of glucagon can be decreased. Usually insulin that is involved in the growth and development of tissues and control of glucose homeostasis by stimulating the glucose transport into muscle and adipose cells is secreted by Beta cells of pancreas. Active insulin molecule consists of two chains of alpha and beta units that are held together by disulfide bond. Sufficient insulin secretion happens by the activation of a particular insulin receptor, tyrosine kinases. Insulin binds alpha subunit of receptor and causes the inhibition of tyrosine auto phosphorylation by the Beta sub unit. Activated insulin receptor phosphorylates insulin receptor substrates (IRS1-4) that are members of IRS family. IRS1/2 is most important for glucose transport. Tyrosine activated IRS proteins serves as a binding sites for those molecules that regulates signal transduction like one's containing SH-2 (Src homolog 2) domain such as P13K, GrB2, SHIP2 that binds to tyrosine residues of IRS proteins. P13K is the chief signal moderator of metabolic and mitogenic acts of insulin. P13K forms signaling complex to mediate downstream signaling. This

complex is composed of p85 subunit that binds to IRS protein and a p110 catalytic subunit. P13k-IRS1/2 increases the catalytic activity of p110 that allows phosphorylation of its substrate PtfIns P2 to generate P3 and recruits PDK1, PKB/Akt and PKC to plasma membrane via PH domain. PIP2 and PIP3 are phospholipid components of cell membrane; they are enriched at plasma membrane and act as substrate for many important signaling proteins. These kinases on activation results in insulin responses, like, GluT4 translocation, glucagon synthesis of GSK3 and lipogenesis by upregulation of fatty acid synthase gene. In case of T2DM subjects, beta cells of pancreas are not releasing enough insulin to transport glucose from blood to tissues and the synthesis of glucagon causes the transport of glucose from tissues to blood due to which blood glucose concentration is increasing constantly. Any alteration in the normal signaling cascade of insulin pathway to the synthesis of glucagon can stop the synthesis of glucagon. Normal insulin signaling cascade start from the activation of IRS1/2 protein, thus the inhibition of these proteins by some other molecule can alter the whole normal pathway that is responsible for the glucagon production. In component 6 SOCS3 that is involved in TNF signaling pathway, inhibits the IRS1/2 that are involved in P13K signaling pathway and insulin signaling pathway and type II diabetes pathway. Similarly Akt that is involved in P13K-Akt signaling pathway and TNF signaling pathway inhibits the GSK3 that is involved in the non alcoholic fatty liver disease pathway. Inhibition of IRS1/2 and GSK3 activity can stops the glucagon synthesis in the normal insulin signaling transduction pathway. GSK3 activates the Ptf1a that is found to be involved in component 3 in our generated results. Ptf1a is involved in insulin signaling pathway indirectly through the regulation of beta cells development organism specific bio system. Inhibition of GSK3 prohibits the activity of Ptf1a and beta cells development stops that release insulin to transport glucose from blood to tissues and causes T2DM. In this way these two components were found to crosstalk with each other to affect the normal insulin signaling pathway in insulin resistant and type 2 diabetes subjects. Figure 4.25 shows the pathway crosstalk of component 2 and component 6.

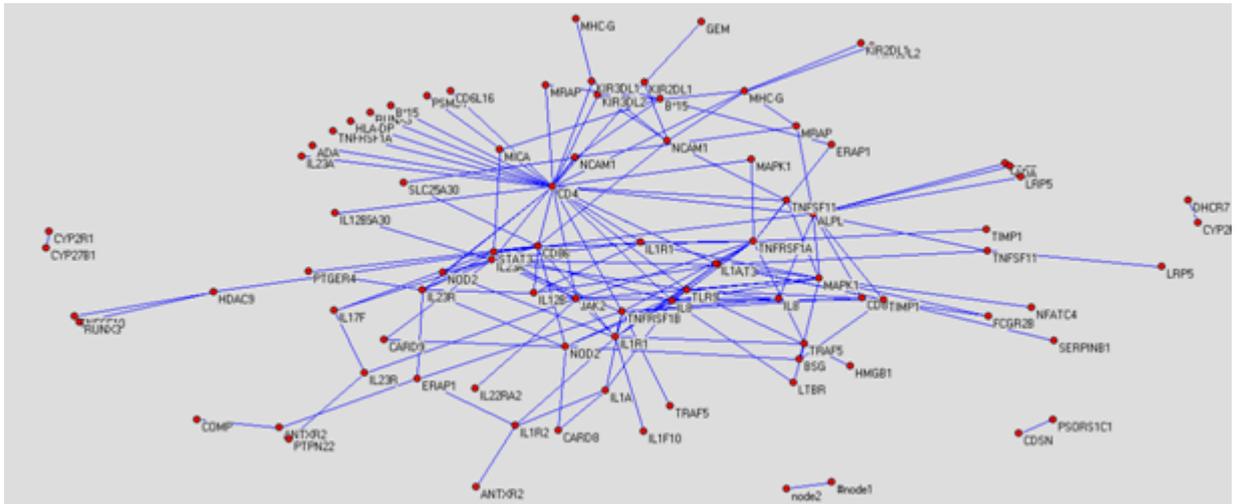


FIGURE 4.27: Extended network includes one giant network and its small components based on Energy level

(if there should arise an occurrence of single-protein input). Grey hubs are proteins associated with our info or second shell of interactors for various and single information individually. The first shell interactors are the proteins legitimately connected with your information protein(s). second shell of interactors are the proteins related with the proteins from the first shell or with our input protein(s). It can happen that a second shell protein can be straightforwardly associated with our input protein(s), however it will normally have a flimsier affiliation and in this way it would not appear among the predetermined number of the first shell interactors. We can perceive which shell the protein has a place with by taking a gander at the shade of the air pocket, as second shell proteins are constantly Grey. The amplified system examined from STRING database in the Pajek programming is indicated in Figure 4.28. The amplified system in Pajek programming on the premise of Fruchterman Reingold vitality incorporates one titan system and four isolated little systems which are gotten individually from the seed protein demonstrated in Figure 4.27. The giant network comprised of 93 nodes associated by means of very nearly 199 edges 4.28. The spine system comprised of 15 nodes joined by means of just about 45 edges 4.29. In the same manner, we examined the estimations charactering system recorded in Table 4.16: number of nodes (N), average degree ($\langle k \rangle$), diameter (D) and longest most limited way length (lsp1).

TABLE 4.16: The general network measurements for networks

S.No	Description	Giant network	Backbone network
1.	Number of Nodes	118	15
2.	Number of Edges	1157	10.6
3.	Average Node degree	19.6	2
4.	Average local clustering coefficient	LRP5 to COMP	node1 to IL23R

4.4.1 Key nodes in the PPI network:

This study partitioned the network on the basis of degree of each node. Figure 4.30 shows the network which is partitioned on the basis of level of degree against each node. Different colors represent different level of degree. The nodes with large degree or high BC were viewed as key nodes, and 5% of the total nodes set of the network was used as the critical point of large degree and high BC nodes. Of the total nodes, 31 nodes have high BC at the threshold of 0.01 (Table A.205), 55 nodes have large degree at the threshold of 1 (Table A.208) and 24 nodes were selected with high BC and large degree. In order to discern their roles in the network, these nodes were highlighted in different color and size (Figure 4.31). CD4 is a hub protein with the largest degree and also a bottleneck protein with the highest BC. CD4 has highest CC value, which indicates that it locates at the centre of the network. Different colors represent different levels of degrees. Nodes with same colors have same degree. Nodes with yellow color have the lowest degree. Greyish node representing CD4 have the highest degree. Grey color with large size node shows the gene with large BC value and large degree. Similarly yellow color with small size shows the genes with lowest BC values and degree for those genes. The signal transduction pathway of high BC proteins and link-

TABLE 4.17: The list of genes extracted from literary database showing association with essential Ankylosing spondylitis [123], [124], [126], [159]

S.No.	Symbol	Description
1.	B27	Human Leukocyte Antigen (HLA) B27
2.	TNF alpha	tumor necrosis factor alpha
3.	C reactive protein	C reactive Protein
4.	ERAP1	Endoplasmic reticulum amino peptidase 1
5.	IL23R	Interleukin-23 Receptor
6.	IL17	Interleukin-17
7.	TNFR1	Tumor-necrosis factor receptor-1
8.	IL 1beta	Interleukin-1 beta
9.	IL12B	Interleukin-12 beta
10.	IL 33	Interleukin-33
11.	STAT3	Signal transducer and activator of transcription 3
12.	JARID1A	Jumonji/ARID domain-containing protein 1A
13.	IL 10	Interleukin-10
14.	ANKH	Human homolog of the murine progressive ankylosis gene
15.	IL 6	Interleukin-6
16.	ANTXR2	Anthrax toxin receptor 2
17.	CIMT	Carotid Intimal Medial Thickness
18.	TBKBP1	TBK1 binding protein 1
19.	PTGER4	Prostaglandin E receptor 4
20.	IL 23	Interleukin-23
21.	TNAP	Tissue-nonspecific Alkaline Phosphatase
22.	ORMDL3	Orosomucoid like 3

S.No.	Symbol	Description
23.	PTPN22	Protein tyrosine phosphatase, non-receptor type 22 (lymphoid)
24.	RUNX3	Runt-related transcription factor 3
25.	KIF21B	Kinesin family member 21B
26.	FCRL4	Fc receptor-like protein 4
27.	ST2	Interleukin 1 receptor-like 1, also known as IL1RL1
28.	ADA	Adenosine deaminase
29.	CARD9	Caspase recruitment domain family, member 9
30.	KIR	Killer cell immunoglobulin-like receptors
31.	CD4	CD4 molecule, This gene encodes a membrane glycoprotein of T lymphocytes
32.	MICA	MHC class I polypeptide-related sequence A
33.	HLA B	Major histocompatibility complex, class I, B,
34.	KIR2DS5	Killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 5
35.	COMP	Cartilage oligomeric matrix protein
36.	SNAPC4	Mall nuclear RNA activating complex, polypeptide 4, 190kDa
37.	PPARGC1B	Peroxisome proliferator-activated receptor gamma, coactivator 1 beta
38.	LRP5	Low density lipoprotein receptor-related protein 5
39.	p38	Mitogen-activated protein kinases are a class of mitogen-activated protein kinases
40.	CD147	CD147/EMMPRIN (Extracellular Matrix Metalloproteinase Inducer), also known as Basigin (BSG)
41.	NOD2	Nucleotide-binding oligomerization domain containing 2
42.	TIMP	TIMP metalloproteinase inhibitor 1
43.	DPB1	HLA class II histocompatibility antigen, DP(W2) beta chain
44.	JAK2	Janus kinase 2
45.	CTLA 4	Cytotoxic T-lymphocyte-associated protein 4
46.	CD86	CD86 molecule

S.No.	Symbol	Description
47.	C3b	Complement component c3b
48.	IL1A	Interleukin-1 alpha
49.	CNOT3	CCR4-NOT transcription complex, subunit 3
50.	CD56	CD56 molecule
51.	TLR9	Toll-like receptor 9
52.	IL1R	Interleukin-1 receptor
53.	at V1	-
54.	CTCL	Cutaneous T-cell lymphoma-associated antigen 1
55.	IL1R2	Interleukin-1 receptor type 2
56.	FCGR2B	Fc fragment of IgG, low affinity IIb, receptor (CD32)
57.	IL 1F7	interleukin 1 family, member 7 (zeta)
58.	Pst 1	Protoplast secreted protein 1
59.	CDSN	Corneodesmosin
60.	PSORS1C1	Psoriasis susceptibility 1 candidate 1
61.	PSMD7	Proteasome (prosome, macropain) 26S subunit, non-ATPase, 7
62.	HDAC	Histone deacetylase 1
63.	HMGB1	High mobility group box 1
64.	RANKL	Receptor activator of nuclear factor kappa-B ligand
65.	CXCL16	Chemokine (C-X-C motif) ligand 16
66.	KIR3DL2	Killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 2
67.	IL 8	Interleukin-8
68.	TNFSF10	Tumor necrosis factor (ligand) superfamily, member 10
69.	IL 4	Interleukin-4
70.	TRAF5	TNF receptor-associated factor 5

S.No.	Symbol	Description
71.	IL 5	Interleukin-5
72.	IL 13	Interleukin-13
73.	DHCR7	7-dehydrocholesterol reductase
74.	ORAI1	ORAI calcium release-activated calcium modulator 1
75.	CARD8	Caspase recruitment domain family, member 8
76.	CYP27B1	Cytochrome P450, family 27, subfamily B, polypeptide 1
77.	LTBR	Lymphotoxin beta receptor (TNFR superfamily, member 3)
78.	CYP2R1	Cytochrome P450, family 2, subfamily R, polypeptide 1
79.	CYP24A1	Cytochrome P450, family 24, subfamily A, polypeptide 1
80.	gout 1	Gout 1
81.	LEI	-
82.	KIR3DL1	Killer cell immunoglobulin-like receptor, three domains, long cytoplasmic tail, 1
83.	IL 22	Interleukin-22
84.	P13	P13
85.	T12	-

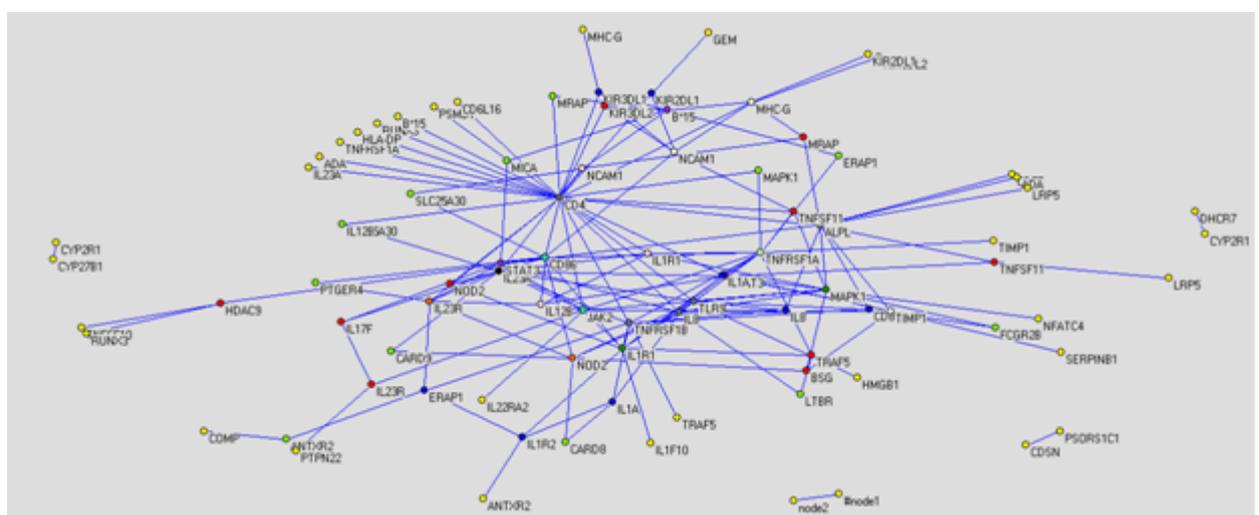


FIGURE 4.30: Partitioning of networks on the basis of degree of each node.

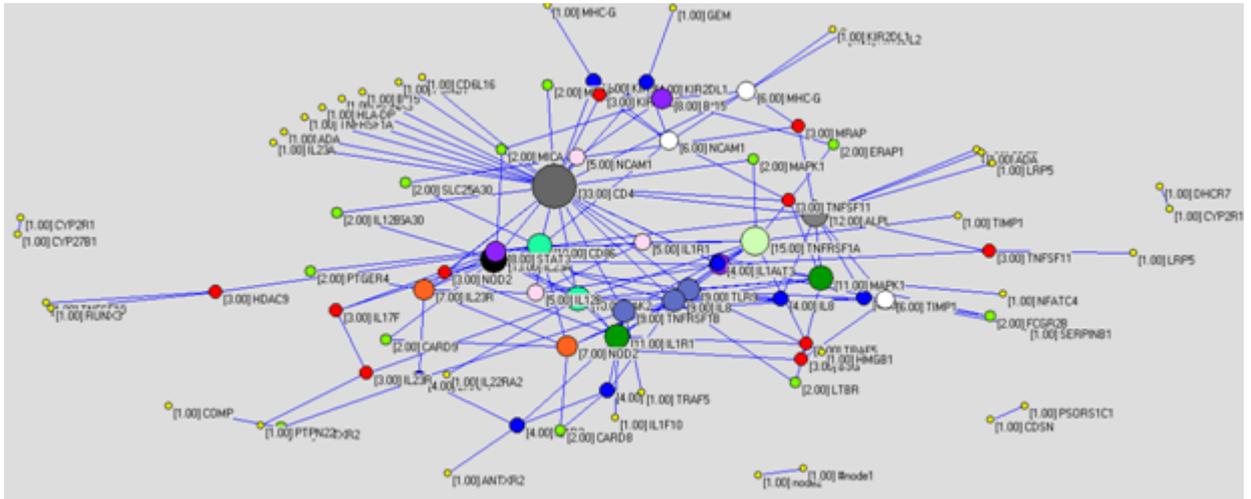


FIGURE 4.31: Vectors of the nodes on the basis of size to show the most required nodes in the network.

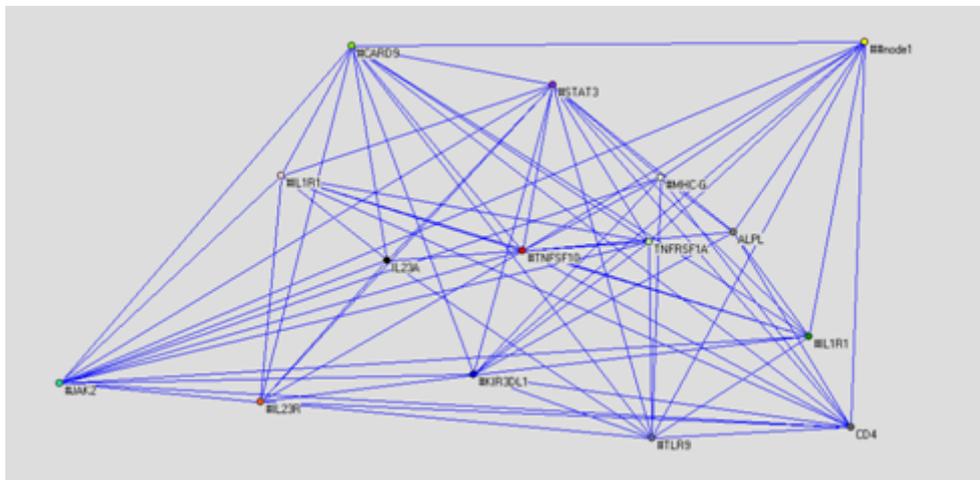


FIGURE 4.32: Topology of the backbone network

age/crosstalk between them is the resultant of backbone network that has from 15 high BC nodes whose size corresponds to their BC value and approximately 45 relations between them Figure 4.32. Devoid of manipulating the values of BC and CC, we can establish that CD4 situates at the centre of the backbone network with the highest BC value and the largest degree. CD4 has 33 neighbors: IL23A, ALPL, JAK2, CD86, TLR9, TNFRSF1B, STAT3, B*15, IL23R, NOD2, MHC-G, NCAM1, IL1R1, KIR3DL1, KIR2DL1, IL1A, IL8, TNFSF10, KIR3DL2, TNFSF11, MRAP, SLC25A30, IL12B, MAPK1, LTBR, TNFRSF1A, ADA, CXCL16,

HLA-DP, PSMD7, CD6, CD55, RUNX3. These proteins also represented immunological pathways, Natural killer cell mediated cytotoxicity system, cell signaling and development of nervous system and inflammatory system etc.

4.4.2 Structural holes and scale freeness of the network topology:

As discussed earlier that structural holes are imperative topological aspect of networks that are significant in the recognition of the function of nodes in the associations amongst sub clusters of graphs. Structural holes split surplus causes of information that are more preservative than overlapping. Figure 4.33 scrutinizes the spot of all nodes in their neighborhood for the incidence of structural holes. Numeral measures (most anticipated by Burt) that portray a variety of phases of the benefit or shortcoming of the nodes are also worked out. The scale freeness of the topology is confirmed by the clustering coefficient and degree of each node. Clustering coefficient of a node declines as degree of scrupulous node amplifies and circulation of clustering coefficient is an essential characteristic of scale-free networks which decreases as the node degree increases, thereby subsequent power law distribution. Scale free topology of the network generated by Pajek software is shown in Figure 4.34.

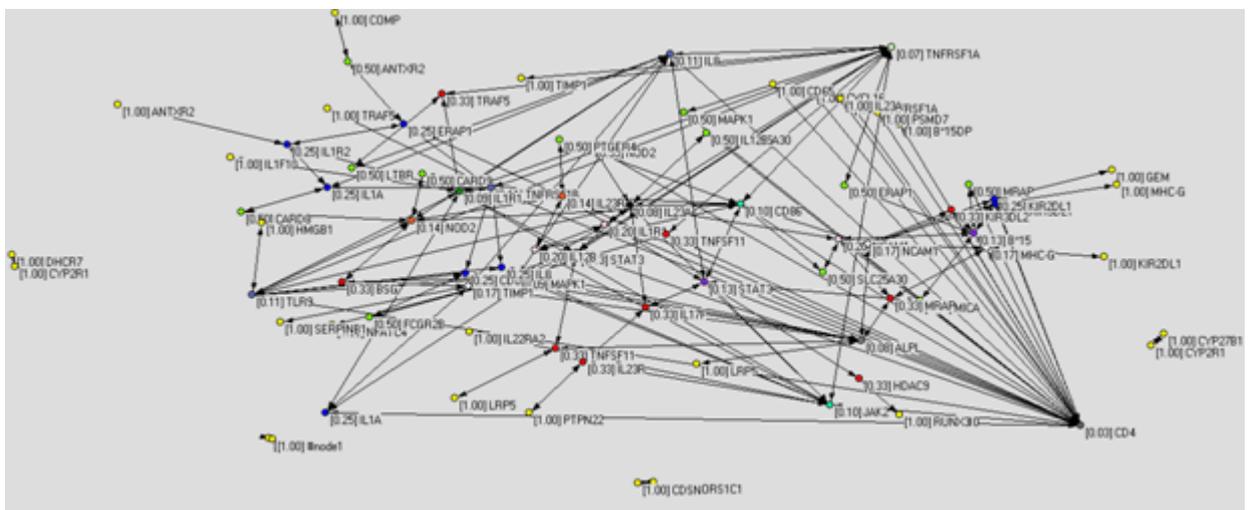


FIGURE 4.33: Structural holes in a network

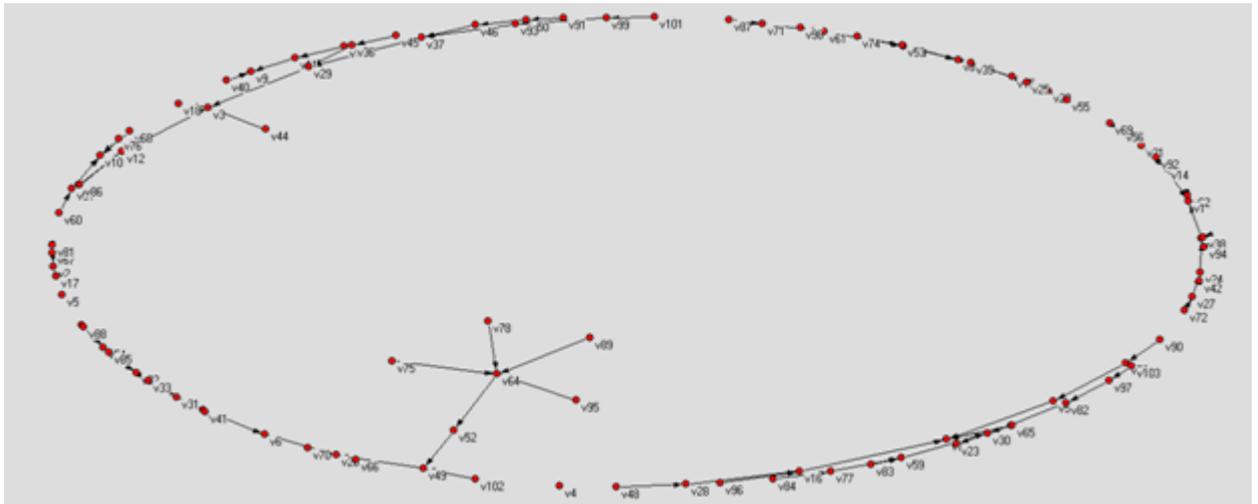


FIGURE 4.34: Scale freeness topology of a network

In the current research, total 86 have been found to be causative to AS. The network generated from the seed proteins have been further divided into one giant network and many small networks. Although there is a possibility that some genes may be missed during literature search as well as novel susceptible genes for AS. According to Gipsi Lima-Mendez and Jacques van Helden, biological networks are lenient to nodes removal, and innovative nodes favor to connect to nodes with large degree. In an additional word, biological networks are stout to arbitrary variation of nodes but perceptive to hub exclusion.

By setting the threshold of 1, 0.0.1, 55 proteins having large degree and 31 proteins having high BC were discovered. Among those, 24 proteins are those that are with both large degree and high BC. Usually in all protein-protein networks, there exist four kinds of proteins that disentangle the effects of betweenness and degree. Yu and his co-workers also alienated proteins into: non-hubnon-bottlenecks having small degree and low BC; hubnon-bottlenecks with large degree but low BC; non-hub bottlenecks that have small degree but high BC; and hubbottlenecks which have large degree and high BC. Hubbottlenecks proteins are also known to be date-hubs, on the other hand hubnon-bottlenecks are known to be party-hubs. Han et al. discerned two subtypes amongst the extremely associated proteins. Party hubs intermingle with majority of their partners concurrently, whereas date hubs connect diverse partners at different periods or positions. We alleged that

auxiliary corroborating the space-time outcome of these proteins, which will assist us to make out drug targets and biomarkers for AS. CD4 with the largest degree also has high BC in proteins list. The backbone network gives apparent illustration of all important genes and crosstalk between them. The main component of the whole network that is also recognized as giant network contains majority of seed proteins that are involved in AS along with their other PPI neighbors. The finding of this research suggests that AS disparity is coordinated by an incorporated PPI network having central hub CD4. Duftner C. et al., [214] demonstrated that circling CD3+CD4+CD28-cells were extended in the fringe blood of AS patients yet not in age-coordinated sound control people. The rates of CD4+CD28-T cells were plainly lower than those of CD8+CD28-T cells in AS patients.

Chapter 5

Conclusion and recommendations

5.1 Conclusion

In recent past, a number of explorations have been done in the literature of diabetic disease utilizing information mining strategies. Various systems with information revelation approaches [215][88][89] have been subjugated in some of these analysis. These advances help a great deal in biomedical science in the distinguishing proof of significant reason for diabetes at hereditary level. The subsequent stage to this is the manner by which the genes that are dependable to diabetes carry on, how they identifies with each other, how much affiliation they have with different genes, what influence they show to different genes and so on. The past work does not have this learning for diabetes.

In this research, data mining techniques have been applied to develop a system not only extracts data from online databases automatically (PubMed Info Extractor) by giving the specific query to the system but also finds the relevant useful information from that collected data. Automatic extraction of such information from printed information can altogether improve natural research efficiency by keeping scientists in the know regarding the cutting edge in their examination area, by assisting them envision organic pathways, and by producing liable innovative speculations regarding novel connections with the possibility that there may exists

some connections can be great contender to encourage natural research and approval. We found the association among GWAS known gene variants of disease of T2DM and the association of these genes with other known protein coding genes of humans chromosome 1 to chromosome 22. In order to accomplish our tasks for the current research, different methods and techniques had been applied for association extraction among genes, identification of shared functionality between and among the genes, connected components in the resulted network and topological analysis of that network and components for the particular disease. Based on the existing knowledge discovery methods, a system has been developed that has identified correlated genes along with their state of functionality on the basis of appropriate mined documents for T2DM. The most commonly used association mining rules such as confidence, support, lift, conviction has been used for the task of discovering associations among genes. The lexicon that contains all the promising relationship terms showing the associations among gene terms have been applied to all extorted sentences from the abstracts to determine the nature of relationship between two genes. Given a gene/protein interaction network, we found strongly connected components based on the generated association graph that may correspond to biological functional modules is today's need in order to understand complex systems and to investigate a system. The topological analysis of extracted connected components was done using Cytoscape and identified the hub proteins from the extorted components. Two case studies for Ankylosing spondylitis and Type II Diabetes Mellitus have been presented over here that indicate the usefulness of the current research. There are many computational approaches for understanding the hidden relationships among gene variants is very important and T2DM is not an exception either. These hidden patterns can also be identified using well know big data approach. The strongly connected components in a Type II Diabetes Mellitus gene network have been revealed by using big data driven approach. Using association rule mining techniques, a genetic network has been built for the gene variants of T2DM from literature database. In the generated network, strongly connected components (SCC), maximal strongly connected subgraphs are found. Due to efficient computations, Tarjans Algorithm

is used to find strongly connected components of a directed graph as it requires only one depth first search (DFS) traversal to implement this algorithm. Cell boundaries of the found gene components are also identified using information from Human protein Atlas. We have selected only those components that show highly expressed genes in the pancreatic endocrine cells and skeletal muscle cell according to T2DM for determining the disrupted pathway after analyzing the normal functional pathway of T2DM. From a network generated by using the association rule mining technique, seven strongly connected components have been identified. These components represent P53, HNF1Alpha, HNF1Beta, INSR, INS, IL-6 and GnRH as the regulators or initiators of seven different biological pathways. These biological pathways are originated to be allied with T2DM by big data approach used for this research. The path followed by the first connected graph is P53→IGFBP3→IGF1→IRS1→SHC1→TRAF2→mTOR→TNFAlpha→P21→P300→P3-SP1→P53. This sub-network is an evidence for the association of T2DM with the genes that are involved in cancer cell metabolism, growth regulation, proliferation control etc. Similarly these connected components demonstrate the association between different metabolic pathways, for example, insulin signaling pathway, mTOR pathway, MODY pathway, glycolysis, lipid homeostasis, Age-rage signaling pathway, MAPK pathway, p53 pathway. Self inhibition of ngn3 is also acknowledged in these components. In diabetic patients, pancreatic islets in case of fasting lessen PKA and mTOR activity and induce Sox2 and Ngn3 expression and insulin production. Self inhibition of Ngn3 can therefore affect the insulin production. The results suggest that genes responsible for T2DM are involved in several pathways such as MAPK, mTOR and p53 signaling pathways. The obtained network through data mining predicts the relationship of T2DM with other diseases through crosstalks. As the common diseases with high incidence, Type 2 diabetes mellitus and Ankylosing spondylitis gains much attention among researchers and has a rather large literature accumulation. Type 2 diabetes and Ankylosing spondylitis have been used as testing disease for system evaluation. The significant findings of the current research are: (i) Development of automated data extraction tool PMIE, with the target of giving a platform

to information extraction and mining helpful data from the extracted information from the biological database PubMed. It adds to the abilities of the current apparatuses and servers for information extraction and analysis. (ii) biological association network that shows specific relations among T2DM known gene variants from GWAS catalog and with other known protein coding genes of human chromosome 1 to chromosome 22. The developed tool also identified the transitive relation between genes along with nature of relationship between the associated genes. Self inhibition of *ngn3* has also been identified (iii) Seven connected components from the biological networks have been identified that show involvement of gene variants of T2DM with different other pathways. (iv) Topological analysis of network for AS and identified hub protein, degree, BC and CC value of each gene involved in AS gene-gene interaction network. The finding suggests that AS disparity is coordinated by an incorporated PPI network having central hub CD4. Circling CD3+CD4+CD28-cells were extended in the fringe blood of AS patients yet not in age-coordinated sound control people. The rates of CD4+CD28-T cells were plainly lower than those of CD8+CD28-T cells in AS patients. Designing of biological networking models will help researchers in unfolding various missing links between the possible interactions. It will help in further resolution of already available data with finer results.

5.2 Future work

As it has been mentioned that two case studies Ankylosing spondylitis and Type II Diabetes have been used to evaluate the developed system. We could not proceed our research in the area of Ankylosing Spondylitis because it has been beyond the scope and very complicated for the current PhD thesis. In future we are planning to use the system on Ankylosing Spondylitis for finding the gene-gene association and even gene disease association. Ankylosing Spondylitis is a well known type of Arthritis which is a common disease in modern world and its specific treatment is still not available. Such type of information as we discovered for Type II Diabetes by using the developed system can help the biologists and researchers

to find appropriate treatment for the disease. The concept of big data has been around for years. Big data analysis approaches can play an imperative part in health research. This can be a helpful asset for diabetes researchers because it can reveal veiled acquaintance from a massive sum of diabetes-related data. The growing availability of genome-wide expression data for T2DM has enabled a big data approach to identify molecular markers by finding robust statistical associations between genes and identifying key regulators in a large network using graph theory. There are many computational approaches for understanding the hidden relationships among gene variants but T2DM is not an exception either. These hidden patterns can also be identified using well know big data approach. We feel that enormous information utilizing should be financially savvy and spotlight on customized drug. In future the ebb and flow research can be stretched out toward customized drug as up till now we have an unmistakable picture of gene-gene co-operation organize, the connection between and among genes, transitive relations among genes and the association of genes as parts.

Bibliography

- [1] D. Benson and I. Karsch-Mizrachi, “Genbank.” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 15–18, 2000.
- [2] A. Driscoll, J. Daugelaite, and R. Sleator., “Big data, hadoop and cloud computing in genomics.” *J Biomed Inf*, vol. 46, no. 5, p. 774781, 2013.
- [3] M. Ashburner, C. Ball, J. Blake, D. Botstein, H. Butler, J. Cherry, A. Davis, K. Dolinski, S. Dwight, J. Eppig, M. Harris, D. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. Matese, J. Richardson, M. Ringwald, G. Rubin, and G. Sherlock, “Gene ontology: tool for the unification of biology. the gene ontology consortium.” *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [4] M. de Heus, “Towards a library of parallel graph algorithm in java,” in *14th Twente Student conference on IT*, January 21st, 2011.
- [5] R. Sedgewick and K. Wayne, “Algorithms,” vol. 4, 2012. [Online]. Available: <http://algs4.cs.princeton.edu/home/retrievedon04-2012>
- [6] J. Barnat, P. Bauch, L. Brim, and M. Ceska, “Computing strongly connected components in parallel on cuda.” IEEE International Parallel and Distributed Processing Symposium, 2011.
- [7] S. Alshomrani and G. Iqbal, “Analysis of strongly connected components (scc) using dynamic graph representation,” *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 4, pp. 94–100, 2012.
- [8] J. I. G. D. S. Graef, John W.; Wolfsdorf, *Manual of Pediatric Therapeutics*. Lippincott Williams and Wilkins (LWW), 2007.

- [9] S. Grundy, B. Hansen, S. J. Smith, J. Cleeman, and R. Kahn, "Clinical management of metabolic syndrome: report of the american heart association/national heart, lung, and blood institute/american diabetes association conference on scientific issues related to management." *Arterioscler Thromb Vasc Biol.*, vol. 24, no. 2, pp. 19–24, 2004.
- [10] T. Aldi, D. I. Kraja, K. E. Chasman, P. R. North, Alexander, R. Lisa, T. O. Yanek, J. A. Kilpelinen, Smith, A. Dehghan, J. Dupuis, D. J. Andrew, F. Mary, F. Feitosa, A. Y. Tekola-Ayele, I. M. Chu, Z. D. Nolte, M. Andrew, A. Sarah, Y. V. Pendergrass, M. D. Sun, Ritchie, V. Ahmad, L. Honghuang, L. Symen, M. Letizia, R. Rebecca, S. Yaming, A. Mark, H. Ziegler, and I. Kyung, "Pleiotropic genes for metabolic syndrome and inflammation." *Mol Genet Metab.*, vol. 112, no. 4, pp. 317–338, 2014.
- [11] A. Galassi, K. Reynolds, and J. He, "Metabolic syndrome and risk of cardiovascular disease: a metaanalysis." *The American journal of medicine.*, vol. 119, no. 2, pp. 812–819, 2006.
- [12] K. Monda, K. North, S. Hunt, D. Rao, and A. Province, MA. Kraja, "The genetics of obesity and the metabolic syndrome." *Endocrine, metabolic and immune disorders drug targets.*, vol. 10, no. 2, pp. 86–108, 2010.
- [13] R. Eckel, S. Grundy, and P. Zimmet, "The metabolic syndrome." *Lancet.*, vol. 365, no. 9468, pp. 1415–1428, 2005.
- [14] A. Lusis, A. Attie, and K. Reue, "Metabolic syndrome: from epidemiology to systems biology." *Nature reviews. Genetics.*, vol. 9, no. 11, pp. 819–830, 2008.
- [15] L. Djousse, H. Padilla, T. Nelson, J. Gaziano, and K. Mukamal, "Diet and metabolic syndrome." *Endocrine, metabolic and immune disorders drug targets.*, vol. 10, no. 2, pp. 124–137, 2010.
- [16] P. Katzmarzyk, A. Leon, J. Wilmore, J. Skinner, D. Rao, T. Rankinen, and C. Bouchard, "Targeting the metabolic syndrome with exercise: evidence

- from the heritage family study.” *Medicine and science in sports and exercise.*, vol. 35, no. 10, pp. 1703–1709, 2003.
- [17] M. Cerf, “Beta cell dysfunction and insulin resistance.” *Frontiers in Endocrinology.*, vol. 4, no. 37.
- [18] C. Nidhi and K. Dr. Kavita, “Theoretical studies of prevalence of obesity and type 2 diabetes mellitus,” *IJAR.*, vol. 1, no. 9, pp. 153–156, 2015.
- [19] K. Maedler, “Beta cells in type 2 diabetes - a crucial contribution to pathogenesis.” *Diabetes Obes Metab.*, vol. 10, no. 5, pp. 408–420, 2008.
- [20] T. Buchanan, A. Xiang, and K. Page, “Gestational diabetes mellitus: Risks and management during and after pregnancy.” *Nature reviews Endocrinology.*, vol. 8, no. 11, pp. 639–649, 2012.
- [21] J. Barry, Goldstein, and M.-W. Dirk, *Type 2 diabetes: principles and practice.* CRC Press ., 2008.
- [22] T. Yorifuji, K. Kurokawa, M. Mamada, T. Imai, M. Kawai, Y. Nishi, S. Shishido, Y. Hasegawa, and T. Nakahata, “Neonatal diabetes mellitus and neonatal polycystic, dysplastic kidneys: Phenotypically discordant recurrence of a mutation in the hepatocyte nuclear factor-1beta gene due to germline mosaicism.” *The Journal of Clinical Endocrinology and Metabolism.*, vol. 89, no. 6, pp. 2905–2908, 2004.
- [23] E. Edghill, C. Bingham, A. Slingerland, J. Minton, C. Noordam, S. Ellard, and A. Hattersley, “Hepatocyte nuclear factor-1 beta mutations cause neonatal diabetes and intrauterine growth retardation: support for a critical role of hnf-1beta in human pancreatic development.” *DiabeticMedicine.*, vol. 23, no. 12, pp. 1301–1306, 2006.
- [24] T. RB, “Mild familial diabetes with dominant inheritance.” *Q J Med.*, vol. 43, no. 170, pp. 339–357, 1974.

- [25] M. Charles and J. M. Clark, "The burden of diabetes: Introductory remarks." *The Worldwide Burden of Diabetes Workshop Proceedings.*, vol. 21, no. 3, pp. 1–2, 1996.
- [26] R. Rubin, W. Altman, and D. Mendelson, "Health care expenditure for people with diabetes mellitus, 1992," *J. Clin. Endocrinol. Metab.*, vol. 78, no. 4, pp. 809A–809F, 1994.
- [27] H. King and M. Rewers, "Global estimates for prevalence of diabetes mellitus and impaired glucose tolerance in adults. who ad hoc diabetes reporting group." *Diabetes Care.*, vol. 16, no. 1, pp. 157–177, 1993.
- [28] K. Ramaiya, V. Kodali, and K. Alberti, "Epidemiology of diabetes in asians of the indian subcontinent." *Diabetes Metab. Rev.*, vol. 6, no. 11, pp. 125–146, 1990.
- [29] A. Amos, D. McCarty, and P. Zimmet, "The rising global burden of diabetes and its complications; estimates and projections to the year 2010." *Diabetic Medicine.*, vol. 14, no. 5, pp. S1–S85, 1997.
- [30] C. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030." *PLoS Med.*, vol. 3, no. 11, pp. 1–54, 2006.
- [31] A. D. Association., "Economic costs of diabetes in the u.s. in 2012." *Diabetes Care.*, vol. 36, no. 6, pp. 1033–1034, 2013.
- [32] D. Alessandro, P. Mary-Elizabeth, and K. C. Ronald, "The emerging genetic architecture of type 2 diabetes." *Cell Metab.*, vol. 8, no. 3, pp. 186–200, 2008.
- [33] T. Waterfield and A. Gloyn, "Monogenic -cell dysfunction in children: clinical phenotypes, genetic etiology and mutational pathways." *Pediatr Health.*, vol. 2, no. 4, pp. 517–532, 2008.
- [34] S. Fajans, G. Bell, and K. Polonsky, "Molecular mechanisms and clinical pathophysiology of maturity-onset diabetes of the young." *N Engl J Med.*, vol. 345, no. 13, pp. 971–980, 2001.

- [35] S. Ellard, C. Bellann-Chantelot, and A. Hattersley, “Best practice guidelines for the molecular genetic diagnosis of maturity-onset diabetes of the young.” *Diabetologia.*, vol. 51, no. 4, pp. 546–553, 2008.
- [36] M. McCarthy and A. Hattersley, “Learning from molecular genetics: novel insights arising from the definition of genes for monogenic and type 2 diabetes.” *Diabetes.*, vol. 57, no. 11, pp. 2889–2898, 2008.
- [37] S. Sunita, “Genetics of type 2 diabetes: Advances and future prospect.” *Diabetes and Metabolism.*, vol. 6, no. 4, pp. 1–8, 2015.
- [38] I. Barroso, M. Gurnell, V. Crowley, M. Agostini, and J. Schwabe, “Dominant negative mutations in human ppargamma associated with severe insulin resistance, diabetes mellitus and hypertension.” *Nature.*, vol. 402, no. 6764, pp. 880–883, 1999.
- [39] Garcia-Montoya and Leticia, “Recent advances in ankylosing spondylitis: understanding the disease and management.” *F1000Research*, vol. 7, no. Rev-1512.
- [40] J. Braun, van den Berg R, X. Baraliakos, H. Boehm, Burgos-VargasR, and E. Collantes-Estevez, “Update of the asas/eular recommendations for the management of ankylosing spondylitis.” *Ann Rheum Dis*, vol. 70, no. 45.
- [41] J. Smolen, J. Braun, M. Dougados, P. Emery, O. FitzGerald, and P. Helliwell, “Treating spondyloarthritis, including ankylosing spondylitis and psoriatic arthritis, to target: recommendations of an international task force.” *Ann Rheum Dis*, vol. 73, no. 6.
- [42] P. Sidiropoulos, G. Hatemi, I. Song, J. Avouac, E. Collantes, and V. Hamuryudan, “Evidence-based recommendations for the management of ankylosing spondylitis: systematic literature search of the 3e initiative in rheumatology involving a broad panel of experts and practising rheumatologists.” *Rheumatology(Oxford)*, vol. 47, no. 23.

- [43] M. Vidal and S. Fields, “The yeast two-hybrid assay: still finding connections after 25 years.” *Nature*, vol. 11, pp. 1203–1206, 2014.
- [44] C. Perez-Iratxeta, P. Bork, and M. Andrade, “Association of genes to genetically inherited diseases using data mining.” *Nature genetics*, vol. 31, no. 3, pp. 316–319, 2002.
- [45] D. Harth, *The Invention of Cultural Memory*, 01 2008, pp. 85–96.
- [46] S. Buettcher, C. L. A. Clarke, and G. V. Cormack., “Information retrieval: Implementing and evaluating search engines.” *Cambridge, MA: MIT Press*, vol. 8, no. 4, pp. 600–632, 2010.
- [47] C. Ravi, M. B. Raju, and N. S. Chandra, “Mining frequent patterns from heterogeneous uncertain data streams using big data.” *Journal of advanced research in dynamical and controlled systems.*, vol. 10, no. 10, pp. 12–16, 2013.
- [48] H. S. Ibrahim, S. M. Abdou, and M. Gheith, “Sentiment analysis for modern standard arabic and colloquial.” *International Journal on Natural Language Computing (IJNLC)*, vol. 4, no. 2, pp. 1–15, 2015.
- [49] B. R, D. M, S. IN, T. K, Z. A, and M. AT, “Data analysis and data mining: current issues in biomedical informatics.” *Methods Inf Med.*, vol. 50, no. 6, p. 536544, 2011.
- [50] V. Rao and E. Smith, “New front-end serverdesign delivers on performance without sucking up power.” *InProceedings of the 2015 ACM SIGMOD International Conference on Management of Data. ACM.*, 2016.
- [51] S. Kulkarni, a. M. F. Nikunj Bhagat, V. Kedigehalli, C. Kellogg, S. Mittal, J. M. Patel, K. Ramasamy, and S. Taneja., “Twitter heron: Stream processing at scale.” *ACM*, 2015.
- [52] S. Moody, J. Boehm, D. Barbie, and W. Hahn, “Functional genomics and cancer drug target discovery.” *Current Opinion in Molecular Therapeutics*, vol. 12, no. 3, pp. 284–293, 2010.

- [53] M. He, M. Xu, B. Zhang, J. Liang, P. Chen, J.-Y. Lee, T. Johnson, H. Li, X. Yang, J. Dai, L. Liang, L. Gui, Q. Qi, J. Huang, Y. Li, L. Adair, T. Aung, Q. Cai, C.-Y. Cheng, M.-C. Cho, Y. Cho, M. Chu, B. Cui, Y.-T. Gao, M. Go, D. Gu, W. Gu, H. Guo, Y. Hao, J. Hong, Z. Hu, Y. Hu, J. Huang, J.-Y. Hwang, M. Ikram, G. Jin, D.-H. Kang, C. Khor, B.-J. Kim, H. Kim, M. Kubo, J. Lee, N. Lee, R. Li, J. Li, J. Liu, J. Longe, W. Lu, X. Lu, X. Miao, Y. Okada, R.-H. Ong, G. Qiu, M. Seielstad, and X. Sim, “Meta-analysis of genome-wide association studies of adult height in east asians identifies 17 novel loci.” *Human Molecular Genetics*, vol. 24, no. 6, pp. 1791–1800, 2015.
- [54] Y.-H. Qiu, F.-Y. Deng, M.-J. Li, and S.-F. Lei, “Identification of novel risk genes associated with type 1 diabetes mellitus using a genome-wide gene-based association analysis.” *Journal of Diabetes Investigation.*, vol. 5, no. 6, pp. 649–656, 2014.
- [55] S. Penttila, M. Jokela, H. Bouquin, A. Saukkonen, J. Toivanen, and B. Udd, “Late-onset spinal motor neuronopathy is caused by mutation in *chchd10*.” *Ann Neurol*, vol. 77, no. 1, pp. 163–172, 2015.
- [56] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, and C. Chen, “The intact molecular interaction database in 2012.” *Nucleic Acids Research.*, vol. 40, no. Database issue, pp. 841–846, 2012.
- [57] G. Bader, D. Betel, and C. Hogue, “Bind: the biomolecular interaction network database,” *Nucleic Acids Research.*, vol. 31, no. 1, pp. 248–250, 2003.
- [58] M. Krallinger, F. Leitner, and A. Valencia, “Analysis of biological processes and diseases using text mining approaches.” *Methods Mol Biol.*, vol. 593, pp. 341–382, 2010.
- [59] L. Adamic, D. Wilkinson, B. Huberman, and E. Adar, “A literature based method for identifying gene-disease connections.” *In Proceedings of the IEEE Computer Society Conference on Bioinformatics, Stanford, CA.*, vol. 1, pp. 109–117, 2002.

- [60] H. Al-Mubaid and R. Singh, “A new text mining approach for finding protein-to-disease associations.” *Am J Biochem Biotechnol.*, vol. 1, no. 3, pp. 145–152, 2005.
- [61] J. Freudenberg and P. Propping, “A similaritybased method for genomewide prediction of disease-relevant human genes.” *Bioinformatics.*, vol. 18, no. 2, pp. 110–115, 2002.
- [62] P. Glenisson, B. Coessens, S. Vooren, J. Mathys, Y. Moreau, and B. De Moor, “Txtgate: profiling gene groups with text-based information.” *Genome Biol.*, vol. 5, no. 6, pp. 1–12, 2004.
- [63] D. Huang, B. Sherman, and R. Lempicki, “Systematic and integrative analysis of large gene lists using david bioinformatics resources.” *Nat. Protoc.*, vol. 4, no. 1, pp. 44–57, 2009.
- [64] A. Subramanian, P. Tamayo, V. Mootha, S. Mukherjee, B. Ebert, M. Gillette, A. Paulovich, S. Pomeroy, T. Golub, E. Lander, and J. Mesirov, “Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.” *Proc. Natl Acad. Sci.*, vol. 102, no. 43, pp. 15 545—15 550, 2005.
- [65] D. Martin, C. Brun, E. Remy, P. Mouren, D. Thieffry, and B. Jacq, “Gotoolbox: functional analysis of gene datasets based on gene ontology.” *Genome Biol.*, vol. 5, no. 12, pp. 1–8, 2004.
- [66] S. Micha, “A strong connectivity algorithm and its applications to data flow analysis. computers and mathematics with applications.” *Ann Neurol*, vol. 7, no. 1, pp. 67–72, 1981.
- [67] R. Tarjan, “Depth-first search and linear graph algorithms.” *SIAM Journal on Computing*, vol. 1, no. 2, pp. 146–160, 1972.
- [68] K. Raman, “Construction and analysis of proteinprotein interaction networks.” *Automated Experimentation.*, vol. 2, no. 2, pp. 1759–4499, 2010.

- [69] S. Hwang, S. Son, S. Kim, Y. Kim, H. Jeong, and D. Lee, “A protein interaction network associated with asthma.” *J Theor Biol*, vol. 25, no. 2, pp. 722–731, 2008.
- [70] T. Lon-Charles, B. Francisco, D. Capdevila, B. Nitsch, P. De Moor, C. De, and M. Yves, “A guide to web tools to prioritize candidate genes,” *Briefings in Bioinformatics*, vol. 12, no. 1, pp. 22–32, 2011.
- [71] T. Nadezhda, T. Doncheva, M. Kacprowski, and Albrecht, “Recent approaches to the prioritization of candidate disease genes.” *WIREs System Biology and Medicine*, vol. 4, no. 5, pp. 429–442, 2012.
- [72] M. Rosario, Piro, and D. C. Ferdinando, “Computational approaches to disease-gene prediction: rationale, classification and successes.” *The FEBS Journal*, vol. 279, no. 5, pp. 678–696, 2012.
- [73] Y. Moreau and L. Tranchevent, “Computational tools for prioritizing candidate genes: boosting disease-gene discovery.” *Nat. Rev. Genet.*, vol. 13, no. 8, pp. 523–536, 2012.
- [74] E. Sarah, T. Leon-Charles, S. Alejandro, A. Amin, D. Jesse, and M. Yves, “Beegle: from literature mining to disease-gene discovery.” *Nucleic Acids Research*, vol. 44, no. 2, pp. 1–8, 2016.
- [75] A. Rakesh and S. Ramakrishnan, “Fast algorithms for mining association rules.” *Proceedings of the 20th International Conference on Very Large Data Bases.*, vol. 1215, pp. 487–499, 2000.
- [76] T. Karthikeyan and N. Ravikumar, “A survey on association rule mining.” *International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE).*, vol. 3, no. 1, pp. 5223–5227, 2014.
- [77] X. Li, “Improvement of apriori algorithm for association rules.” *World Automation Congress (WAC).*, 2012.

- [78] A. Zgr, T. Vu, G. Erkan, and D. Radev, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network." *Bioinformatics.*, vol. 24, no. 13, p. 277285, 2008.
- [79] Y. Akane, M. Yusuke, T. Yuka, and T. Junichi, "Biomedical information extraction with predicate-argument structure patterns," in *In Proceedings of the Eleventh annual meeting of the association for natural language processing.*, 2005, pp. 60–69.
- [80] J. Temkin and M. Gilder, "Extraction of protein interaction information from unstructured text using a context-free grammar." *Bioinformatics.*, vol. 19, no. 16, pp. 2046–2053, 2003.
- [81] G. Erkan, D. Radev, and A. Ozgur, "Semisupervised classification for extracting protein interaction sentences using dependency parsing," in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 228–237.
- [82] K. Fundel, R. Kuffner, and R. Zimmer, "Relex-relation extraction using dependency parse trees." *Bioinformatics.*, vol. 23, no. 3, pp. 365–371, 2006.
- [83] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, L.-c. C. Tranchevent, B. De Moor, P. Marynen, B. Hassan, P. Carmeliet, and Y. Moreau, "Gene prioritization through genomic data fusion," *Nature biotechnology*, vol. 24, no. 5, pp. 537–544, 2006.
- [84] J. Chen, C. Shen, and A. Sivachenko, "Mining alzheimer disease relevant proteins from integrated protein interactome data." *Pac. Symp. Biocomput.*, vol. 11, pp. 367–378, 2006.
- [85] X. Ma, H. Lee, L. Wang, and F. Sun, "Cgi: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data," *Bioinformatics*, vol. 23, no. 2, pp. 215–221, 2007.

- [86] J. Hutz, A. Kraja, H. McLeod, and M. Province, "Candid: a flexible method for prioritizing candidate genes for complex human traits." *Genet Epidemiol.*, vol. 32, no. 8, pp. 779–790, 2008.
- [87] J. Morrison, R. Breitling, D. Higham, and D. Gilbert, "Generank: using search engine technology for the analysis of microarray experiments." *BMC Bioinformatics.*, vol. 6, pp. 233–247, 2005.
- [88] Y. Oner, T. Tunc, E. Egrioglu, and Y. Atasoy, "Comparisons of logistic regression and artificial neural networks in lung cancer data," *Am. J. Intell. Syst.*, vol. 3, no. 2, pp. 71–74, 2013.
- [89] Y. Guo, G. Bai, and Y. Hu, "Using bayes network for prediction of type-2 diabetes," *Proceeding of the IEEE International Conference for Internet Technology and Secured Transactions.*, pp. 471–472, 2012.
- [90] H. Ji, H. Dong, B. Young, and B. Seoung, "A text mining approach to find patterns associated with diseases and herbal materials in oriental medicine." *International Journal of Information and Education Technology.*, vol. 2, no. 3, pp. 224–226, 2012.
- [91] M. Sajid, M. Shahbaz, and A. G., "Negative and positive association rules mining fromtext using frequent and infrequent itemsets." *The scientific world journal.*, vol. 2014, no. 2, pp. 1–11, 2014.
- [92] Y. Tsuruoka, J. Kim, R. Shiba, N. Nagata, T. Hishiki, and J. Tsujii, "Extraction of genedisease relations from medline using domain dictionaries and machine learning." *In Proceedings of the Pacific Symposium on Biocomputing.*, pp. 4–15, 2006.
- [93] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, "Mint: a molecular interaction database." *FEBS Letters*, vol. 513, no. 1, pp. 135–140, 2002.

- [94] T. Hwang, W. Zhang, M. Xie, J. Liu, and R. Kuang, “Inferring disease and gene set associations with rank coherence in networks.” *Bioinformatics.*, vol. 27, no. 19, pp. 2692–2699, 2011.
- [95] V. McKusick, “Mendelian inheritance in man and its online version, omim.” *Am. J. Hum. Genet.*, vol. 80, no. 8, pp. 588–604, 2007.
- [96] S. Wuchty, Z. Oltvai, and A. Barabasi, “Evolutionary conservation of motif constituents in the yeast protein interaction network.” *Nat. Genet.*, vol. 35, no. 2, pp. 176–179, 2003.
- [97] B. Schwikowski, P. Uetz, and S. Fields, “A network of protein-protein interactions in yeast.” *Nat. Biotechnol.*, vol. 18, no. 12, pp. 1257–1261, 2000.
- [98] d. Boyd and N. Ellison, “Social network sites: Definition, history, and scholarship.” *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2007.
- [99] M. Kurt, N. Stefan, and S. Peter, “Engineering dfs based graph algorithms,” *Partially supported by DFG grant SA*, vol. 933, 2007.
- [100] H. Gabow, “Path-based depth first search strong and biconnected components,” *Information Processing Letters*, vol. 74, no. 3 and 4, pp. 107–114, 2000.
- [101] D. Swati, S. Poorvi, Dodwad, and M. Meghna, “Finding strongly connected components in a social network graph,” *International Journal of Computer Applications*, vol. 136, no. 7, pp. 0975–8887, 2016.
- [102] B. Surender, C. Keerti, and R. Liam, “An efficient strongly connected components algorithm in the fault tolerant model.” *44th International Colloquium on Automata, Languages, and Programming*, vol. 72, 2017.
- [103] C. Jean-Michel, *On-the-Fly Verification of Linear Temporal Logic*, 1999, vol. 1708.

- [104] J. Rual, K. Venkatesa, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. Berriz, F. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. Goldberg, L. Zhang, S. Wong, G. Franklin, S. Li, J. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. Sikorski, J. Vandenhaute, H. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. Cusick, D. Hill, F. Roth, and M. Vidal, "Towards a proteome-scale map of the human protein-protein interaction network." *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.
- [105] S. Lee, T. Tsao, K. Yang, H. Lin, Y. Kuo, C. Hsu, W. Lee, K. Huang, and C. Kao, "Construction and analysis of the protein-protein interaction networks for schizophrenia, bipolar disorder, and major depression." *BMC Bioinformatics*, vol. 12, no. 13, pp. 13–20, 2011.
- [106] J. Ran, H. Li, J. Fu, L. Liu, Y. Xing, X. Li, H. Shen, Y. Chen, X. Jiang, and Y. Li, "Construction and analysis of the protein-protein interaction network related to essential hypertension," *BMC Syst Biol*, vol. 7, no. 32, pp. 7–32, 2013.
- [107] H. Rakshit, N. Rathi, and D. Roy, "Construction and analysis of the protein-protein interaction networks based on gene expression profiles of parkinson's disease." *PLoS One*, vol. 9, no. 8, pp. 1–17, 2014.
- [108] G. Jaco and V. Antti, "More efficient on-the-fly ltl verification with tarjan's algorithm," *Theoretical Computer Science*, vol. 345, no. 7, pp. 60–82, 2005.
- [109] C. Chen, S. Hong, Z. Li-guo, L. Jian, C. Xiao-ge, Y. An-liang, K. Shao-san, G. Wei-xing, H. Hui, C. Feng-hong, and L. Zhi-guo, "Construction and analysis of protein-protein interaction networks based on proteomics data of prostate cancer." *INTERNATIONAL JOURNAL OF MOLECULAR MEDICINE*, vol. 37, no. 6, pp. 1576–1586, 2016.
- [110] S. Soudabeh and S. Mohd Shahir, "Systematic analysis of protein interaction network associated with azoospermia," *Int. J. Mol. Sci*, vol. 17, no. 11, pp. 1–10, 2016.

- [111] S. Akram, R. Mostafa, A. Afsaneh, Z. Mona, M. Seyed Reza, and N. Abdol Rahim, "Protein-protein interaction network analysis of cirrhosis liver disease," *Gastroenterol Hepatol Bed Bench*, vol. 9, no. 2, pp. 114–123, 2016.
- [112] T. A. Kumbhare and P. S. V. Chobe., "An overview of association rule mining algorithms," *International Journal of Computer Science and Information Technologies*, vol. 5, no. 1, pp. 927–930, 2014.
- [113] K. Komal and S. Simple, "A comparative analysis of association rules mining algorithms." *International Journal of Scientific and Research Publications.*, vol. 3, no. 5, pp. 2250–3153, 2013.
- [114] K. Gagandeep and A. Shruti, "Performance analysis of association rule mining algorithms." *International Journal of Advanced Research in Computer Science and Software Engineering.*, vol. 3, no. 8, pp. 1–28, 2013.
- [115] S. Pratiksha and G. Tina, "Comparitive study of apriori and fp growth algorithms." *Indian journal of research.*, vol. 2, no. 3, pp. 134–145, 2013.
- [116] P. Parita and W. Dinesh, "Comparative study of association rule mining algorithms." *UNIASCIT*, vol. 2, no. 1, pp. 170–172, 2012.
- [117] U. Ferraro Petrillo, G. Roscigno, G. Cattaneo, and R. Giancarlo, "Fastdoop: a versatile and efficient library for the input of fasta and fastq files for mapreduce hadoop bioinformatics applications." *Bioinformatics*, vol. 33, no. 10, p. 15751577, 2017.
- [118] A. Amir and V. Virginia, "Popular conjectures imply strong lowerbounds for dynamic problems." 2014, p. 434443.
- [119] C. Shiri, D. Thomas, I. Giuseppe, L. Jakub, and P.-s. Nikos, "Decremental single-source reachability and strongly connected components in total update time." 2016, p. 315324.
- [120] G. Loukas, I. Giuseppe, and P. Nikos, "Strong connectivity in directedgraphs under failures, with applications." 2017, p. 18801899.

- [121] K. Bruce, K. Valerie, and M. Ben, “Dynamic graph connectivity in polylogarithmic worst case time.” 2013, p. 11311142.
- [122] R. Roberts, “Pubmed central: The genbank of the published literature.” *Proceedings of the National Academy of Sciences of the United States of America.*, vol. 98, no. 2, pp. 381–382, 2001.
- [123] A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, and E. Mountjoy, “The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.” *Nucleic Acids Research.*, vol. 47, no. D1, pp. 1362—4962, 2019.
- [124] M. Uhln, L. Fagerberg, B. Hallstrm, C. Lindskog, P. Oksvold, and A. Mardinoglu, “Tissue-based map of the human proteome.” *Science*, vol. 347, no. 6220, pp. 394–400, 2015.
- [125] Y. Liu, Y. Liang, and D. Wishart, “Polysearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more.” *Nucleic Acids Res.*, vol. 43, no. W1, pp. 535–542, 2015.
- [126] A. Hamosh, A. Scott, J. Amberger, and V. Bocchini, CA.and McKusick, “Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders.” *Nucleic Acids Res.*, vol. 33, no. D1, pp. 514–517, 2005.
- [127] I. Barroso, M. Gurnell, V. Crowley, M. Agostini, and J. Schwabe, “Dominant negative mutations in human ppargamma associated with severe insulin resistance, diabetes mellitus and hypertension.” *Nature*, vol. 402, no. 6764, pp. 880–883, 1999.
- [128] D. Altshuler, J. Hirschhorn, M. Klannemark, C. Lindgren, and M. Vohl, “The common ppargamma pro12ala polymorphism is associated with decreased risk of type 2 diabetes.” *Nat Genet*, vol. 26, no. 1, pp. 76–80, 2000.

- [129] M. Agostini, M. Gurnell, and D. Savage, "Tyrosine agonists reverse the molecular defects associated with dominant-negative mutations in human peroxisome proliferator-activated receptor gamma." *Endocrinology*, vol. 145, no. 4, pp. 1527–1538, 2004.
- [130] A. Gloyn, M. Weedon, K. Owen, M. Turner, and B. Knight, "Large scale association studies of variants in genes encoding the pancreatic betacell k-atp channel subunits kir6.2 (kcnj11) and sur1 (abcc8) confirm that the kcnj11 e23k variant is associated with type 2 diabetes." *Diabetes*, vol. 52, no. 2, pp. 568–572, 2003.
- [131] J. Florez, N. Burt, P. de Bakker, P. Almgren, and T. Tuomi, "Haplotype structure and genotype phenotype correlations of the sulfonylurea receptor and the islet atp-sensitive potassium channel gene region." *Diabetes*, vol. 53, no. 5, pp. 1360–1368, 2004.
- [132] S. Grant, G. Thorleifsson, I. Reynisdottir, R. Benediktsson, and A. Manolescu, "Variant of transcription factor 7-like 2 (tcf7l2) gene confers risk of type 2 diabetes." *Nat Genet*, vol. 38, pp. 320–323, 2006.
- [133] C. Zhang, L. Qi, D. Hunter, J. Meigs, and J. Manson, "Variant of transcription factor 7-like 2 (tcf7l2) gene and the risk of type 2 diabetes in large cohorts of u.s. women and men." *Diabetes*, vol. 55, no. 9, pp. 2645–2648, 2006.
- [134] R. Saxena, L. Gianniny, N. Burt, V. Lyssenko, and C. Giuducci, "Common single nucleotide polymorphisms in tcf7l2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals." *Diabetes*, vol. 55, no. 10, pp. 2890–2895, 2006.
- [135] C. Groves, E. Zeggini, J. Minton, T. Frayling, and M. Weedon, "Association analysis of 6,736 u.k. subjects provides replication and confirms tcf7l2 as a type 2 diabetes susceptibility gene with a substantial effect on individual risk." *Diabetes*, vol. 55, no. 12, pp. 2640–2644, 2006.

- [136] J. Florez, K. Jablonski, N. Bayley, T. Pollin, and P. de Bakker, "Tcf7l2 polymorphisms and progression to diabetes in the diabetes prevention program." *N Engl J Med*, vol. 355, no. 3, pp. 241–250, 2006.
- [137] C. M. Damcott, T. I. Pollin, L. J. Reinhart, S. H. Ott, H. Shen *et al.*, "Polymorphisms in the transcription factor 7-like 2 (tcf7l2) gene are associated with type 2 diabetes in the amish: replication and evidence for a role in both insulin secretion and insulin resistance." *Diabetes*, vol. 55, no. 9, pp. 2654–2659, 2006.
- [138] L. J. Scott, L. L. Bonnycastle, C. J. Willer, A. G. Sprau, A. U. Jackson *et al.*, "Association of transcription factor 7-like 2 (tcf7l2) variants with type 2 diabetes in a finnish sample." *Diabetes*, vol. 55, no. 9, pp. 2649–2653, 2006.
- [139] S. Cauchi, D. Meyre, C. Dina, H. Choquet, C. Samson *et al.*, "Transcription factor tcf7l2 genetic study in the french population: Expression in human beta-cells and adipose tissue and strong association with type 2 diabetes." *Diabetes*, vol. 55, no. 10, pp. 2903–2908, 2006.
- [140] T. Hayashi, Y. Iwamoto, K. Kaku, H. Hirose, and S. Maeda, "Replication study for the association of tcf7l2 with susceptibility to type 2 diabetes in a japanese population." *Diabetologia*, vol. 50, no. 5, pp. 980–984, 2007.
- [141] M. Horikoshi, K. Hara, C. Ito, R. Nagai, and P. Froguel, "A genetic variation of the transcription factor 7-like 2 gene is associated with risk of type 2 diabetes in the japanese population." *Diabetologia*, vol. 50, no. 4, pp. 747–751, 2007.
- [142] D. Lehman, K. J. Hunt, R. J. Leach, J. Hamlington, and R. Arya, "Haplotypes of transcription factor 7-like 2 (tcf7l2) genes and its upstream region are associated with type 2 diabetes and age of onset in mexican americans." *Diabetes*, vol. 56, no. 2, pp. 389–393, 2007.

- [143] M. S. Sandhu, M. N. Weedon, K. A. Fawcett, J. Wasson, and S. L. Debenham, "Common variants in wfs1 confer risk of type 2 diabetes." *Nat Genet*, vol. 39, no. 8, pp. 951–953, 2007.
- [144] E. Zeggini, M. N. Weedon, C. M. Lindgren, T. M. Frayling, and K. S. Elliott, "Replication of genome-wide association signals in uk samples reveals risk loci for type 2 diabetes." *Science*, vol. 316, no. 5829, pp. 1336–1341, 2007.
- [145] R. Saxena, B. F. Voight, V. Lyssenko, N. P. Burtt, and P. I. de Bakker, "Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels." *Science*, vol. 316, no. 5829, pp. 1331–1336, 2007.
- [146] L. J. Scott, K. L. Mohlke, L. L. Bonnycastle, C. J. Willer, and Y. Li, "A genomewide association study of type 2 diabetes in finns detects multiple susceptibility variants." *Science*, vol. 316, no. 5829, pp. 1341–1345, 2007.
- [147] V. Steinthorsdottir, G. Thorleifsson, I. Reynisdottir, R. Benediktsson, and T. Jonsdottir, "A variant in cdkal1 influences insulin response and risk of type 2 diabetes." *Nat Genet*, vol. 39, no. 6, pp. 770–775, 2007.
- [148] R. Sladek, G. Rocheleau, J. Rung, C. Dina, and L. Shen, "A genome-wide association study identifies novel risk loci for type 2 diabetes." *Nature*, vol. 445, no. 7130, pp. 881–885, 2007.
- [149] M. B. Schulze, H. Al-Hasani, H. Boeing, E. Fisher, and F. Doring, "Variation in the hhex- ide gene region predisposes to type 2 diabetes in the prospective, population based epic- potsdam cohort." *Diabetologia*, vol. 50, no. 11, pp. 2405–2407, 2007.
- [150] W. Winckler, R. R. Graham, P. I. de Bakker, M. Sun, and P. Almgren, "Association testing of variants in the hepatocyte nuclear factor 4alpha gene with risk of type 2 diabetes in 7,883 people." *Diabetes*, vol. 54, no. 3, pp. 886–892, 2005.
- [151] T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, and R. M. Freathy, "A common variant in the fto gene is associated with body mass index and

- predisposes to childhood and adult obesity.” *Science*, vol. 316, no. 5826, pp. 889–894, 2007.
- [152] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, and P. Deloukas, “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.” *Nature*, vol. 447, no. 7145, pp. 661–678, 2007.
- [153] S. L. J. Zeggini E., R. Saxena, T. Hu, and B. F. Voight, “Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes.” *Nat Genet*, vol. 40, no. 5, pp. 638–645, 2008.
- [154] S. Omori, Y. Tanaka, M. Horikoshi, A. Takahashi, and K. e. a. Hara, “Replication study for the association of new meta analysis derived risk loci with susceptibility to type 2 diabetes in 6,244 japanese individuals.” *Diabetologia*, vol. 52, no. 8, pp. 1554–1560, 2009.
- [155] B. F. Voight, L. J. Scott, V. Steinthorsdottir, A. P. Morris, and C. Dina, “Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis.” *Nat Genet*, vol. 42, no. 7, pp. 579–589, 2010.
- [156] J. Dupuis, C. Langenberg, I. Prokopenko, R. Saxena, and N. Soranzo, “New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk.” *Nat Genet*, vol. 42, no. 2, pp. 105–116, 2010.
- [157] N. Bouatia-Naji, A. Bonnefond, C. Cavalcanti-Proenca, T. Spars, and J. Holmkvist, “A variant near *mtnr1b* is associated with increased fasting plasma glucose levels and type 2 diabetes risk.” *Nat Genet*, vol. 41, no. 1, pp. 89–94, 2009.
- [158] K. Pruitt, J. Harrow, and R. Harte, “The consensus coding sequence (ccds) project: Identifying a common protein-coding gene set for the human and mouse genomes.” *Genome Res.*, vol. 19, no. 7, pp. 1316–1323, 2009.

- [159] P. Shannon, A. Markiel, and O. Ozier, “Cytoscape: a software environment for integrated models of biomolecular interaction networks.” *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.
- [160] D. Auber, “Tulip a huge graph visualization framework, in graph drawing software,” in *Mathematics and Visualization*, Springer, Berlin, Germany, 2004, pp. 105–126.
- [161] M. Jacomy, T. Venturini, S. Heymann, and M. Bastianz, “Forceatlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software,” *Plos ONE*, vol. 9, no. 6, pp. 986–979, 2014.
- [162] A. Mrvar and V. Batagelj, “Analysis and visualization of large networks with program package pajek.” *Complex Adaptive Systems Modeling*, vol. 4, no. 1, p. 6.
- [163] M. Kanehisa and S. Goto, “Kegg: Kyoto encyclopedia of genes and genomes.” *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, 2000.
- [164] Z. Fu, E. R. Gilbert, and D. Liu, “Regulation of insulin synthesis and secretion and pancreatic beta-cell dysfunction in diabetes.” *Current diabetes reviews*, vol. 9, no. 1, pp. 25–53, 2013.
- [165] K. Tsuruzoe, R. Emkey, K. M. Kriauciunas, K. Ueki, and K. Cr., “Insulin receptor substrate 3 (irs-3) and irs-4 impair irs-1- and irs-2-mediated signaling.” *Molecular and Cellular Biology*, vol. 21, no. 1, pp. 26–38, 2001.
- [166] J. Boucher, A. Kleinridders, and K. Cr., “Insulin receptor signaling in normal and insulin-resistant states,” *Cold Spring Harbor Perspectives in Biology.*, vol. 6, no. 1, pp. 1–23, 2014.
- [167] R. S. Salamon and B. Jm., “Pip3:tool of choice for the class i pi 3-kinases,” *BioEssays: news and reviews in molecular, cellular and developmental biology*, vol. 35, no. 7, pp. 602–611, 2013.
- [168] B. D. Manning and T. A.

- [169] Q. L. Zhou, Z. Y. Jiang, and J. Holik, "Akt substrate tbc1d1 regulates glut1 expression through the mtor pathway in 3t3-l1 adipocytes," *The Biochemical journal*, vol. 411, no. 3, pp. 647–655, 2008.
- [170] A. Tzatsos, "Raptor binds the shc and irs-1 npxy binding) domain of insulin receptor substrate-1 (irs-1) and regulates the phosphorylation of irs-1 at ser-636/639 by mtor," *The Journal of Biological Chemistry*, vol. 284, no. 34, pp. 22 525–22 534, 2009.
- [171] W. M. Chu, "Tumor necrosis factor." *Cancer Lett*, vol. 328, no. 2, pp. 222–225, 2013.
- [172] M. Shuh, H. Bohorquez, L. G. E. Jr, and C. Aj., "Tumor necrosis factor-alpha: Life and death of hepatocytes during liver ischemia/reperfusion injury." *Ochsner J*, vol. 13, no. 1, pp. 119–130, 2013.
- [173] D. MacEwan, "Tnf receptor subtype signalling: differences and cellular consequences." *Cell Signal*, vol. 14, no. 6, pp. 477–492, 2002.
- [174] P. Itkin-Ansari, E. Marcora, I. Geron, B. Tyrberg, C. Demeterco, E. Hao, C. Padilla, C. Ratineau, A. Leiter, J. E. Lee, and F. Levine, "Neurod1 in the endocrine pancreas: localization and dual function as an activator and repressor," *Dev DynJul;*, vol. 233, no. 3, pp. 946–953, 2005.
- [175] K. D. Copps and W. Mf., "Regulation of insulin sensitivity by serine/threonine phosphorylation of insulin receptor substrate proteins irs1 and irs2," *Diabetologia*, vol. 55, no. 10, pp. 2565–2582, 2012.
- [176] C. S. Chao, Z. L. Loomis, J. E. Lee, and L. Sussel, "Genetic identification of a novel neurod1 function in the early differentiation of islet a, pp and e cells," *Developmental biology*, vol. 312, no. 2, pp. 523–532, 2007.
- [177] R. Dassaye, S. Naidoo, and C. Me., "Transcription factor regulation of pancreatic organogenesis, differentiation and maturation," *Islets*, vol. 8, no. 1, pp. 13–34, 2016.

- [178] H. P. Shih, J. L. Kopp, and M. Sandhu, “A notch-dependent molecular circuitry initiates pancreatic endocrine and ductal cell differentiation,” *Development (Cambridge, England)*, vol. 139, no. 14, pp. 2488–2499, 2012.
- [179] M. Ejarque, S. Cervantes, G. Pujadas, A. Tutusaus, L. Sanchez, and R. Gasa, “Neurogenin3 cooperates with foxa2 to autoactivate its own expression.” *J Biol Chem.*, vol. 288, no. 17, pp. 11 705–11117, 2013.
- [180] R. G. Jones, D. R. Plas, and S. Kubek, “Amp-activated protein kinase induces a p53-dependent metabolic checkpoint.” *Molecular Cell*, vol. 18, no. 3, pp. 283–293, 2005.
- [181] Z. Feng, W. Hu, and E. de Stanchina, “The regulation of ampkbeta1, tsc2, and pten expression by p53: Stress, cell and tissue specificity, and the role of these gene products in modulating the igf-1-akt-mtor pathways,” *Cancer Research*, vol. 67, no. 7, pp. 3043–3053, 2007.
- [182] A. J. Levine, Z. Feng, T. W. Mak, H. You, and J. S., “Coordination and communication between the p53 and igf-1-akt-tor signal transduction pathways.” *Genes and Development*, vol. 20, no. 3, pp. 267–275, 2006.
- [183] J. Zhang, “The direct involvement of sirt1 in insulin-induced insulin receptor substrate-2 tyrosine phosphorylation.” *J Biol Chem. EpubSep*, vol. 282, no. 47, pp. 34 356–34 364, 2007.
- [184] W. C. Comb, J. E. Hutti, P. Cogswell, L. C. Cantley, and B. As., “p85a sh2 domain phosphorylation by ikk promotes feedback inhibition of pi3k and akt in response to cellular starvation,” *Molecular Cell*, vol. 45, no. 6, pp. 719–730, 2012.
- [185] B. Wang, Z. Jie, D. Joo, A. Ordureau, P. Liu, W. Gan, J. Guo, J. Zhang, B. J. North, X. Dai, and X. Cheng, “Bian x zhang l, harper jw, sun sc, wei w,” *Nature*, vol. 18, no. 545, pp. 365–369, May 2017.

- [186] A. Ito, L. C-h, and X. Zhao, “p300/cbp-mediated p53 acetylation is commonly induced by p53-activating agents and inhibited by mdm2,” *The EMBO Journal*, vol. 20, no. 6, pp. 1331–1340, 2001.
- [187] J. Eeckhoutte, P. Formstecher, and B. Laine, “Hepatocyte nuclear factor 4a enhances the hepatocyte nuclear factor 1a-mediated activation of transcription,” *Nucleic Acids Research*, vol. 32, no. 8, pp. 2586–2593, 2004.
- [188] J.-s. Bae, T.-h. Kim, M.-y. Kim, J.-m. Park, and Y.-h. Ahn, “Transcriptional regulation of glucose sensors in pancreatic cells and liver: An update,” *Sensors (Basel, Switzerland)*, vol. 10, no. 5, pp. 5031–5053, 2010.
- [189] N. . Tanimizu and A. Miyajima, “Notch signaling controls hepatoblast differentiation by altering the expression of liver-enriched transcription factors,” *Send to J Cell SciJul*, vol. 1, no. 117, pp. 3165–74, 2004.
- [190] N. Grarup, G. Andersen, and N. T. Krarup, “Association testing of novel type 2 diabetes risk alleles in the jazf1, cdc123/camk1d, tspan8, thada, adamts9, and notch2 loci with insulin release, insulin sensitivity, and obesity in a population-based sample of 4,516 glucose-tolerant middle-aged danes,” *Diabetes*, vol. 57, no. 9, pp. 2534–2540, 2008.
- [191] B. Lee and J. Kim, “Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning.” *IEEE J Biomed Health Inform*, vol. 20, no. 1, pp. 39–46, 2016.
- [192] B. Lee, B. Ku, J. Nam, D. Pham, and J. Kim, “Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes.” *IEEE J Biomed Health Inform*, vol. 18, no. 2, pp. 555–561, 2014.
- [193] N. Khan, A. Ahmad, R. Tiwari, M. Kamal, and G. Mushtaq, “Current challenges to overcome in the management of type 2 diabetes mellitus and associated neurological disorders.” *CNS Neurol Disord Drug Targets.*, vol. 13, no. 3, pp. 1440—1457, 2014.

- [194] P. De Meyts and J. Whittaker, "Structural biology of insulin and igf1 receptors: implications for drug design." *Nat .Rev. Drug Discov.*, vol. 1, no. 10, pp. 769–783, 2002.
- [195] M. F. White, "Irs proteins and the common path to diabetes." *Am. J. Physiol Endocrinol. Metab.*, vol. 283, no. 3, pp. 413–422, 2002.
- [196] D. R. Alessi and C. P. Downes, "The role of pi 3-kinase in insulin action." *Biochim Biophys. Acta*, vol. 1436, no. 1-2, pp. 151–164, 1998.
- [197] B. Vanhaesebroeck and D. R. Alessi, "The pi3k-pdk1 connection: more than just a road to pkb." *Biochem J.*, vol. 346, no. 3, pp. 561–576, 2000.
- [198] M. Beeson, M. P. Sajan, M. Dizon, D. Grebenev, J. Gomez-Daspert, A. Miura, Y. Kanoh, J. Powe, G. Bandyopadhyay, M. L. Standaert, and R. V. Farese, "Activation of protein kinase c-zeta by insulin and phosphatidylinositol-3,4,5-(po4)3 is defective in muscle in type 2 diabetes and impaired glucose tolerance: amelioration by rosiglitazone and exercise." *Diabetes*, vol. 52, no. 8, pp. 1926–1934, 2003.
- [199] J. A. L. Good, W. H. Ziegler, D. B. Parekh, D. R. Alessi, P. Cohen, and P. J. Parker, "Protein kinase c isotypes controlled by phosphoinositide 3-kinase through the protein kinase pdk1." *Science*, vol. 281, no. 5385, pp. 2042–2045, 1998.
- [200] N. Pullen, P. B. Dennis, M. Andjelkovic, A. Dufner, S. C. Kozma, B. A. Hemmings, and G. Thomas, "Phosphorylation and activation of p70s6k by pdk1." *Science*, vol. 279, no. 5351, pp. 707–710, 1998.
- [201] L. Stephens, K. Anderson, D. Stokoe, H. Erdjument-Bromage, G. F. Painter, A. B. Holmes, P. R. Gaffney, C. B. Reese, F. McCormick, P. Tempst, J. Coadwell, and P. T. Hawkins, "Protein kinase b kinases that mediate phosphatidylinositol 3,4,5- trisphosphate-dependent activation of protein kinase b." *Science*, vol. 279, no. 5351, pp. 710–714, 1998.

- [202] D. A. Cross, D. R. Alessi, P. Cohen, M. Andjelkovich, and B. A. Hemmings, "Inhibition of glycogen synthase kinase-3 by insulin mediated by protein kinase b." *Nature*, vol. 378, no. 6559, pp. 785–789, 1995.
- [203] S. Kane, H. Sano, S. C. Liu, J. M. Asara, W. S. Lane, C. C. Garner, and G. E. Lienhard, "A method to identify serine kinase substrates akt phosphorylates a novel adipocyte protein with a rab gtpase-activating protein (gap) domain." *J. Biol Chem.*, vol. 277, no. 2002, p. 2211522118, 1998.
- [204] A. S. M.J. Brady, F.J. Bourbonais, "The activation of glycogen synthase by insulin switches from kinase inhibition to phosphatase activation during adipogenesis in 3t3-l1 cells." *J. Biol. Chem.*, vol. 273, no. 23, pp. 14 063–14 066, 1998.
- [205] H. Sano, S. Kane, E. Sano, C. P. Miinea, J. M. Asara, W. S. Lane, C. W. Garner, and G. E. Lienhard, "Insulin-stimulated phosphorylation of a rab gtpase-activating protein regulates glut4 translocation." *J.Biol. Chem*, vol. 278, no. 17, pp. 14 599—14 602, 2003.
- [206] M. Beeson, M. P. Sajan, J. G. Daspet, V. Luna, M. Dizon, D. Grebenev, J. L. Powe, S. Lucidi, A. Miura, Y. Kanoh, G. Bandyopadhyay, M. L. Standaert, T. R. Yeko, and R. V. Farese, "Defective activation of protein kinase c-z in muscle by insulin and phosphatidylinositol-3, 4,5,-(po(4))(3) in obesity and polycystic ovary syndrome," *Metab. Syndr. Relat. Disord.*, vol. 2, pp. 49–56, 2004.
- [207] E. V. Obberghen, V. Baron, L. Delahaye, B. Emanuelli, N. Filippa, S. Giorgetti-Peraldi, P. Lebrun, I. Mothe-Satney, P. Peraldi, S. Rocchi, D. Sawka-Verhelle, S. Tartare-Deckert, and J. Giudicelli, "Surfing the insulin signaling," *J. Clin. Invest*, vol. 31, pp. 966–977, 2001.
- [208] J. L. Evans, I. D. Goldfine, B. A. Maddux, and G. M. Grodsky, "Oxidative stress and stressactivated signaling pathways: a unifying hypothesis of type 2 diabetes." *Endocr Rev.*, vol. 23, no. 5, pp. 599–622, 2002.

- [209] G. Liu and C. M. Rondinone, “Jnk: bridging the insulin signaling and inflammatory pathway,” *Opin Curr Investig. Drugs*, vol. 6, no. 10, pp. 979–987, 2005.
- [210] R. Somwar, M. Perreault, S. Kapur, C. Taha, G. Sweeney, T. Ramlal, D. Y. Kim, J. Keen, C. H. Cote, A. Klip, and A. Marette, “Activation of p38 mitogen-activated protein kinase alpha and beta by insulin and contraction in rat skeletal muscle: potential role in the stimulation of glucose transport.” *Diabetes*, vol. 49, no. 11, pp. 1794–1800, 2000.
- [211] K. Mussig, H. Fiedler, H. Staiger, C. Weigert, R. Lehmann, E. D. Schleicher, and H. U. Haring, “Insulin-induced stimulation of jnk and the pi 3-kinase/mtor pathway leads to phosphorylation of serine 318 of irs-1 in c2c12 myotubes.” *Biophys. Res Biochem Commun.*, vol. 335, no. 3, pp. 819–825, 2005.
- [212] Z. Mirza, A. Ali, G. Ashraf, M. Kamal, and A. Abuzenadah, “Proteomics approaches to understand linkage between alzheimers disease and type 2 diabetes mellitus.” *CNS Neurol Disord Drug Targets.*, vol. 13, no. 3, pp. 213–225, 2014.
- [213] E. Georga, V. Protopappas, D. Polyzos, and D. Fotiadis, “Evaluation of short-term predictors of glucose concentration in type 1 diabetes combining feature ranking with regression models.” *Med Biol Eng Comput.*, vol. 53, no. 12, pp. 1305—1318, 2015.
- [214] C. Duftner, C. Goldberger, A. Falkenbach *et al.*, “Prevalence, clinical relevance and characterization of circulating cytotoxic cd4+cd28- t cells in ankylosing spondylitis,” *Arthritis Research and Therapy*, vol. 5, no. 5, pp. 292–300, 2003.
- [215] S. Kumari and A. Singh, “A data mining approach for the diagnosis of diabetes mellitus.” 2013, pp. 373–375.

Appendix A

TABLE A.1: Common risk gene variants for T2DM, identified by GWAS

Year of Discovery	Gene	Chromosomal Loci	Disease Mechanism	Study
2000	PPARG	3p25.2	Insulin sensitivity	GWAS
2003	KCNJ11/ABCC8	11p15.1	Beta cell dysfunction	GWAS
2006	TCF7L2	10q25.2	Beta cell dysfunction	GWAS
2007	WFS1	4p16.1	Beta cell dysfunction	GWAS
2007	CDKN2A/2B	9p21.3	Beta cell dysfunction	GWAS
2007	CDKAL1	6p22.3	Beta cell dysfunction	GWAS
2007	SLC30A8	8q24.11	Decreased Beta Cell Function	GWAS
2007	HHEX/IDE	10q23.33	Beta cell dysfunction	GWAS
2007	HNF1Beta	17q12	Beta Cell Dysfunction	GWAS, M.A.
2007	FTO	16q12.2	Obesity. Increased triglycerides and cholesterol	GWAS, M.A.
2007	IGF2BP2	3q28	Beta cell dysfunction	GWAS, M.A.
2008	CDC123/CAMK1D	10p13	Beta cell dysfunction	GWAS, M.A.
2008	JAZF1	7p15.1	Beta cell dysfunction	GWAS, M.A.
2008	TSPAN8/LGR5	12q21.1	Beta Cell dysfunction	GWAS, M.A.
2008	THADA	2p21	Beta Cell dysfunction	GWAS, M.A.
2008	ADAMTS9	3p14.1	Decreased insulin sensitivity	GWAS, M.A.
2008	NOTCH2	1p11.2	Unknown(Membrane receptor)	GWAS, M.A.
2008	KCNQ1	11p15.5	Beta-cell dysfunction	GWAS, M.A.
2009	IRS1	2q36.3	Increased insulin Resistance	GWAS
2010	DGKB-TMEM195	7p21.2	Decreased -cell function	GWAS, M.A.
2010	GCK	7p13	Insulin sensitivity	GWAS, M.A.
2010	GCKR	2p23.3	Increased insulin resistance	GWAS, M.A.
2010	PROX1	1q32.3	Decreased -cell function	GWAS, M.A.
2010	ADCYS	3q21.1	Decreased insulin Sensitivity	GWAS, M.A.
2010	DUSP9	Xq28	Phosphatase	GWAS, M.A.
2010	BCL11A	2p16.1	Beta-cell dysfunction	GWAS, M.A.
2010	ZBED3	5q13.3	Beta-cell dysfunction	GWAS, M.A.
2010	KLF14	7q32.3	Insulin action	GWAS, M.A.
2010	TP53INP1	8q22.1	Unknown	GWAS, M.A.

Year of Discovery	Gene	Chromosomal Loci	Disease Mechanism	Study
2010	CHCHD9	9q21.31	Unknown	GWAS, M.A.
2010	CENTD2/ ARAP1	11q13.4	Beta-cell dysfunction	GWAS, M.A.
2010	HMGA2	12q14.3	Transcriptional Regulator	GWAS, M.A.
2010	HNF1A	12q24.31	Pancreatic and liver transcriptional regulator	GWAS, M.A.
2010	ZFAND6	15q25.1	Beta-cell dysfunction	GWAS, M.A.
2010	PRC1	15q26.1	Unknown (Cytokinesis regulator)	GWAS, M.A.
2010	MTNR1B	11q14.3	Decreased -cell function	GWAS, M.A.
1967	INS	11p15.5	permanent neonatal Diabetes Mellitus	M.A.
1999	Akt	1q43-q44	Increased chemical signaling	M.A.
2004	mTOR	1p36.22	cell growth and division, cellular processes	GWAS, M.A.
2001	TGFB	19q13.2	increase in signal transduction	GWAS, M.A.
2003	FN1	2q35	Dysfunctionality	GWAS, M.A.
1987	FST	5q11.2	Inhibitory activity	M.A.
2005	STAT1	2q32.2	Signaling pathway	GWAS, M.A.
1995	IRS2	13q34	Phosphorylation	M.A.
2013	HNF4A	20q13.12	Transcriptional regulation	GWAS, M.A.
2007	SIRT1	10q21.3	Unknown	GWAS, M.A.
2004	LEP	7q32.1	reduced production of hormones	GWAS, M.A.
2012	NF-KB	1p34.3	Cell proliferation	GWAS, M.A.
1993	JNK	15q15.1	Dysfunctional	M.A.
1989	GluT4	17p13.1	Decreased Beta Cell Functionality	M.A.
2009	PLcBeta	15q15.1	Signal Transduction pathway	GWAS, M.A.
1995	NeuroD	2q31.3	Transcriptional regulation	M.A.

TABLE A.2: Akt Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Akt	→	AS160	1.04	0.77	2.60
2	Akt	→	PDE-3B	1	0.69	2.42
3	Akt	→	GSK-3	1.02	0.73	1.76
4	Akt	→	Foxa1/2	0.97	0.67	1.25
5	Akt	→	PDX1	1.06	0.63	1.09
6	Akt	→	FST	0.80	0.71	2.93
7	Akt	→	BAD	0.75	0.79	2.9
8	Akt	→	Glut4	0.83	0.75	1.5
9	Akt	→	AdipoR	1.03	0.70	1.16
10	Akt	→	INS	0.98	0.49	1.15

TABLE A.3: mTOR Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	mTOR	→	TNFAlpha	0.91	0.74	1.19
2	mTOR	→	IRS1	0.85	0.72	1.10
3	mTOR	→	HNF4Alpha	0.76	0.69	1.34

TABLE A.4: TGFBR Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	TGFBR	→	TGFBR1/2	0.97	0.77	1.27
2	TGFBR	→	Foxa1/2	1.03	0.70	1.16
3	TGFBR	→	FST	0.93	0.69	1.23

TABLE A.5: TGFB Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	TGFB	\rightarrow	TGFBR	1	0.77	1.30

TABLE A.6: THADA Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	THADA	\rightarrow	PPARG	0.73	0.49	1.8
2	THADA	\rightarrow	CAMK1D	0.65	0.54	2.29
3	THADA	\rightarrow	JAZF1	0.84	0.63	2.41

TABLE A.7: P300 Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	P300	\rightarrow	P300-SP1	0.89	0.68	2.12

TABLE A.8: P300-SP1 Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	P300-SP1	\rightarrow	P53	0.97	0.71	1.21
2	P300-SP1	\rightarrow	P21	0.85	0.76	1.32

TABLE A.9: P53 Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	P53	\rightarrow	IGFBP3	0.94	0.56	1.51
2	P53	\rightarrow	P300-SP1	0.80	0.71	2.93
3	P53	\rightarrow	P21	0.75	0.79	2.9

TABLE A.10: IGFBP3 Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	IGFBP3	\rightarrow	STAT1	0.76	0.64	1.94
2	IGFBP3	\rightarrow	FN1	0.72	0.68	1.64
3	IGFBP3	\rightarrow	IGF1	0.83	0.71	2.0

TABLE A.11: FN1 Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	FN1	\rightarrow	FST	0.87	0.55	2.07
2	FN1	\rightarrow	Akt	0.94	0.59	1.22
3	FN1	\rightarrow	TNFR1	0.76	0.76	1.25
4	FN1	\rightarrow	TNFAlpha	0.83	0.71	1.33
5	FN1	\rightarrow	mTOR	0.79	0.69	1.29

TABLE A.12: STAT1 Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	STAT1	\rightarrow	P21	0.91	0.62	1.18
2	STAT1	\rightarrow	BAD	1.02	0.77	1.76
3	STAT1	\rightarrow	C-MYC	0.66	0.62	3.9

TABLE A.13: MIZ1 Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	MIZ1	\rightarrow	P21	0.85	0.59	1.23
2	MIZ1	\rightarrow	UCP1	0.57	0.77	1.16

TABLE A.14: TGFBR1/2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	TGFBR1/2	→	SMAD2	0.74	0.71	1.24
2	TGFBR1/2	→	FN1	0.76	0.64	1.94
3	TGFBR1/2	→	TRAF2	0.72	0.68	1.64

TABLE A.15: SMAD2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	SMAD2	→	SMAD3	0.66	0.74	1.54
2	SMAD2	→	TRAF2	0.62	0.78	1.44
3	SMAD2	→	HNF1Alpha	0.75	0.64	1.44
4	SMAD2	→	Akt	0.78	0.75	1.34

TABLE A.16: FST Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	FST	→	Act	0.79	0.74	1.54
2	FST	→	SHC1	0.72	0.68	1.63

TABLE A.17: Act Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Act	→	TGFB	1.02	0.77	1.
2	Act	→	TGFBR2	0.82	0.74	1.16
3	Act	→	SMAD3	0.93	0.69	1.53

TABLE A.18: TRAF2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	TRAF2	→	mTOR	0.92	0.77	1.79
2	TRAF2	→	NeuroD	0.76	0.64	1.94

TABLE A.19: PIP3 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PIP3	→	mTOR	0.89	0.79	1.59

TABLE A.20: ADAMTs9 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	ADAMTs9	→	THADA	1.04	0.54	1.26
2	ADAMTs9	→	CAMK1D	0.98	0.49	1.15
3	ADAMTs9	→	Act	1.02	0.77	1.76
4	ADAMTs9	→	JAZF1	0.97	0.71	1.21
5	ADAMTs9	→	CDC123	0.85	0.76	1.32
6	ADAMTs9	→	PPARG	0.92	0.77	1.79

TABLE A.21: GCK Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	GCK	→	Foxa1/2	0.92	0.77	1.79
2	GCK	→	SLC30A8	0.94	0.55	1.26
3	GCK	→	INS	1.02	0.77	1.76
4	GCK	→	HNF4Alpha	0.97	0.71	1.21
5	GCK	→	PDX1	0.85	0.76	1.32
6	GCK	→	Hes1	0.74	0.59	1.62
7	GCK	→	NeuroD	0.93	0.73	1.67

TABLE A.22: C-MYC Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	C-MYC	→	CDKN2B	0.94	0.55	1.26
2	C-MYC	→	IL-6	1.02	0.77	1.76

TABLE A.23: P13k Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	P13k	→	Akt	0.87	0.79	1.45
2	P13k	→	IGFBP3	0.92	0.77	1.76
3	P13k	→	PDK1/2	0.79	0.87	1.39

TABLE A.24: SHIP2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	SHIP2	→	PIP3	1.04	0.54	1.26
2	SHIP2	→	IGF2BP2	0.94	0.66	1.25

TABLE A.25: PDK1/2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PDK1/2	→	APkc	1.02	0.77	1.76
2	PDK1/2	→	CDC123	0.97	0.77	1.27
3	PDK1/2	→	PDE-3B	1.03	0.70	1.16
4	PDK1/2	→	Akt	0.93	0.69	1.23

TABLE A.26: SIRT1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	SIRT1	→	Akt	1.03	0.67	1.76
2	SIRT1	→	UCP2	0.99	0.87	1.06

TABLE A.27: TCF7L2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	TCF7L2	→	Akt	1.02	0.74	1.56
2	TCF7L2	→	IRS1	1.04	0.54	1.26
3	TCF7L2	→	TNFAlpha	0.97	0.77	1.27
4	TCF7L2	→	IL-6	1.03	0.70	1.16
5	TCF7L2	→	GSK-3	0.93	0.69	1.23
6	TCF7L2	→	AdipoR	0.97	0.71	1.21
7	TCF7L2	→	PPARG	0.85	0.76	1.32
8	TCF7L2	→	HSL	0.94	0.55	1.26
9	TCF7L2	→	FTO	1.02	0.77	1.76
10	TCF7L2	→	SLC30A8	1.03	0.67	1.69
11	TCF7L2	→	GCKR	0.72	0.87	1.32

TABLE A.28: SOCS3 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	SOCS3	→	IRS1	0.97	0.76	1.59
2	SOCS3	→	GluT4	0.92	0.77	1.76

TABLE A.29: IGF1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	IGF1	→	IGF2	1.03	0.73	1.45
2	IGF1	→	TGFBR2	0.97	0.74	1.56
3	IGF1	→	MIZ1	0.93	0.72	1.66

TABLE A.30: CDKN2A Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	CDKN2A	→	HNF4Alpha	0.94	0.55	1.26
2	CDKN2A	→	LEP	0.98	0.78	1.66

TABLE A.31: TNFAlpha Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	TNFAlpha	→	TNFR1	0.83	0.67	1.56
2	TNFAlpha	→	Ngn3	0.78	0.69	1.69

TABLE A.32: SLC30A8 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	SLC30A8	→	HNF4Alpha	0.97	0.71	1.21
2	SLC30A8	→	ObR	0.85	0.76	1.32

TABLE A.33: TNFR1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	TNFR1	→	NF-KB	0.89	0.71	1.36
2	TNFR1	→	GCK	0.93	0.69	1.23

TABLE A.34: IGF2BP2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	IGF2BP2	→	SLC30A8	0.97	0.89	1.54
2	IGF2BP2	→	JAZF1	0.94	0.66	1.25
3	IGF2BP2	→	HHEX	0.87	0.65	1.34

TABLE A.35: CDKAL1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	CDKAL1	→	IGF2BP2	0.85	0.67	1.23
2	CDKAL1	→	SLC30A8	0.543	0.75	1.61
3	CDKAL1	→	WFS1	1.02	0.77	1.76
4	CDKAL1	→	PEPCK	1.06	0.67	1.46
5	CDKAL1	→	HHEX	1.23	0.72	1.56
6	CDKAL1	→	FTO	0.96	0.76	1.64

TABLE A.36: MTNR1B Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	MTNR1B	→	DGKB	0.97	0.71	1.21
2	MTNR1B	→	CAMK1D	0.85	0.76	1.32
3	MTNR1B	→	PL1N	0.94	0.66	1.25
4	MTNR1B	→	HNF4Alpha	0.82	0.73	1.62

TABLE A.37: PROX1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PROX1	→	HNF1A	1.02	0.77	1.76
2	PROX1	→	G6PC	0.99	0.72	1.45

TABLE A.38: HHEX Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	HHEX	→	Foxa1/2	0.86	0.73	1.37
2	HHEX	→	CDKN2B	0.92	0.71	1.26
3	HHEX	→	SLC30A8	0.89	0.71	1.32

TABLE A.39: NOTCH Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	NOTCH	→	HNF1B	0.85	0.67	1.23
2	NOTCH	→	Hes1	0.543	0.75	1.61
3	NOTCH	→	CAMK1D	1.02	0.77	1.76
4	NOTCH	→	Sox9	0.97	0.71	1.21
5	NOTCH	→	Ngn3	0.85	0.76	1.25

TABLE A.40: GL1S3 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	GL1S3	→	Ngn3	0.85	0.76	1.32
2	GL1S3	→	BCL11A	0.93	0.67	1.22

TABLE A.41: LEP Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	LEP	→	ObR	0.94	0.66	1.25
2	LEP	→	PROX1	0.92	0.71	1.36

TABLE A.42: ACDC Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	ACDC	\rightarrow	AdipoR	0.95	0.72	1.56
2	ACDC	\rightarrow	EGFR	0.82	0.75	1.16

TABLE A.43: NF-KB Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	NF-KB	\rightarrow	IL-6	0.97	0.77	1.27
2	NF-KB	\rightarrow	IP3R	1.03	0.70	1.16
3	NF-KB	\rightarrow	P13K	0.93	0.69	1.23

TABLE A.44: C-JUN Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	C-JUN	\rightarrow	DUSP9	1.02	0.59	2.37
2	C-JUN	\rightarrow	Mafa	1.04	0.54	1.26

TABLE A.45: GNRHR Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	GNRHR	\rightarrow	PLA2	0.92	0.78	1.56
2	GNRHR	\rightarrow	GnRH	1.02	0.69	1.49

TABLE A.46: GluT4 Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	GluT4	\rightarrow	Grb2	0.99	0.87	1.06
2	GluT4	\rightarrow	ATGL	0.87	0.53	2.36

TABLE A.47: EGFR Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	EGFR	→	Src	0.78	0.69	1.69
2	EGFR	→	TCF7L2	0.98	0.72	1.36

TABLE A.48: GnRH Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	GnRH	→	Gq/11	0.97	0.71	1.21
2	GnRH	→	PAA	0.85	0.76	1.32
3	GnRH	→	GS	0.82	0.67	1.56

TABLE A.49: PAA Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PAA	→	GCKR	0.85	0.76	1.32
2	PAA	→	PLA2	0.92	0.78	1.56

TABLE A.50: Gq/11 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Gq/11	→	PLcBeta	1.02	0.77	1.56
2	Gq/11	→	PAA	0.82	0.76	1.32

TABLE A.51: PLcBeta Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PLcBeta	→	PKC	0.97	0.77	1.27
2	PLcBeta	→	AC	1.03	0.70	1.16
3	PLcBeta	→	IP3R	0.93	0.69	1.23

TABLE A.52: NeuroD Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	NeuroD	→	INS	0.85	0.67	1.23
2	NeuroD	→	TSHR	0.543	0.75	1.61
3	NeuroD	→	Sox9	1.02	0.72	1.49

TABLE A.53: Foxa1/2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Foxa1/2	→	Ngn3	0.92	0.78	1.56
2	Foxa1/2	→	PDX1	0.89	0.72	1.67
3	Foxa1/2	→	HNF4Alpha	0.97	0.71	1.21
4	Foxa1/2	→	PyK	0.85	0.76	1.32
5	Foxa1/2	→	IGF2	1.01	0.73	1.76
6	Foxa1/2	→	GK	0.97	0.77	1.27
7	Foxa1/2	→	TGFBR2	1.03	0.70	1.16
8	Foxa1/2	→	FBP	0.93	0.69	1.23
9	Foxa1/2	→	PEPCK	1.03	0.67	1.16
10	Foxa1/2	→	IGF1	0.93	0.69	1.27
11	Foxa1/2	→	G6PC	0.96	0.77	1.76
12	Foxa1/2	→	IRS	1.02	0.77	1.36

TABLE A.54: PPARG Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PPARG	→	FTO	0.91	0.67	1.56
2	PPARG	→	INS	0.97	0.77	1.27

TABLE A.55: Ptf1a Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Ptf1a	→	PDX1	1.02	0.77	1.76
2	Ptf1a	→	Ngn3	0.97	0.77	1.27
3	Ptf1a	→	FABP	1.03	0.70	1.16
4	Ptf1a	→	NeuroD	0.93	0.69	1.23

TABLE A.56: HNF6Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	HNF6	→	HNF1Beta	1.02	0.77	1.76
2	HNF6	→	PDX1	1.03	0.70	1.16
3	HNF6	→	HNF4Alpha	0.97	0.72	1.46
4	HNF6	→	Ngn3	0.93	0.69	1.23

TABLE A.57: HNF1Beta Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	HNF1Beta	→	HNF4Alpha	1.02	0.67	1.46
2	HNF1Beta	→	CMYC	1.03	0.70	1.16
3	HNF1Beta	→	Ngn3	0.93	0.69	1.23

TABLE A.58: HNF4Alpha Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	HNF4Alpha	→	HNF6	0.93	0.69	1.23
2	HNF4Alpha	→	GCKR	1.02	0.77	1.76
3	HNF4Alpha	→	PPARG	1.03	0.70	1.16
4	HNF4Alpha	→	HNF1Alpha	0.93	0.69	1.23

TABLE A.59: Ngn3 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Ngn3	→	NeuroD	0.93	0.69	1.23
2	Ngn3	→	Ngn3	1.02	0.77	1.76
3	Ngn3	→	FTO	0.85	0.67	1.23
4	Ngn3	→	PDX1	0.85	0.67	1.23
5	Ngn3	→	HNF6	0.543	0.75	1.61
6	Ngn3	→	CAPN10	0.92	0.67	1.36

TABLE A.60: HNF1Alpha Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	HNF1Alpha	→	PDX1	1.02	0.77	1.76
2	HNF1Alpha	→	STAT1	0.85	0.67	1.23
3	HNF1Alpha	→	HNF4Alpha	0.543	0.75	1.61

TABLE A.61: Mafa Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Mafa	→	PDX1	1.02	0.77	1.76
2	Mafa	→	GK	0.85	0.67	1.23

TABLE A.62: PDX1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PDX1	→	INS	1.02	0.77	1.76
2	PDX1	→	Hes1	0.85	0.67	1.23
3	PDX1	→	TSH	0.97	0.77	1.27
4	PDX1	→	Ngn3	1.03	0.70	1.16
5	PDX1	→	PDX1	0.93	0.69	1.23

TABLE A.63: Hes1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Hes1	→	Ngn3	1.03	0.70	1.16
2	Hes1	→	Mafa	0.93	0.69	1.23
3	Hes1	→	NOTCH	1.02	0.77	1.76

TABLE A.64: JAZF1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	JAZF1	→	PPARG	1.02	0.77	1.76
2	JAZF1	→	C-JUN	1.03	0.70	1.16
3	JAZF1	→	CAMK1D	0.93	0.69	1.23
4	JAZF1	→	CDC123	0.92	0.67	1.36

TABLE A.65: CDC123 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	CDC123	→	PPARG	0.93	0.69	1.23
2	CDC123	→	ACC	1.02	0.77	1.76
3	CDC123	→	Ngn3	1.03	0.70	1.16
4	CDC123	→	DGKB	0.96	0.79	1.23

TABLE A.66: CAMk1D Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	CAMk1D	→	CDC123	0.93	0.69	1.23
2	CAMK1D	→	PKa	1.02	0.77	1.76
3	CAMK1D	→	FAS	0.95	0.79	1.12

TABLE A.67: aPKC Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	aPKC	→	SREBP-1C	0.85	0.67	1.23
2	aPKC	→	FBP	1.02	0.77	1.76

TABLE A.68: ADCYS Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	ADCYS	→	CAMK1D	0.85	0.67	1.53
2	ADCYS	→	Pka	0.543	0.75	1.51
3	ADCYS	→	Gq/11	0.89	0.74	1.46
4	ADCYS	→	CDC123	0.59	0.71	1.69

TABLE A.69: IL-6 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	IL-6	→	C-JUN	0.78	0.69	1.54
2	IL-6	→	Src	1.04	0.54	1.26
3	IL-6	→	IL-6R	1.02	0.73	1.76
4	IL-6	→	UCP2	0.94	0.64	1.62
5	IL-6	→	TCF7L2	0.97	0.77	1.27
6	IL-6	→	UCP1	0.93	0.74	1.36
7	IL-6	→	UCP3	1.03	0.70	1.16
8	IL-6	→	FTO	0.99	0.69	1.23

TABLE A.70: IL-6R Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	IL-6R	→	AMPK	1.03	0.70	1.16
2	IL-6R	→	SOCS3	0.92	0.73	1.67

TABLE A.71: ObR Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	ObR	→	AMPK	0.97	0.77	1.27
2	ObR	→	MTNR1B	1.03	0.70	1.16
3	ObR	→	GnRHR	0.93	0.69	1.23

TABLE A.72: AdipoR Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	AdipoR	→	AMPK	0.97	0.67	1.76

TABLE A.73: IP3R Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	IP3R	→	PLA2	0.93	0.69	1.23
2	IP3R	→	BAR	0.96	0.73	1.56

TABLE A.74: PKC Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PKC	→	PLA2	1.03	0.70	1.16
2	PKC	→	P53	0.93	0.69	1.23
3	PKC	→	Src	1.02	0.77	1.76

TABLE A.75: PLA2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PLA2	→	PLA2	1.04	0.65	1.26
2	PLA2	→	P300	0.99	0.77	1.76
3	PLA2	→	GS	0.94	0.54	1.66

TABLE A.76: Sox9 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Sox9	→	HNF6	1.04	0.54	1.26
2	Sox9	→	PDK1/2	0.94	0.66	1.25
3	Sox9	→	Ngn3	0.99	0.74	1.46

TABLE A.77: TSH Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	TSH	→	TSHR	0.82	0.67	1.39
2	TSH	→	GS	1.02	0.77	1.76

TABLE A.78: TSHR Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	TSHR	→	GS	0.94	0.72	1.49
2	TSHR	→	GSK3	0.89 0.69	1.52	

TABLE A.79: UCP3 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	UCP3	→	UCP1	0.92	0.67	1.46
2	UCP3	→	UCP2	0.89	0.71	1.67
3	UCP3	→	Sox9	0.95	0.76	1.34

TABLE A.80: UCP2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	UCP2	→	UCP1	0.92	0.78	1.56
2	UCP2	→	HNF6	1.01	0.73	1.43

TABLE A.81: BCL11A Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	BCL11A	\rightarrow	JNK1	0.87	0.74	1.16
2	BCL11A	\rightarrow	Sox9	0.93	0.69	1.63

TABLE A.82: PkG Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	PkG \rightarrow	JNK2	0.85	0.67	1.23	
2	PkG \rightarrow	HSL	0.53	0.75	1.61	

TABLE A.83: PkG Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	PkG \rightarrow	JNK2	0.85	0.67	1.23	
2	PkG \rightarrow	HSL	0.73	0.75	1.21	

TABLE A.84: BAR Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	BAR	\rightarrow	GS	0.94	0.66	1.25
2	BAR	\rightarrow	TSH	0.82	0.69	1.46

TABLE A.85: SREBP-1C Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	SREBP-1C	\rightarrow	ACC	0.87	0.67	1.26
2	SREBP-1C	\rightarrow	FAS	0.93	0.71	1.76
3	SREBP-1C	\rightarrow	ADCYS	1.01	0.74	1.48

TABLE A.86: AMPK Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	AMPK	→	ACC	0.85	0.76	1.32
2	AMPK	→	FAS	1.03	0.75	1.29
3	AMPK	→	AC	0.92	0.67	1.56
4	AMPK	→	GSK-3	0.94	0.54	1.26
5	AMPK	→	MTNR1B	1.02	0.77	1.76

TABLE A.87: HSL Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	HSL	→	FABP	1.02	0.77	1.76
2	HSL	→	ASK1	0.99	0.76	1.35

TABLE A.88: GS Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	GS	→	AC	0.92	0.67	1.56
2	GS	→	HHEX	0.89	0.71	1.32
3	GS	→	UCP2	1.02	0.77	1.76
4	GS	→	PL1N	0.97	0.67	1.27
5	GS	→	Akt	1.03	0.70 1.16	
6	GS	→	GnRHR	0.93	0.69	1.23

TABLE A.89: AC Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	AC	\rightarrow	Pka	0.543	0.75	1.61
2	AC	\rightarrow	DGKB	1.02	0.77	1.76
3	AC	\rightarrow	BAR	0.64	0.75	1.61
4	AC	\rightarrow	INSR	0.85	0.67	1.23
5	AC	\rightarrow	PAA	0.73	0.75	1.59

TABLE A.90: Pka Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	Pka	\rightarrow	PL1N	0.65	0.75	1.61
2	Pka	\rightarrow	HSL	1.02	0.77	1.76
3	Pka	\rightarrow	AMPK	1.03	0.70	1.16
4	Pka	\rightarrow	AP-1	0.93	0.69	1.23

TABLE A.91: GSK3 Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	GSK3	\rightarrow	GS	0.94	0.66	1.25
2	GSK3	\rightarrow	CDKN2A	1.03	0.70	1.16
3	GSK3	\rightarrow	IL-6	0.93	0.69	1.23
4	GSK3	\rightarrow	TNFA	1.02	0.67	1.49

TABLE A.92: ACC Subgroup and its association with other genes

Rule	Antecedent	\rightarrow	Consequent	Confidence	Support	Lift
1	ACC	\rightarrow	CAPN10	1.04	0.54	1.26

TABLE A.93: CAPN10 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	CAPN10	→	PKG	0.97	0.69	1.45
2	CAPN10	→	DGKB	0.92	0.77	1.32

TABLE A.94: WFS1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	WFS1	→	Pka	1.02	0.77	1.76

TABLE A.95: BAD Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	BAD	→	WFS1	1.04	0.54	1.26
2	BAD	→	PEPCK	1.01	0.67	1.36

TABLE A.96: DUSP9 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	DUSP9	→	GrB2	1.01	0.74	1.28

TABLE A.97: GrB2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	GrB2	→	CGI-58	0.82	0.81	1.76

TABLE A.98: SMAD3 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	SMAD3	→	HNF1Beta	1.04	0.54	1.26
2	SMAD3	→	SMAD2	1.03	0.75	1.29
3	SMAD3	→	P300	1.02	0.77	1.76

TABLE A.99: G6PC Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	G6PC	→	SOCS3	0.85	0.67	1.23
2	G6PC	→	FBP	0.543	0.75	1.61

TABLE A.100: UCP1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	UCP1	→	Mc4R	0.79	0.73	1.37

TABLE A.101: Mc4R Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Mc4R	→	PPARG	1.02	0.77	1.76

TABLE A.102: RXR Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	RXR	→	AS160	0.53	0.71	1.16

TABLE A.103: GK Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	GK	→	RXR	0.92	0.78	1.56
2	GK	→	G6PC	0.63	0.75	1.61

TABLE A.104: AP-1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	AP-1	→	aPKC	0.92	0.74	1.76

TABLE A.105: IGF2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	IGF2	→	FST	0.89	0.77	1.76
2	IGF2	→	SIRT1	0.93	0.75	1.61

TABLE A.106: PyK Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PyK	→	GnRH	0.94	0.66	1.25
2	PyK	→	FBP	0.85	0.67	1.23
3	PyK	→	PEPCK	0.63	0.75	1.61
4	PyK	→	G6PC	1.04	0.54	1.26
5	PyK	→	GK	0.93	0.71	1.39

TABLE A.107: AdPLA Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	AdPLA	→	PyK	0.82	0.71	1.26
2	AdPLA	→	FABP	0.88	0.76	1.19

TABLE A.108: ASK1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	ASK1	→	AdPLA	0.80	0.71	1.33

TABLE A.109: SHC1 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	SHC1	→	UCP3	1.03	0.79	1.64

TABLE A.110: CGI-58 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	CGI-58	→	P53	0.86	0.69	1.43
2	CGI-58	→	ATGL	1.02	0.77	1.76

TABLE A.111: PEPCK Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PEPCK	→	NFKB	0.94	0.69	1.45
2	PEPCK	→	FAS	0.82	0.73	2.13

TABLE A.112: GnRH Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	GnRH	→	PLcBeta	0.86	0.72	1.67

TABLE A.113: PL1N Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PL1N	→	DUSP9	1.02	0.67	1.54
2	PL1N	→	ATGL	1.04	0.54	1.26
3	PL1N	→	CG1-58	0.80	0.71	2.93

TABLE A.114: CDKN2B Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	CDKN2B	→	CDKAL1	0.80	0.77	1.34

TABLE A.115: DGKB Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	DGKB	→	SREBP-1C	1.03	0.67	1.56

TABLE A.116: JNK1/2 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	JNK1/2	→	PKC	1.03	0.70	1.16
2	JNK1/2	→	Src	0.93	0.69	1.23

TABLE A.117: Src Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	Src	→	TNFR1	0.92	0.67	1.56
2	Src	→	EGFR	0.87	0.62	1.34

TABLE A.118: AS-160 Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	AS-160	→	FN1	0.86	0.79	1.75

TABLE A.119: ATGL Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	ATGL	→	AC	1.04	0.76	1.25

TABLE A.120: GCKR Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	GCKR	→	SHIP2	0.87	0.57	1.13
2	GCKR	→	NOTCH	0.59	0.73	1.06

TABLE A.121: FABP Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	FABP	→	ACDC	1.06	0.63	1.09
2	FABP	→	CGI-58	1.01	0.67	1.24

TABLE A.122: FBP Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	FBP	→	PEPCK	0.87	0.69	1.34

TABLE A.123: ACC Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	ACC	→	Akt	0.98	0.49	1.15

TABLE A.124: PDE-3B Subgroup and its association with other genes

Rule	Antecedent	→	Consequent	Confidence	Support	Lift
1	PDE-3B	→	AC	1.04	0.54	1.26

TABLE A.125: Transitive association of INS with other gene variants

Antecedent	→	Consequent
INS	→	IRS1
INS	→	PL1N
INS	→	AdPLA
INS	→	HNF1A
INS	→	HNF4A
INS	→	Ngn3
INS	→	PDX1
INS	→	PyK
INS	→	GK
INS	→	FBP
INS	→	PEPCK
INS	→	G6PC

TABLE A.126: Transitive association of INSR with other gene variants

Antecedent	→	Consequent
INSR	→	P13K
INSR	→	SHC1
INSR	→	PDX1

TABLE A.127: Transitive association of IRS1 with other gene variants

Antecedent	→	Consequent
IRS1	→	Akt
IRS1	→	PDK1/2

TABLE A.128: Transitive association of Akt with other gene variants

Antecedent	→	Consequent
Akt	→	GS
Akt	→	NGN3
Akt	→	HNF4Alpha
Akt	→	PyK
Akt	→	GK
Akt	→	FBP
Akt	→	PEPCK
Akt	→	G6PC
Akt	→	ACT
Akt	→	GrB2

TABLE A.129: Transitive association of mTOR with other gene variants

Antecedent	→	Consequent
mTOR	→	TNFR1
mTOR	→	P13K
mTOR	→	SHC1

TABLE A.130: Transitive association of THADA with other gene variants

Antecedent	→	Consequent
THADA	→	FTO
THADA	→	INS
THADA	→	CDC123
THADA	→	Pka

TABLE A.131: Transitive association of P300 with other gene variants

Antecedent	→	Consequent
P300	→	P53

TABLE A.132: Transitive association of ADAMTs9 with other gene variants

Antecedent	→	Consequent
ADAMTs9	→	Pka
ADAMTs9	→	FTO
ADAMTs9	→	INS

TABLE A.133: Transitive association of P53 with other gene variants

Antecedent	→	Consequent
P53	→	STAT1
P53	→	FN1
P53	→	IGF1

TABLE A.134: Transitive association of IGFBP3 with other gene variants

Antecedent	→	Consequent
IGFBP3	→	P21
IGFBP3	→	C-MYC
IGFBP3	→	FST
IGFBP3	→	Akt
IGFBP3	→	IGF2

TABLE A.135: Transitive association of FN1 with other gene variants

Antecedent	→	Consequent
FN1	→	Act
FN1	→	AS160
FN1	→	PDE-3B
FN1	→	GSK-3
FN1	→	Foxa1/2
FN1	→	BAD
FN1	→	GluT4

TABLE A.136: Transitive association of FST with other gene variants

Antecedent	→	Consequent
FST	→	TGFB

TABLE A.137: Transitive association of Act with other gene variants

Antecedent	→	Consequent
Act	→	TGFBR

TABLE A.138: Transitive association of TGF β with other gene variants

Antecedent	\rightarrow	Consequent
TGF β	\rightarrow	TGFBR1/2

TABLE A.139: Transitive association of TGFBR with other gene variants

Antecedent	\rightarrow	Consequent
TGFBR	\rightarrow	SMAD2

TABLE A.140: Transitive association of TGFBR1/2 with other gene variants

Antecedent	\rightarrow	Consequent
TGFBR1/2 \rightarrow NF-KB		

TABLE A.141: Transitive association of TRAF2 with other gene variants

Antecedent	\rightarrow	Consequent
TRAF2	\rightarrow	TNF α
TRAF2	\rightarrow	IRS1

TABLE A.142: Transitive association of SIRT1 with other gene variants

Antecedent	\rightarrow	Consequent
SIRT1	\rightarrow	FST
SIRT1	\rightarrow	BAD
SIRT1	\rightarrow	GluT4

TABLE A.143: Transitive association of GCK with other gene variants

Antecedent	→	Consequent
GCK	→	Ngn3
GCK	→	PyK
GCK	→	GK
GCK	→	FBP
GCK	→	PEPCK
GCK	→	G6PC
GCK	→	INSR
GCK	→	CDKN2A
GCK	→	JNK1/2
GCK	→	HNF6
GCK	→	PPARG
GCK	→	HNF1Alpha

TABLE A.144: Transitive association of STAT1 with other gene variants

Antecedent	→	Consequent
STAT1	→	CDKN2B

TABLE A.145: Transitive association of SLC30A8 with other gene variants

Antecedent	→	Consequent
SLC30A8	→	HNF6
SLC30A8	→	PPARG
SLC30A8	→	HNF1Alpha

TABLE A.146: Transitive association of IGF2BP2 with other gene variants

Antecedent	→	Consequent
IGF2BP2	→	HNF4Alpha
IGF2BP2	→	Foxa1/2
IGF2BP2	→	SLC30A8

TABLE A.147: Transitive association of TCF7L2 with other gene variants

Antecedent	→	Consequent
TCF7L2	→	AS160
TCF7L2	→	PDE-3B
TCF7L2	→	Foxa1/2
TCF7L2	→	FST
TCF7L2	→	BAD
TCF7L2	→	GluT4
TCF7L2	→	P13K
TCF7L2	→	SHC1
TCF7L2	→	TNFR1
TCF7L2	→	C-JUN
TCF7L2	→	Src
TCF7L2	→	IL-6R
TCF7L2	→	UCP1
TCF7L2	→	UCP2
TCF7L2	→	UCP3
TCF7L2	→	GS
TCF7L2	→	AMPK
TCF7L2	→	INS
TCF7L2	→	MC4R
TCF7L2	→	HNF4Alpha

TABLE A.148: Transitive association of CDKN2A with other gene variants

Antecedent	→	Consequent
CDKN2A	→	HNF6
CDKN2A	→	PPARG
CDKN2A	→	HNF1Alpha

TABLE A.149: Transitive association of GLIS3 with other gene variants

Antecedent	→	Consequent
GLIS3	→	NeuroD
GLIS3	→	HNF6

TABLE A.150: Transitive association of NOTCH with other gene variants

Antecedent	→	Consequent
NOTCH	→	HNF4Alpha
NOTCH	→	HNF6

TABLE A.151: Transitive association of P53 with other gene variants

Antecedent	→	Consequent
P53	→	STAT1
P53	→	FN1
P53	→	IGF1

TABLE A.152: Transitive association of HNF1Beta with other gene variants

Antecedent	→	Consequent
HNF1Beta	→	HNF6
HNF1Beta	→	PPARG
HNF1Beta	→	HNF1Alpha
HNF1Beta	→	NeuroD

TABLE A.153: Transitive association of HNF6 with other gene variants

Antecedent	→	Consequent
HNF6	→	INS
HNF6	→	Hes1
HNF6	→	MafA
HNF6	→	PPARG
HNF6	→	HNF6
HNF6	→	HNF1Alpha
HNF6	→	NeuroD

TABLE A.154: Transitive association of NeuroD with other gene variants

Antecedent	→	Consequent
NeuroD	→	INSR
NeuroD	→	CDKN2A
NeuroD	→	Foxa1/2
NeuroD	→	JNK1/2

TABLE A.155: Transitive association of Hes1 with other gene variants

Antecedent	→	Consequent
Hes1	→	NeuroD
Hes1	→	Ngn3
Hes1	→	HN F6

TABLE A.156: Transitive association of Sox9 with other gene variants

Antecedent	→	Consequent
Sox9	→	HNF1Beta
Sox9	→	PDX1
Sox9	→	HNF4Alpha
Sox9	→	Ngn3
Sox9	→	NeuroD
Sox9	→	HNF6

TABLE A.157: Transitive association of Ngn3 with other gene variants

Antecedent	→	Consequent
Ngn3	→	INS
Ngn3	→	NeuroD
Ngn3	→	Ngn3
Ngn3	→	HNF6
Ngn3	→	HNF1Beta
Ngn3	→	PDX1
Ngn3	→	HNF4Alpha

TABLE A.158: Transitive association of HNF1Alpha with other gene variants

Antecedent	→	Consequent
HNF1Alpha	→	INS
HNF1Alpha	→	Hes1
HNF1Alpha	→	Ngn3
HNF1Alpha	→	PDX1
HNF1Alpha	→	MafA

TABLE A.159: Transitive association of HNF4Alpha with other gene variants

Antecedent	→	Consequent
HNF4Alpha	→	HNF1Beta
HNF4Alpha	→	PDX1
HNF4Alpha	→	HNF4Alpha
HNF4Alpha	→	Ngn3
HNF4Alpha	→	FTO
HNF4Alpha	→	INS

TABLE A.160: Transitive association of PDX1 with other gene variants

Antecedent	→	Consequent
PDX1	→	INSR
PDX1	→	CDKN2A
PDX1	→	Foxa1/2
PDX1	→	JNK1/2
PDX1	→	Ngn3
PDX1	→	NeuroD
PDX1	→	HNF6
PDX1	→	INS
PDX1	→	Hes1
PDX1	→	PDX1

TABLE A.161: Transitive association of Ptf1a with other gene variants

Antecedent	→	Consequent
Ptf1a	→	INS
Ptf1a	→	Hes1
Ptf1a	→	Ngn3
Ptf1a	→	PDX1
Ptf1a	→	MafA

TABLE A.162: Transitive association of MafA with other gene variants

Antecedent	→	Consequent
MafA	→	INS
MafA	→	Hes1
MafA	→	Ngn3
MafA	→	PDX1
MafA	→	MafA

TABLE A.163: Transitive association of CDKAL1 with other gene variants

Antecedent	→	Consequent
CDKAL1	→	SLC30A8
CDKAL1	→	HHEX
CDKAL1	→	HNFA1alpha
CDKAL1	→	Foxa1/2
CDKAL1	→	Mc4R
CDKAL1	→	UCP1
CDKAL1	→	UCP2
CDKAL1	→	UCP3

TABLE A.164: Transitive association of HHEX with other gene variants

Antecedent	→	Consequent
HHEX	→	Ngn3
HHEX	→	PDX1
HHEX	→	HNF4A1alpha
HHEX	→	PyK
HHEX	→	GK
HHEX	→	FBP
HHEX	→	PEPCK
HHEX	→	G6PC

TABLE A.165: Transitive association of PROX1 with other gene variants

Antecedent	→	Consequent
PROX1	→	PDX1

TABLE A.166: Transitive association of IL-6 with other gene variants

Antecedent	→	Consequent
IL-6	→	DUSP9
IL-6	→	SOCS3
IL-6	→	UCP1
IL-6	→	UCP2
IL-6	→	mC4R
IL-6	→	UCP3

TABLE A.167: Transitive association of LEP with other gene variants

Antecedent	→	Consequent
LEP	→	AMPK

TABLE A.168: Transitive association of IL-6R with other gene variants

Antecedent	→	Consequent
IL-6R	→	IRS1

TABLE A.169: Transitive association of ObR with other gene variants

Antecedent	→	Consequent
ObR	→	ACC
ObR	→	FAS
ObR	→	GSK-3

TABLE A.170: Transitive association of SOCS3 with other gene variants

Antecedent	→	Consequent
SOCS3	→	P13K
SOCS3	→	SHC1

TABLE A.171: Transitive association of ACDC with other gene variants

Antecedent	→	Consequent
ACDC	→	AMPK

TABLE A.172: Transitive association of AdipoR with other gene variants

Antecedent	→	Consequent
AdipoR	→	ACC
AdipoR	→	FAS
AdipoR	→	GSK-3

TABLE A.173: Transitive association of NF-KB with other gene variants

Antecedent	→	Consequent
NF-KB	→	C-JUN
NF-KB	→	Src
NF-KB	→	IL-6R
NF-KB	→	UCP1
NF-KB	→	UCP2
NF-KB	→	UCP3
NF-KB	→	FTO

TABLE A.174: Transitive association of GnRHR with other gene variants

Antecedent	→	Consequent
GnRHR	→	PAA
GnRHR	→	Gq/11
GnRHR	→	GS

TABLE A.175: Transitive association of GnRH with other gene variants

Antecedent	→	Consequent
GnRH	→	PLA2
GnRH	→	PLcBeta
GnRH	→	AC
GnRH	→	UCP2

TABLE A.176: Transitive association of PAA with other gene variants

Antecedent	→	Consequent
PAA	→	PLA2

TABLE A.177: Transitive association of Gq/11 with other gene variants

Antecedent	→	Consequent
Gq/11	→	Pkc
Gq/11	→	IP3R

TABLE A.178: Transitive association of PLA2 with other gene variants

Antecedent	→	Consequent
PLA2	→	PLA2

TABLE A.179: Transitive association of PLcBeta with other gene variants

Antecedent	→	Consequent
PLcBeta	→	PLA2
PLcBeta	→	Src

TABLE A.180: Transitive association of Pkc with other gene variants

Antecedent	→	Consequent
Pkc	→	PLA2

TABLE A.181: Transitive association of IP3R with other gene variants

Antecedent	→	Consequent
IP3R	→	PLA2

TABLE A.182: Transitive association of TNFR1 with other gene variants

Antecedent	→	Consequent
TNFR1	→	IL-6

TABLE A.183: Transitive association of PPARG with other gene variants

Antecedent	→	Consequent
PPARG	→	MC4R
PPARG	→	UCP1
PPARG	→	UCP2
PPARG	→	UCP3
PPARG	→	INSR
PPARG	→	CDKN2A
PPARG	→	Foxa1/2
PPARG	→	JNK1/2

TABLE A.184: Transitive association of AMPK with other gene variants

Antecedent	→	Consequent
AMPK	→	GS

TABLE A.185: Transitive association of GSK-3 with other gene variants

Antecedent	→	Consequent
GSK-3	→	AC
GSK-3	→	UCP2

TABLE A.186: Transitive association of P13k with other gene variants

Antecedent	→	Consequent
P13k	→	AS160
P13k	→	PDE-3B
P13k	→	GSK-3
P13k	→	Foxa1/2
P13k	→	FST
P13k	→	BAD
P13k	→	GluT4
P13k	→	aPKC
P13k	→	Akt

TABLE A.187: Transitive association of PDK1/2 with other gene variants

Antecedent	→	Consequent
PDK1/2	→	SREBP-1C
PDK1/2	→	AS160
PDK1/2	→	PDE-3B
PDK1/2	→	GSK-3
PDK1/2	→	Foxa1/2
PDK1/2	→	FST
PDK1/2	→	BAD
PDK1/2	→	GluT4

TABLE A.188: Transitive association of aPKC with other gene variants

Antecedent	→	Consequent
aPKC	→	ACC
aPKC	→	FAS

TABLE A.189: Transitive association of SHIP2 with other gene variants

Antecedent	→	Consequent
SHIP2	→	mTOR

TABLE A.190: Transitive association of PIP3 with other gene variants

Antecedent	→	Consequent
PIP3	→	TNFAlpha
PIP3	→	IRS1

TABLE A.191: Transitive association of JAZF1 with other gene variants

Antecedent	→	Consequent
JAZF1	→	FTO
JAZF1	→	INS
JAZF1	→	CDC123
JAZF1	→	PKA
JAZF1	→	PPARG

TABLE A.192: Transitive association of CAMK1D with other gene variants

Antecedent	→	Consequent
CAMK1D	→	PPARG

TABLE A.193: Transitive association of CDC123 with other gene variants

Antecedent	→	Consequent
CDC123	→	FTO
CDC123	→	INS

TABLE A.194: Transitive association of ADCYS with other gene variants

Antecedent	→	Consequent
ADCYS	→	CDC123
ADCYS	→	PL1n
ADCYS	→	HSL
ADCYS	→	AMPK

TABLE A.195: Transitive association of PKa with other gene variants

Antecedent	→	Consequent
PKa	→	FABP
PKa	→	ACC
PKa	→	FAS
PKa	→	GSK-3

TABLE A.196: Transitive association of UCP3 with other gene variants

Antecedent	→	Consequent
UCP3	→	UCP1

TABLE A.197: Transitive association of TSH with other gene variants

Antecedent	→	Consequent
TSH	→	GS

TABLE A.198: Transitive association of TSHR with other gene variants

Antecedent	→	Consequent
TSHR	→	AC
TSHR	→	UCP2

TABLE A.199: Transitive association of GS with other gene variants

Antecedent	→	Consequent
GS	→	Pka
GS	→	UCP1

TABLE A.200: Transitive association of AC with other gene variants

Antecedent	→	Consequent
AC	→	PL1N
AC	→	HSL
AC	→	AMPK

TABLE A.201: Transitive association of BAR with other gene variants

Antecedent	→	Consequent
BAR	→	UCP2

TABLE A.202: Transitive association of BCL11A with other gene variants

Antecedent	→	Consequent
BCL11A	→	HNF6
BCL11A	→	Ngn3

TABLE A.203: Transitive association of PKG with other gene variants

Antecedent	→	Consequent
PKG	→	FABP

TABLE A.204: List of nodes with their CC values

SN	Symbol	CC value	SN	Symbol	CC value
1.	CD4	0.297574	48.	ERAP1	0.297574
2.	JAK2	0.288625	49.	MRAP	0.288625
3.	IL8	0.285406	50.	ERAP1	0.285406
4.	CD86	0.282258	51.	B*15	0.282258
5.	ALPL	0.282258	52.	IL23R	0.282258
6.	TLR9	0.282258	53.	PTGER4	0.282258
7.	STAT3	0.281224	54.	IL17F	0.281224
8.	TNFRSF1B	0.280198	55.	LTBR	0.280198
9.	IL23R	0.280198	56.	SLC25A30	0.280198
10.	TNFRSF1A	0.278167	57.	IL22RA2	0.278167
11.	IL23A	0.277163	58.	MICA	0.277163
12.	STAT3	0.271287	59.	BSG	0.271287
13.	IL1R1	0.271287	60.	TRAF5	0.271287
14.	IL1R1	0.270332	61.	CD55	0.270332
15.	IL1A	0.270332	62.	ADA	0.270332
16.	MAPK1	0.270332	63.	LRP5	0.270332
17.	TNFSF10	0.268441	64.	HMGB1	0.268441
18.	NCAM1	0.268441	65.	TRAF5	0.268441
19.	TNFSF11	0.265655	66.	CARD9	0.265655
20.	NOD2	0.265655	67.	IL1R2	0.265655
21.	MHC-G	0.260252	68.	TIMP1	0.260252
22.	KIR3DL1	0.258499	69.	HDAC9	0.258499
23.	KIR2DL1	0.258499	70.	FCGR2B	0.258499
24.	NCAM1	0.25338	71.	IL1F10	0.25338
25.	IL8	0.24846	72.	NFATC4	0.24846
26.	LTBR	0.246071	73.	KIR3DL2	0.246071
27.	KIR3DL2	0.246071	74.	KIR2DL1	0.246071
28.	MAPK1	0.243728	75.	CARD8	0.243728
29.	MRAP	0.242956	76.	GEM	0.242956
30.	SLC25A30	0.242956	77.	MHC-G	0.242956
31.	IL12B	0.234783	78.	SERPINB1	0.234783

SN	Symbol	CC value	SN	Symbol	CC value
32.	CD86	0.226473	79.	ANTXR2	0.226473
33.	NOD2	0.225144	80.	LRP5	0.225144
34.	TNFRSF1A	0.216265	81.	PTPN22	0.216265
35.	RUNX3	0.206382	82.	ANTXR2	0.206382
36.	ADA	0.202037	83.	TNFSF10	0.202037
37.	CXCL16	0.202037	84.	RUNX3	0.202037
38.	HLA-DP	0.181929	85.	COMP	0.181929
39.	B*15	0.021505	86.	- node1	0.021505
40.	CD55	0.021505	87.	PSORS1C1	0.021505
41.	PSMD7	0.021505	88.	CYP2R1	0.021505
42.	IL23A	0.021505	89.	DHCR7	0.021505
43.	CD6	0.021505	90.	node2	0.021505
44.	IL12B	0.021505	91.	CDSN	0.021505
45.	TIMP1	0.021505	92.	CYP27B1	0.021505
46.	IL1A	0.021505	93.	CYP2R1	0.021505
47.	TNFSF11	0.297574			

TABLE A.205: List of nodes with their BC values

SN	Symbol	BC value	SN	Symbol	BC value
1.	CD4	0.406135	48.	SLC25A30	0.001155
2.	ALPL	0.120117	49.	IL12B	0.001155
3.	STAT3	0.090223	50.	LTBR	0.000855
4.	JAK2	0.087467	51.	CARD9	0.000654
5.	TNFRSF1B	0.076514	52.	CARD8	0.000585
6.	TNFRSF1A	0.076096	53.	SLC25A30	0.000546
7.	IL8	0.070691	54.	FCGR2B	0.00047
8.	IL1R1	0.069583	55.	PTGER4	0.000281
9.	CD86	0.064858	56.	node1	0
10.	MAPK1	0.05817	57.	PSORS1C1	0
11.	IL23A	0.056031	58.	LRP5	0
12.	TLR9	0.050731	59.	TNFRSF1A	0
13.	IL23R	0.049848	60.	RUNX3	0
14.	MHC-G	0.047051	61.	ADA	0
15.	STAT3	0.045567	62.	PTPN22	0
16.	ERAP1	0.044783	63.	CXCL16	0
17.	HDAC9	0.039417	64.	IL1F10	0
18.	TIMP1	0.028557	65.	NFATC4	0
19.	NOD2	0.027935	66.	HLA-DP	0
20.	IL1R2	0.023804	67.	CYP2R1	0
21.	KIR3DL1	0.023632	68.	B*15	0
22.	KIR2DL1	0.023632	69.	DHCR7	0
23.	IL23R	0.020664	70.	SERPINB1	0
24.	TNFSF11	0.020596	71.	CD55	0
25.	ANTXR2	0.019828	72.	PSMD7	0
26.	NCAM1	0.01838	73.	IL23A	0
27.	B*15	0.017511	74.	CD6	0
28.	NCAM1	0.017422	75.	TIMP1	0
29.	IL1A	0.01291	76.	node2	0
30.	IL1A	0.011089	77.	COMP	0
31.	CD86	0.010578	78.	CDSN	0

TABLE A.206: List of nodes with their BC values

SN	Symbol	BC value	SN	Symbol	BC value
32.	IL1R1	0.009889	79.	GEM	0
33.	IL8	0.007709	80.	KIR3DL2	0
34.	IL17F	0.00629	81.	MHC-G	0
35.	NOD2	0.004623	82.	CYP27B1	0
36.	TNFSF10	0.004594	83.	CYP2R1	0
37.	TNFSF11	0.004594	84.	TNFSF10	0
38.	MRAP	0.004581	85.	IL22RA2	0
39.	IL12B	0.004467	86.	HMGB1	0
40.	KIR3DL2	0.003804	87.	CD55	0
41.	MICA	0.00322	88.	KIR2DL1	0
42.	TRAF5	0.003163	89.	ANTXR2	0
43.	BSG	0.002642	90.	ADA	0
44.	ERAP1	0.002424	91.	LRP5	0
45.	MRAP	0.00235	92.	TRAF5	0
46.	LTBR	0.001403	93.	RUNX3	0
47.	MAPK1	0.001403			

TABLE A.207: The list of proteins with both high BC and large degree and their functions at the threshold of 0.01

Symbol	Function Description
CD4	CD4 is a co-receptor that assists the T cell receptor (TCR) in communicating with an antigen-presenting cell.
ALPL	Plays an important role in the growth and development of bones and teeth.
STAT3	STAT3 protein transmits signals that help control the body's response to foreign invaders such as bacteria and fungi.
JAK2	Provides instructions for making a protein that promotes the growth and division (proliferation) of cells.
TNFRSF1B	Function as a controller of inflammation.
TNFRSF1A	This protein is one of the major receptors for the tumor necrosis factor-alpha. This receptor can activate the transcription factor NF-B, mediate apoptosis, and function as a regulator of inflammation.
IL8	IL-8 is believed to play a role in the pathogenesis of bronchiolitis, a common respiratory tract disease caused by viral infection.
IL1R1	It is an important mediator involved in many cytokine induced immune and inflammatory responses.
CD86	Expressed on antigen-presenting cells that provides co-stimulatory signals necessary for T cell activation and survival.
MAPK1	Act as an integration point for multiple biochemical signals, and are involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation and development.
IL23A	IL-23 is an important part of the inflammatory response against infection. It promotes up regulation of the matrix metallo-protease MMP9, increases angiogenesis and reduces CD8+ T-cell infiltration.

Symbol	Function Description
TLR9	Plays a fundamental role in pathogen recognition and activation of innate immunity.
IL23R	Promote inflammation and help coordinate the immune system's response to foreign invaders such as bacteria and viruses.
MHC-G	Mediate cellular immunity.
ERAP1	Reduces their ability to transmit chemical signals into the cell, which affects the process of inflammation.
HDAC9	Play a role in hematopoiesis.
TIMP1	The encoded protein is able to promote cell proliferation in a wide range of cell types, and may also have an anti-apoptotic function.
NOD2	The protein is involved in recognizing certain bacteria and stimulating the immune system to respond properly.
IL1R2	Control many different cellular functions including proliferation, differentiation and cell survival/apoptosis but are also involved in several pathophysiological processes including viral infections and autoimmune diseases.
KIR3DL1	Play an important role in regulation of the immune response.
TNFSF11	This protein has a role in the regulation of cell apoptosis.
ANTXR2	This protein is involved in the formation of tiny blood vessels, important for maintaining the structure of basement membranes.
NCAM1	Involved in development of the nervous system, and for cells involved in the expansion of T cells and dendritic cells which play an important role in immune surveillance.
IL1A	It stimulates the activity of genes involved in inflammation and immunity. Plays a critical role in protecting the body from foreign invaders such as bacteria and viruses.

TABLE A.208: List of nodes with their Degree values

SN	Symbol	Degree value	SN	Symbol	Degree value
1.	CD4	33	48.	IL12B	2
2.	TNFRSF1A	15	49.	CARD8	2
3.	IL23A	13	50.	FCGR2B	2
4.	ALPL	12	51.	MAPK1	2
5.	IL1R1	11	52.	ERAP1	2
6.	MAPK1	11	53.	LTBR	2
7.	JAK2	10	54.	PTGER4	2
8.	CD86	10	55.	SLC25A30	2
9.	TLR9	9	56.	node1	1
10.	TNFRSF1B	9	57.	PSORS1C1	1
11.	IL8	9	58.	LRP5	1
12.	STAT3	8	59.	TNFRSF1A	1
13.	B*15	8	60.	RUNX3	1
14.	STAT3	8	61.	ADA	1
15.	IL23R	7	62.	PTPN22	1
16.	NOD2	7	63.	CXCL16	1
17.	MHC-G	6	64.	IL1F10	1
18.	NCAM1	6	65.	NFATC4	1
19.	TIMP1	6	66.	HLA-DP	1
20.	IL1R1	5	67.	CYP2R1	1
21.	NCAM1	5	68.	B*15	1
22.	IL12B	5	69.	DHCR7	1
23.	KIR3DL1	4	70.	SERPINB1	1
24.	KIR2DL1	4	71.	CD55	1
25.	IL1R2	4	72.	PSMD7	1
26.	IL1A	4	73.	IL23A	1
27.	CD86	4	74.	CD6	1
28.	ERAP1	4	75.	TIMP1	1
29.	IL8	4	76.	node2	1
30.	IL1A	4	77.	COMP	1
31.	TNFSF10	3	78.	CDSN	1

SN	Symbol	Degree value	SN	Symbol	Degree value
32.	KIR3DL2	3	79.	GEM	1
33.	BSG	3	80.	KIR3DL2	1
34.	IL17F	3	81.	MHC-G	1
35.	TRAF5	3	82.	CYP27B1	1
36.	HDAC9	3	83.	CYP2R1	1
37.	TNFSF11	3	84.	TNFSF10	1
38.	NOD2	3	85.	IL22RA2	1
39.	TNFSF11	3	86.	HMGB1	1
40.	IL23R	3	87.	CD55	1
41.	MRAP	3	88.	KIR2DL1	1
42.	CARD9	2	89.	ANTXR2	1
43.	LTBR	2	90.	ADA	1
44.	MRAP	2	91.	LRP5	1
45.	ANTXR2	2	92.	TRAF5	1
46.	SLC25A30	2	93.	RUNX3	1
47.	MICA	2			