Cluster-based Analysis of Affected Insulin Signaling Genes in Type 2 Diabetes Mellitus

Bilal Nizami*, Hetal Damani
Department of Biotechnology and Bioinformatics
Dr. D.Y. Patil University, CBD Belapur
Navi Mumbai 400614, INDIA

Dhani Ram Mahato
Department of Biotechnology, Birla Institute of Technology, Kolkata centre,
Kolkata 700107, INDIA

ABSTRACT

Type 2 Diabetes Mellitus (T2DM) being a complex metabolic disease is recognized as one of the potential threat to the human health in the 21st century. Etiologically it is characterized by insulin resistance and diminished insulin secretion. Advances in gene-expression studies related to T2DM have revealed altered expression of a large number of metabolic genes in a variety of tissues. Through a cluster based analysis of microarray datasets, we have identified altered genes associated with insulin signaling. We have also elucidated the application of self-organizing maps (SOMs); a type of mathematical cluster analysis technique that is pertinent for the recognition and classification features in a complex multidimensional gene-expression data. In order to investigate T2DM related alterations in expression of influenced Insulin signaling genes and transcription factors, we have implemented a network-centric methodology. It is also analyzed that these gene-sets share one or more transcription factor binding sites in the promoter regions of the corresponding genes enabling the determination of regulatory mechanisms that lead to gene expression changes in gene network. Furthermore, Gene Set Enrichment Analysis (GSEA) was used to interpret gene expression data to find gene sets sharing common biological function and regulation. Finally, we calculated gene evolutionary rate to explore the lineage distribution amongst all insulin signaling genes.

Keywords

Gene Expression, K-means clustering, Principal Component Analysis, Self-Organizing Maps, Type 2 Diabetes Mellitus

1. INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is a metabolic disorder characterized by abnormal hepatic glucose output, insulin resistance and impaired insulin production inducing elevated blood glucose levels. The chronic hyperglycemia is associated with persistent damage, dysfunction impelling the failure of various organs. The number of affected individuals is ever-increasing due to ageing, urbanization, prevalence of obesity attributed primarily due to a sedentary life style and genetic predisposition. The demographic studies estimate the projected rise from 171 million in 2000 to 366 million in 2030. Dismally, the total number of diabetic people in developing countries is expected to double between 2000 and 2030 [1]. Diabetes is caused by multiple genetic and environmental factors inflicting it to be complex in nature. Recent studies of gene expression levels in humans have indicated that, in T2DM many metabolic pathways are affected owing to the repercussion of metabolic dysfunction [2]. This multi-factorial nature of T2DM indicates that it can be considered as a biological systems disease [3]. There are many pathways which are associated with insulin signaling which include:

[1] MAPK signaling pathway,

[2] phosphatidyl inositol signaling system,

[3] Starch and sucrose metabolism,

[4] Apoptosis,

[5] glycolysis/gluconeogenesis,

[6] Fatty acid biosynthesis.

These pathways are extensively connected and embedded as network, so there is a need to study these insulin signaling pathways from a systems biology perspective [4]. Peculiarly, a system-level approach explicitly studies interactions of systemic components and their resultant effect on function and behavior of that system. Undeniably, the traditional molecular biology approach may not lead to satisfactory understanding of underlying molecular mechanisms vital for curing the disease. Thus we have applied a systems biology approach involving the integration of co-expressed gene network with corresponding protein-protein interaction network to identify signature network. In the present study, we have worked with microarray data from cultured abdominal subcutaneous preadipocytes obtained from lean and obese Pima Indians. However, in spite of having insignificant P- value, these two dataset are easily distinguishable by their gene expression value in Gene clustering as well as in Self organized maps (SOM). Provided the created network in SOM learns from their classification of the training data without any supervision, thus it is a case of unsupervised learning. At present, available human tissue transcriptome data related to T2DM provides an opportunity for identification of novel molecular mechanisms underlying the metabolic phenotype. We have scrutinized these microarray gene expression datasets that identify transcriptionally altered key genes involved in insulin resistance in T2DM. In addition, study of protein-protein interactions network gives insights of protein complexes and their molecular organization in the overall topology of the networks [5]. We began our investigation by focusing on insulin-signaling genes as it is well established method for finding a signature set of genes associated with insulin resistance in diabetes. Genes of insulin signaling network are highly interconnected and tightly regulated, any perturbations at a single node (gene) can thus rapidly diffuse to the rest of

the network (gene network) [5]. Thus, rather than looking for significant aberrations in expression of individual insulin-signaling genes, we focused at the collective expression of a set of insulin-signaling genes whose protein products form a connected protein–protein interaction network. Through this approach we tried to assess the interactions of known T2DM candidate genes with other molecules in different biological pathways and a few unique interactions which are statistically significant. By examining the microarray data, we attempted to detect statistically significant transcriptional alterations in a set of insulin signaling genes in diabetic tissue compared to normal one. We analyzed datasets of tissue-specific and microarray expression profiles of total RNA transcript of obese and non-obese diabetic individuals. These profiles were derived from the type 2 Diabetes database. Cluster analysis methods have been used to group systematically, related gene expression patterns into closely related expression profiles. The created clusters were used to group the genes according to their expression data in Type 2 diabetes under Obese and non-obese condition. We found sub network of highly interconnected insulin signaling genes in our clustered network. In protein interaction network we also identified those genes which perform important function in insulin sensitivity like glucose and glycogen metabolic processes and signaling pathways collectively. Among all other clustering techniques, the peculiarity of Self Organizing Maps (SOM) is that it allows easy visualization of complex data. Moreover it is robust to minor experimental variation. This technique clusters points in multidimensional space. SOM algorithm can be implemented on gene expression data by considering the quantitative expression levels of n genes in k samples as defining n points in k-dimensional space [6]. In order to group together genes with similar expression patterns a simple technique like direct visual inspection can be employed [7]. In SOM, neighboring genes tend to define related clusters. We have also utilized a network-centered methodology to link diabetes-related alterations in gene expression in insulin signaling genes and transcription factors potentially responsible for gene expression changes. These gene-sets share one or more transcription factors binding sites in the promoter regions of the corresponding genes. In order to identify such potential regulatory players, promoter sequences of the genes associated with the reporter metabolites were tested for enrichment of known transcription factor binding motifs and these identified transcription factors provide clues to the regulatory mechanisms that lead to observed gene expression changes in the gene network. We also used Gene Set Enrichment Analysis (GSEA) to evaluate microarray data at the level of gene sets. The gene sets are defined based on prior biological knowledge, e.g., published information about biochemical pathways or co-expression in previous experiments [8]. The goal of GSEA is to determine whether members of a gene set S tend to occur toward the top (or bottom) of the list L, in which case the gene set is correlated with the phenotypic class distinction (Here, we used microarray data condition as phenotypic distinction). Grouping of these datasets can reveal which of the gene-sets correspond to the same biological processes and which represent distinct processes. A comprehensive list of diabetes-related targets (genes/proteins) was compiled from the biomedical literature through text mining technology. SciMiner, a web-based literature mining tool, was used to retrieve and process documents and identify targets from the text [9]. Functional enrichment analyses were performed on these targets to identify significantly over-represented biological functions in terms of Gene Ontology (GO) terms and pathways [10]. In a separate line of analysis, it has been shown that the length of the protein encoded by a gene is significantly and negatively correlated with the evolution rate that is, longer proteins, on average, are more highly conserved in evolution than short ones [11]. There is a negative correlation existing between gene-expression level and sequence evolution rate and it has been proven that highly expressed genes evolve significantly slower than lowly expressed genes [12]. Gene evolution can be characterized not only by the sequence evolution rate, but also by the propensity of a gene to be lost or gain during evolution [13].

2. MATERIALS AND METHODS

Datasets are retrieved from the GEO (Gene Expression Omnibus) database available at http://www.ncbi.nlm.nih.gov/geo/. 28 data sets of 7/7 male/female sample of total RNA transcript are taken. This is divided into two categories, lean and obese condition 14 datasets in each. These samples are the microarray analysis of gene expression profiles of primary cultured preadipocytes from non-diabetic lean and non-diabetic obese Pima Indian subjects. The prevalence of obesity in Pima Indians is among the highest of any human population [14]. For this study, only insulin signaling genes were selected.

These microarray datasets are analyzed using Gene Cluster. In this study, initially hierarchical clustering is performed to calculate the distance matrix between the gene expression data. This study is done to identify highly expressed genes and also to establish their relation with the neighboring genes as depicted in the trees where the two closely occurring items (those with the smallest distance) are joined by a node/branch of a tree. Basically, single linkage clustering was done in which the distance between two gene is minimum of all pair wise distances, in correlation with complete linkage. Where the distances between the two genes is maximum .In complete linkage data were not much distinguishable. By default, Gene Weight of 1 is considered. This method assembles a set of genes into a tree, where entities are joined by very short branches if they are very similar to each other. As the similarity decreases, the branch length increases. Here, hierarchical clustering of 36(n) genes results in 35(n-1) nested clusters. Java Tree View is used to analyze hierarchically clustered genes. These clustered genes are then grouped for better understanding, using Self Organizing Maps (SOM) in MATLAB. Gene expression profile data of insulin signaling genes are subjected to the SOM algorithm. For this, data are trained in batch wise fashion and logarithmically normalized. Each neuron of the SOM was randomly assigned a weight vector with variables corresponding to an expression profile of random values (Fig. 3). During the learning phase, the expression profile of insulin signaling genes are repeatedly compared with the random profiles in neurons. Best Matching Unit (BMU) was found by following method:

$$\| x - mc \| = \min_i \{\| x - mi \|\} \quad (1)$$

Here,

'|| ||' is the distance measure, usually Euclidean distance; and BMU is here denoted by 'c' as shown in Fig. 2.

SOM updation rule for the weight vector of unit 'i' is:

$$Mi(t+1) = mi(t) + \alpha(t)\, hci(t)\, [x(t) - mi(t)] \quad (2)$$

Here, x (t) = an input vector randomly drawn from the input data set at time t, hci (t) = neighborhood kernel around the winner unit c, and α (t) = learning rate at time t.

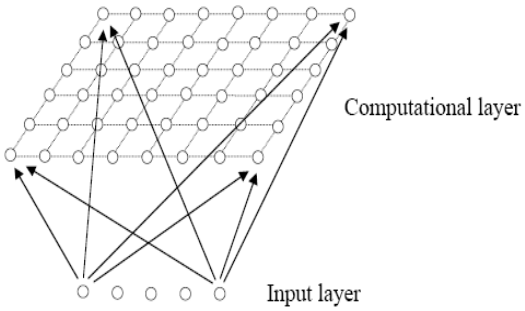

Fig 1: Principle of SOMs: Learning process of Self-Organizing Maps

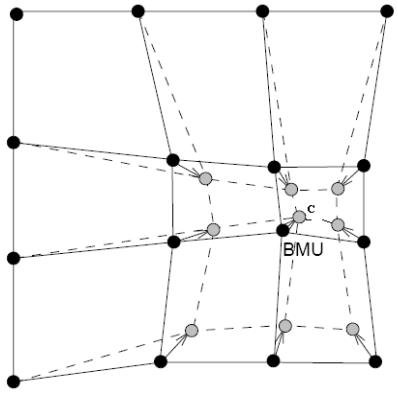

Fig 2: Principle of SOMs: Updation of the Best Matching Unit (BMU) and its neighbors towards the input sample method (marked X). Solid and dashed lines correspond to the situation before and after the updating process, respectively

Next, the Promoter regions, upstream of transcription start site (TSS), of each of the selected genes are analyzed for the enrichment of known transcription factor (TF) binding sequence motifs and thus, TFs were ranked according to their association with the gene sets. Top 10 TFs are selected for each gene. Out of which, common TFs are selected and counted for each gene and represented graphically based on their interaction with insulin signaling genes using Cytoscape [15].

Further, Functional enrichment analysis of each gene is done using GSEA (Gene Set Enrichment Analysis). The microarray expression profiles of insulin signaling genes (grouped in to two classes lean and obese, labeled A and B), are taken as input data. These genes are ranked based on the correlation of their expression values with a given target pattern (Expression pattern of a particular gene). GSEA calculates an Enrichment score (ES) that reflects the degree to which a gene set is over-represented at the extremes (top or bottom) of the entire ranked list. The magnitude of the increment of the ES depends on the correlation of the gene with the phenotype (Expression value in particular condition). In order to identify functional enrichment of the networks, BiNGO (a cytoscape plugin) was used. It is based on a hyper geometric test with Benjamini and Hochberg false discovery rate (FDR) corrected p-value; and displays the over-represented functional gene ontology categories in a gene network [16]. This finds over-representation of statistically significant Gene Ontology (GO) categories in gene sets (Fig. 3).

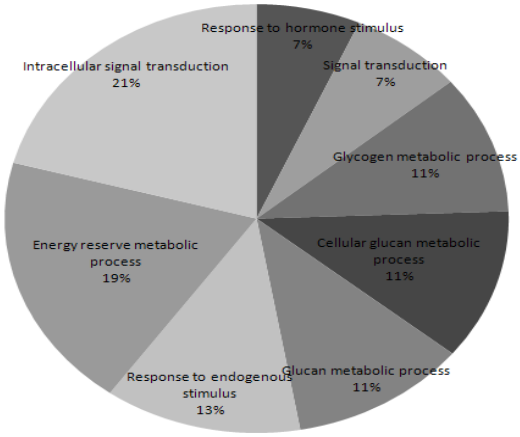

Fig 3: Pie-representation of the functionally enriched GO categories in GSEA identified gene sets

Next, sequences of gene product of insulin signaling genes are retrieved from the NCBI protein database. Each protein sequence is subjected to a pBlast search against the NCBI RefSeq protein sequence database in order to find the homologous sequences in Homo sapiens (Paralogous). Putative paralogous protein sequences are identified from the Blast hits with cutoff value of < 0.1. These paralogous protein sequences are aligned using the CLUSTAL program in MEGA 5.0.

P-distance (Amino acid) for each protein is calculated [17]. This distance is the proportion (p) of amino acid sites at which the two sequences to be compared are different. It is obtained by dividing the number of amino acid differences by the total number of sites compared. Finally, overall mean distance was calculated by averaging these values.

$$p = nd/L \qquad (3)$$

Here, nd = number of amino acids that are different between two aligned sequences, and L=number of sites compared. And finally, maximum likelihood estimates of the amino acid distances between the aligned paralogous sequences are calculated using the JTT (Jones-Taylor-Thornton) evolutionary mode [18].

3. RESULTS AND ANALYSIS

In this study, we have performed cluster based analysis of affected insulin signaling genes. Significance of differential expression for each gene is used as a scoring matrix. The results are summarized as genetic signatures (affected genes) and regulatory signatures (transcription factors) for T2DM [19]. We have performed a t-test for lean and obese condition data in MATLAB and it is found that in most of the condition p-values are > 0.5, suggesting that there is no significant difference in data. Theoretical quartile values are also along the cut-set (Sample Quartile value) and the plot is close to linearity, depicting that in both the conditions data are quite

similar. This is also supported by the K-mean clustering for microarray expression profile data of insulin signaling genes in different tissue samples for lean and obese conditions, as their expressions are quite similar.

A number of genes are observed for differentially expression in the disease state. We looked for significant aberrations in the collective expression of a set of insulin signaling genes. In tissue specific and microarray expression profiling datasets, AKT2, FRAP1, GRB2, PDK4, and RPS6 genes were observed to be distantly related whereas SLC2A4, INS, PPP1R3D, INPPL1, G6PC2, CBLB, SOCS1, GYS1, IGF2BP2, IRS2, PPP1R3C, MAPK10, PTPN11 and PRKAB1 genes were similarly expressed. These datasets exhibits differential expression of AKT2 which is capable of phosphorylating several other known genes and a tyrosine phosphatase PTPN11, that regulate a variety of cellular processes including cell growth, differentiation, mitotic cycle and oncogenic transformation; thus has been reported to be highly associated with T2DM. These results also provide insight into the role of pre-adipocytes in obesity and obesity-related inflammation [20, 21]. The alterations in insulin signaling were identified across adipose, liver, skeletal muscle and pancreas tissues. This information can be used to differentiate highly expressed genes in insulin signaling pathway by analyzing tissue specific expression profiling datasets. The results for the insulin signaling sets suggest that different members of these sets were transcriptionally altered under Type 2 Diabetes. These small sets of genes may independently lead to the insulin resistance phenotype and also consistent with other metabolic disorders [5]. In our results, it is observed that SOM rapidly and reliably clusters this gene expression data into groups that not only show similar gene expression profiles but also contain functionally related genes. SOM is further modified using Sammon's mapping algorithm, where spatial distance correlates with the difference in average expression profile and the circle size with the number of genes within the neuron. SOM results group the data into 9 clusters. Similar genes are grouped in the same cluster based on having similar expression values as shown in Fig. 4 and 5. The most frequent Transcription factors (TFs) associated with insulin signaling genes are identified best on their p-value and association score with their target genes (Table 1).

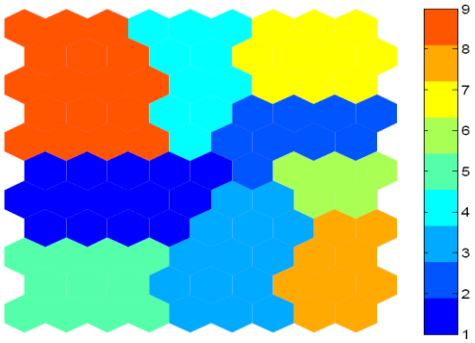

Fig 4: Clustering of lean and obese data condition together in SOMs

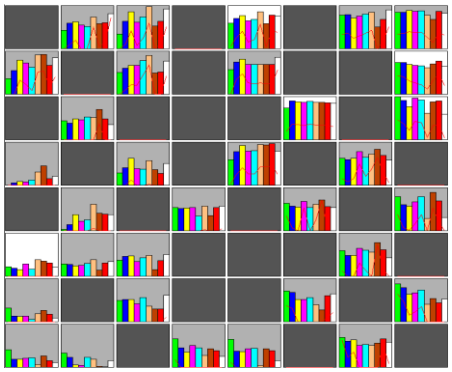

Fig 5: SOM analysis of 28 microarray data sets: Visualization of Self-Organizing Maps organized genes; Grayscale according to clusters; Intensity of grey background shows no. of genes present within each cluster; Line diagram shows average values between each nodes of a cluster; Genes with a similar expression profiles are clustered in the same neuron and genes with closely related profiles are in neighboring neurons.

TABLE 1: IDENTIFIED TRANSCRIPTION FACTORS ASSOCIATED WITH INSULIN SIGNALLING GENES

Transcription Factors	Association Score	p-value
Ap-2gamma	4.953	2.28e-04
Ap-2alpha	4.725	3.36e-04
Sp1	4.013	1.59e-03
Arnt	3.888	2.04e-03
E2f-1	3.62	3.53e-03
Hic-1	3.309	6.20e-03
Cp2	3.123	9.52e-03
Gata-1	3.123	9.52e-03
Stra13	3.049	1.08e-02
E47	3.005	1.23e-02

These TFs were mainly supposed to target PPP1R3C, PDK4, TRIP10, CAPN10, and MAPK1. Highly associated genes (in insulin signaling pathway from Nature pathway database) are analyzed in cytoscape with MCODE program which identified highly interconnected region in a clustered network with a high score of 3.538. These genes are RXRA (seed gene), PPARG, MED1, EP300, MAPK8, PPARGC1A, NR0B2, SIRT1, NCOR1, NCOR2 and FOXO1 as shown in Fig. 6.

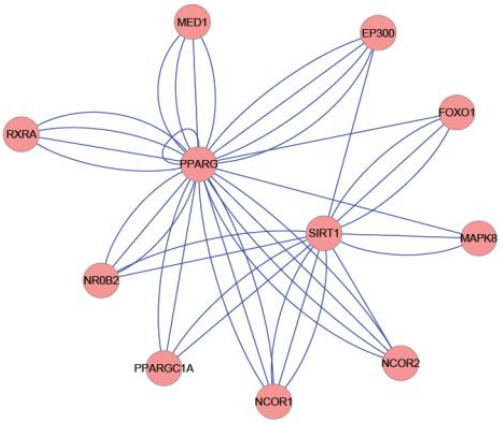

Fig 6: Highly clustered genes in Insulin Signaling set (found with MCODE)

The significance of network topology is that it enables a node to communicate with other genes in the network. It considers varied parameters such as average clustering, average degree, and betweenness centrality [22]. These low degree nodes (genes) are supposed to form highly cohesive network in insulin signaling network. In present analysis, the betweenness Centrality appears to be less as compared to the average node degree. This network follows the power law distribution with highest R-squared value of 0.962. As R-squared value approach unity, it implies that the regression approaches a power fit (Fig. 7).

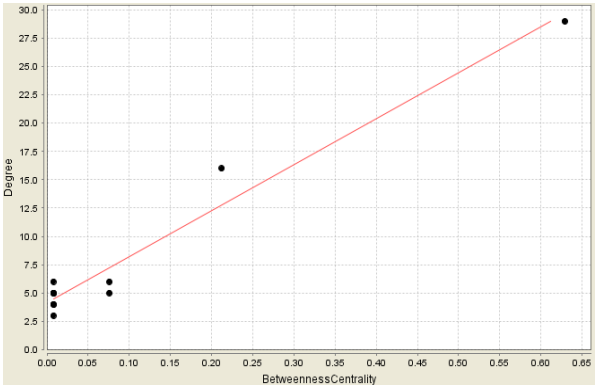

Fig 7: Network analysis: The distribution of Betweeness centrality is plotted on X-axis and the Degree is plotted on Y-axis. (R – Squared value – 0.962 and Correlation value – 0.981)

This suggests an overall scale free network, as network connectivity in these datasets closely resembles a scale free topology. All the information on network topology provides high confidence to their scale-freeness property supporting the fact that the underlying model is linear. Therefore, it can be suggested that the networks analyzed here are sensitive to the perturbation of more highly connected hubs rather than removal of less connected nodes. BiNGO results finds highly annotated GO categories in the insulin signaling protein-protein interaction network derived from the Nature pathway database. Our results find GO categories of those genes which perform important function in insulin sensitivity like glucose and glycogen metabolic processes and signaling pathway collectively. Most of the GO categories are involved in signaling pathways mediated by kinases, signal transducers, transcription factors, and nuclear receptors. Therefore, it can be assumed that the networks represent the clustering of signaling components, which upon alteration may cause T2DM. Previous works have shown that genes having low evolution rates are supposed to be younger in comparison to older genes having high evolution rates [23]. The difference in expression between insulin signaling genes is independent of the difference in length and sequence evolution rate (Table 2).

TABL 2: EVOLUTIONARY RATE (LOGL) OF INSULIN SIGNALING GENES

Gene rank	Gene list	Over-all mean distance	log-l value
1	RPS6KB2	1.358	-96969.76
2	IKBKB	1.316	-92460.29
3	MAP2K1	1.271	-78595.19
4	MAPK1	1.216	-60081.89
5	TRIP10	2.243	-51995.45
6	SOCS1	2.279	-50666.31
7	PTPN11	2.390	-41370.81
8	SLC2A4	1.692	-35721.34
9	CBLB	2.308	-34773.60
10	CAPN10	1.608	-33823.95
11	G6PC2	2.339	-30543.21
12	IRS1	1.956	-23956.44
13	GYS1	1.341	-18670.46
14	PPP1R3D	2.124	-16534.86
15	IDE	1.963	-15818.74

4. DISCUSSIONS AND CONCLUSIONS

Majorities of genes in insulin-signaling belong to the other biological processes like signal transduction, phosphorylation and positive regulation of transcription. In SOMs, expression profiles of a group of genes are represented by a common weight vector. Examination of the gene contents of individual neurons demonstrates that in many cases, the clustering achieved by SOM reliably predicts functional similarity and these similarities reflect participation of common regulatory elements. Applying SOM to these different datasets may reveal conditions under which the particular gene is regulated. As clustering of regulated genes appear to predict functional similarity, comparison with other genes in the same neurons may give valuable hints about the possible functional role of the unknown gene. Our results demonstrate that SOM is a fast and convenient method to organize gene expression data.

Expression profiles of a group of genes are represented by a common weight vector and the data are easily visualized as a two-dimensional matrix. Application of Sammon's mapping, transforms the data into a format where the relationships between individual neuron is even more clearly visualized. Finally, SOM provide easy visualization and interpretation of patterns in expression of gene as compared to other clustering techniques like K-means clustering and Principal Component Analysis.

The distinctions between the gene ages fit the concept of a gene's status. The old and new genes are supposed to be high and low status genes respectively. A high status of gene entails large protein size; evolutionary conservation including both the low propensity of the gene to be lost and slow sequence evolution, high expression level and numerous physico- genetic interactions. The paralogous gene distances within human data do not give very much insight into the gene age in comparison to orthologous genes in different organisms though it gives a clue to their development and their evolution. PPARG emerged as one of the possible candidate gene which encodes targets for anti-diabetes medications. A usual assumption in a gene-expression analysis is that sets of genes are partially co-expressed due to transcriptional co-regulation. Also, gene expressed in a certain category such as different cell types or stress conditions are expected to share binding signals for the TFs.

Conclusively, our results suggest a framework of metabolic disruption observed with insulin resistance which is used to test the role of insulin signaling pathway in mediating T2DM pathophysiology and for the potential biomarkers for the diabetes monitoring. We identified some candidate genes which play significant roles in insulin signaling pathway. Gene age calculation and their all possible interconnections in a network with their functional annotations provide a method to find dependency and development of concerning genes in insulin signaling gene networks. Interacting pairs within a gene network also suggest functional dependency. Our results suggest an alternative approach to find a functional role for an unknown gene entails large protein size; evolutionary conservation including both the low propensity of the gene to be lost and slow sequence evolution, high expression level and numerous physico- genetic interactions. The paralogous gene distances within human data do not give very much insight into the gene age in comparison to orthologous genes in different organisms though it gives a clue to their development and their evolution. PPARG emerged as one of the possible candidate gene which encodes targets for anti-diabetes medications. A usual assumption in a gene-expression analysis is that sets of genes are partially co-expressed due to transcriptional co-regulation. Also, gene expressed in a certain category such as different cell types or stress conditions are expected to share binding signals for the TFs.

5. ACKNOWLEDGMENT

We express our sincere gratitude to Mr. Somnath Tagore (Assistant Professor at D.Y Patil University) for providing us an opportunity to work on this project. We are grateful for his persistent support, guidance and motivation throughout the duration of entire project. We express our sincere thanks to our class mate Sneha Verma for her timely feedback. Lastly, we thank our family members, teachers and friends for their constant support in all our endeavors.

6. REFERENCES

[1] A. Salehzadeh-Yazdi, S. Akbari-Birgani, and A. Masoudi-Nejad, "Diabetes and systems biology", J. Iran. Chem. Soc., vol. 8, suppl. 3, pp. A63-A67, 2011.

[2] E. Loghmani, Chapter 14, Diabetes mellitus: type 1 and type 2, Guidelines for adolescent nutrition services 2005.

[3] K.R. Patil, J. Nielsen, "Uncovering transcriptional regulation of metabolism by using metabolic network topology", PNAS, vol. 102, no. 8, pp. 2685–2689, 2005.

[4] F. Folli, T. Okada, C. Perego, J. Gunton, C.W. Liew et al., "Altered insulin receptor signaling and β-cell cycle dynamics in Type 2 Diabetes mellitus", PLoS ONE, vol. 6, issue 11, pp. e28050, 2011.

[5] M. Liu et al., "Network-based analysis of affected biological processes in type 2 diabetes models", PLoS Genet, vol. 3, issue 6, pp. e96, 2007.

[6] P. Toronen, M. Kolehmainen, G. Wong, and E. Castren, "Analysis of gene expression data using self-organizing maps", FEBS Letters, vol. 451, pp. 142-146, 1999.

[7] P. Tamayo et. al., "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation: Proc. Natl. Acad. Sci. USA, Genetics, vol. 96, pp. 2907–2912, 1999.

[8] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, and L.B. Ebert, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles", PNAS, vol. 102, no. 43, pp. 15545–15550, 2005.

[9] J. Hur, A.D. Schuyler, D.J. States, and E.L. Feldman, "SciMiner: a web-based literature mining tool for target identification and functional enrichment analysis", Bioinformatics, vol. 25, issue 6, pp. 838-840, 2009.

[10] M. Ashburner et al, "Gene ontology: tool for the unification of biology, The Gene Ontology Consortium. Nat Genet", vol. 25, issue 1, pp. 25-29, 2000.

[11] D.J. Lipman, A. Souvorov, E.V. Koonin, A.R. Panchenko, and T.A. Tatusova, "The relationship of protein conservation and sequence length", BMC Evol Biol, vol. 2, pp. 20, 2002.

[12] D.M. Krylov, Y.I. Wolf, I.B. Rogozin, and E.V. Koonin, "Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution", Genome Res, vol. 13, pp. 2229–2235, 2003.

[13] Y.I. Wolf, P.S. Novichkov, G.P. Karev, E.V. Koonin, David J. Lipman, "The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages", PNAS, vol. 106, no. 18, 7273–7280, 2009

[14] S. Nair et al., "Increased expression of inflammation-related genes in cultured preadipocytes/stromal vascular cells from obese compared with non-obese Pima Indians", Diabetologia, vol. 48, issue 9, pp.1784-8, 2005.

[15] P. Shannon et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks", Genome Research, vol. 13, issue 11, pp. 2498-504, 2003.

[16] Y. Benjamini, and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing", Journal of the Royal Statistical Society Series B (Methodological), vol. 57, issue 1, pp. 289-300, 1995.

[17] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar, "MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods: Molecular Biology and Evolution (submitted), 2011.

[18] D.D. Sears, G. Hsiao, A. Hsiao, J.G. Yu, C.H. Courtney, J.M. Ofrecio, J. Chapman, "Mechanisms of human insulin resistance and thiazolidinedione-mediated insulin sensitization", PNAS, vol. 106, no. 44, pp. 18745–18750, 2009.

[19] H.G. Roider, A. Kanhare, T. Manke, Martin Vingron et al, "Predicting transcription factor affinities to DNA from a biophysical Model", Bioinformatics, vol. 23, issue 2, pp. 134-141, 2007.

[20] M.V. DiLeo, G.D. Strahan, M.D. Bakker, and O.A. Hoekenga, "Weighted Correlation Network Analysis (WGCNA) applied to the Tomato Fruit Metabolome", PLoS ONE, vol. 6, issue 10, pp. e26683, 2011.

[21] A. Guilherme, J.V. Virbasius, V. Puri, M.P. Czech, "Adipocyte dysfunctions linking obesity to insulin resistance and type 2 diabetes", Nat Rev Mol Cell Biol, vol. 9, issue 5, pp. 367–377, 2008.

[22] U. Sengupta, S. Ukil, N. Dimitrova, and S. Agrawal, "Expression-based network biology identifies alteration in key regulatory pathways of type 2 diabetes and associated risk/complications", PLoS ONE, vol. 4, issue 12, pp. e8100, 2009.

[23] E.V. Koonin, Y.I. Wolf, "Evolutionary systems biology: Links between gene evolution and function", Curr Opin Biotechnol, vol. 17, pp. 481–487, 2006