

## STRUCTURE PREDICTION TO ANALYSE GENETIC ALTERATION IMPLEMENTING Q MEAN VARIATION ANALYSIS AND PHYLOGENETIC TREE CONSTRUCTION IN PREDIABETIC AND DIABETIC SUBJECT

<sup>1</sup>\*Talib Yusuf A., <sup>2</sup>Talib S. H. and <sup>1</sup>Sanjay Harke

<sup>1</sup>Dept. of Biotechnology, MGM Institute of Health Sciences, Aurangabad.

<sup>2</sup>Dept. of Medicine, MGM Medical College and hospital, Aurangabad.

\*Corresponding Author: Dr. Talib Yusuf A.

Dept. of Biotechnology, MGM Institute of Health Sciences, Aurangabad.

Article Received on 29/08/2019

Article Revised on 19/09/2019

Article Accepted on 09/10/2019

### ABSTRACT

Diabetes is metabolic disorder occurs when pancreas fail to produce enough insulin and when the body cannot use the insulin it produces. The classic sign and symptoms includes polyuria (frequent urination), they will become increasingly thirsty (polydipsia) and hungry (polyphagia), unexplained weight loss when appetite is good, or it may remain asymptomatic. It is classified into two basic form type I and type II diabetes. Complications of diabetes include neuropathy, nephropathy, and retinopathy may develop prediabetic and Diabetes stages of disease or many years after onset of diabetes mellitus. In type I diabetes there is islet cell destruction; absolute insulin deficiency, ketosis in absence of insulin, in type II diabetes having insulin resistance and relative insulin decrease, no HLA (human Leukocyte antigen) association. When hyperglycaemic when doesn't satisfy the diagnostic criteria for Diabetes mellitus is generally called prediabetic. Diabetes Mellitus is diagnosed when either HbA1c > 6.5, fasting glc > 126 mg/dl, or glc 2 hour after OGTT > 200 mg/dl or single random glc > 200 mg/dl. Nearly 45% of all diabetics have peripheral vascular disease may develop macro and / or micro vascular. Thus prediabetic and diabetes mellitus are said to be complex entities caused by complex interplay between genetic and epigenetic and environmental factors. The identification of genetic factor possible now with NGS (next generation sequencing) and GWAs (Genome Wide Associations). Broadly speaking genetic alteration comprises by SNP's, Pseudogenes inaccurate duplication and mutations. In this study an attempt is made to evaluate the Quality Score in the samples of prediabetic (PDS), Diabetic (DS) compared with Healthy (HS) utilising NCBI BLAST (Basic Local Alignment Search Tool), PDB (Protein Databank), ExpASy (Expert Protein Assessment/Analysis System) and metabolic pathway analysis by KEGG and KAAS Server.

**KEYWORDS:** - Diabetes Mellitus, Type I and Type II, Prediabetic, genetic alteration, similarity, Q-Mean (Z score). Glycolysis Pathway.

### INTRODUCTION

#### Diabetes

Diabetes is devastating disease that is characterized by high glucose levels in the blood and has been recorded in the medical literature since as early as 1500 BC.<sup>[2]</sup> Diabetics either do not produce enough insulin to process their intake of glucose or the body does not use the insulin efficiently enough to control glucose levels.<sup>[2]</sup> Untreated, diabetes can cause a number of health problems including, blindness, loss of circulation resulting in limb amputation, high blood pressure, heart disease, and kidney failure.<sup>[2]</sup>

#### Alignment

Best method to analyze the sequence is similarity searching which gives the values to correlate with other fragments, organisms, species, phylum and kingdom,

here The BLAST program (Stephen Altschul) was utilized for analysis, these program works on Needleman Wunsch Algorithm, where sequences are broken into fragments, then Scoring matrices (BLOSUM 62) were generated, which led to log odd ratio, and ended in shortlisting the sequence similarity by KA equation (E-Value) is produced.

The target is to find high-scoring ungapped segments among related sequences. The existence of such segments above a given threshold indicates pair wise similarity beyond random chance, which helps to discriminate related sequences from unrelated sequences in a database.<sup>[1]</sup>

### Structure Prediction

Protein Structure Prediction is the process of prediction of the three dimensional structure of a protein from its amino acid sequence. It is the inference of the three-dimensional structure of a protein from its amino acid sequence—that is, the prediction of its folding and its secondary and tertiary structure from its primary structure. Structure prediction is fundamentally different from the inverse problem of protein design. Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). The QMEAN Z-score in the provides an estimate of the "degree of nativeness" of the structural features observed in the model on a global scale. QMEAN Z-scores around zero (0) indicate good agreement between the model structure and experimental structures of similar size. Scores of -4.0 or below are an indication of models with low quality [Graph: 1].

### Phylogenetic tree

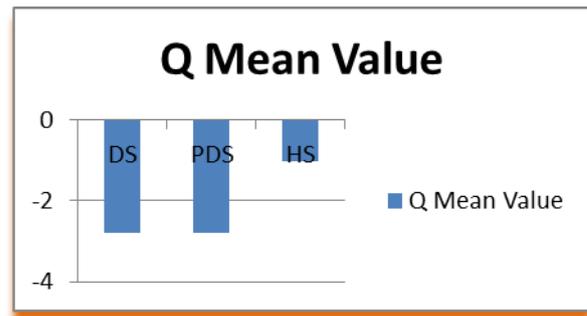
Phylogeny is classified as relationship between two species. The resulting relationship is represented as binary tree. Two main types of trees are: (i) Rooted trees—where all nodes are derived from single node. And (ii) Unrooted trees—where it is not clear that where the nodes originated from.

Our goal in this paper is to implement the computational biology, Bioinformatics to analyse, correlate the disturbances caused by various outcomes like SNP (single nucleotide polymorphism), Pseudogene, EST (expressed Sequence Tag), etc. and to highlight their impact in molecular Biology and its functioning.

### 2. Programs

Different software's and methods are available by which information of common ancestor can be drawn for that firstly target sequence is obtained in specific fasta file form biological database (NCBI), obtained sequences are processed by Different software's. These sequence has processed for sequence alignment, similar sequence has been shortlisted on the basis of Expectation Value (E-Value) (Table 1), then those shortlisted sequences are performed by using FASTA33, SWISS-PROT (ExPASy) for to predict the structure of protein sequence for homology modeling using SWISS-MODEL server.

For phylogenetic tree construction those shortlisted sequence has been performed by using CLUSTALX drawn in the form of fasta using ClustalX (Provides output in .dnd, .phy) and Phylip.



Graph: 1

Q mean is composite estimator based on different geometrical properties and provides both global and local absolute quality estimates on the basis of one single model. The Q-Mean, Z score provides an estimates of the "degree of nativeness" of the structural features observed in the sample. Q-mean Z score around Zero (0) indicates good agreement between the model structure and experimental structures of similar size. Score of -4.0 or below are an indication of models with low quality.

### 3. Phylogenetic Tree Construction

Once the sequence alignment is completed the samples were processed for the Phylogenetic tree construction, for which here we used Clustal X and PHYLIP program to get the pedigree analysis.

After completion of again offline sequence similarity search using Clustal X, three output files are generated named. dnd., aln., phy, that will help to construct phylogenetic tree using Maximum Likelihood [In fig. 2] and Maximum Parsimony method [In fig. 3], in rooted and unrooted form[In fig. 1]. The sampled output shows the correlation based on similarity and dissimilarity between three samples [HS, DS, PDS]. Those outputs show the different branch length from obtained sequence gives the evolutionary divergence.

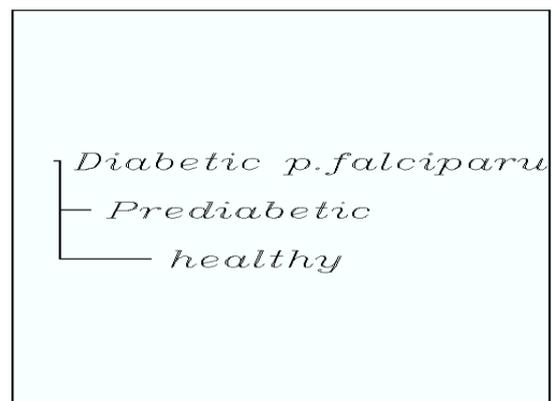


Fig. 1: Rooted Phylogenetic tree.

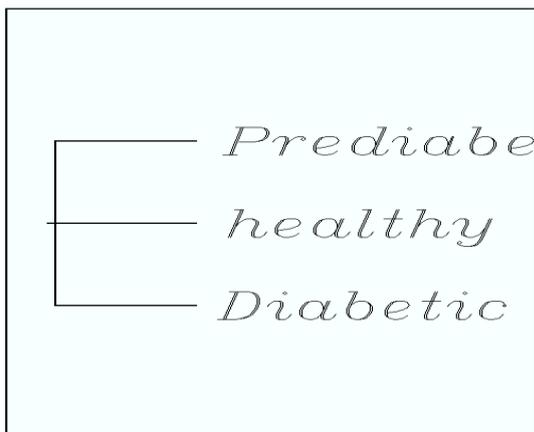


Fig. 2: Rooted Phylogenetic tree by using Maximum Likelihood.

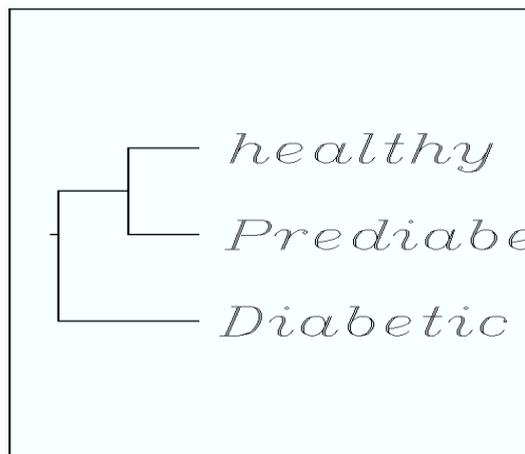


Fig. 3: Rooted Phylogenetic tree by using Maximum Parsimony.

4. Structure Prediction

Predicting the secondary structure is one of the important aspect in Biological world, that will help to understand the availability of atoms in the molecule, energy level, stability of the molecule, and mobility. In this work, Based on similarity search and expected E-values, some sequences have been shortlisted, and ExPasy was utilized to predict the secondary structure, by which different chains of helix, sheets and thread can be seen. Q-Mean provided from the Swiss-Prot shows the accuracy and similarity between the said different sequences of samples [Table-1].

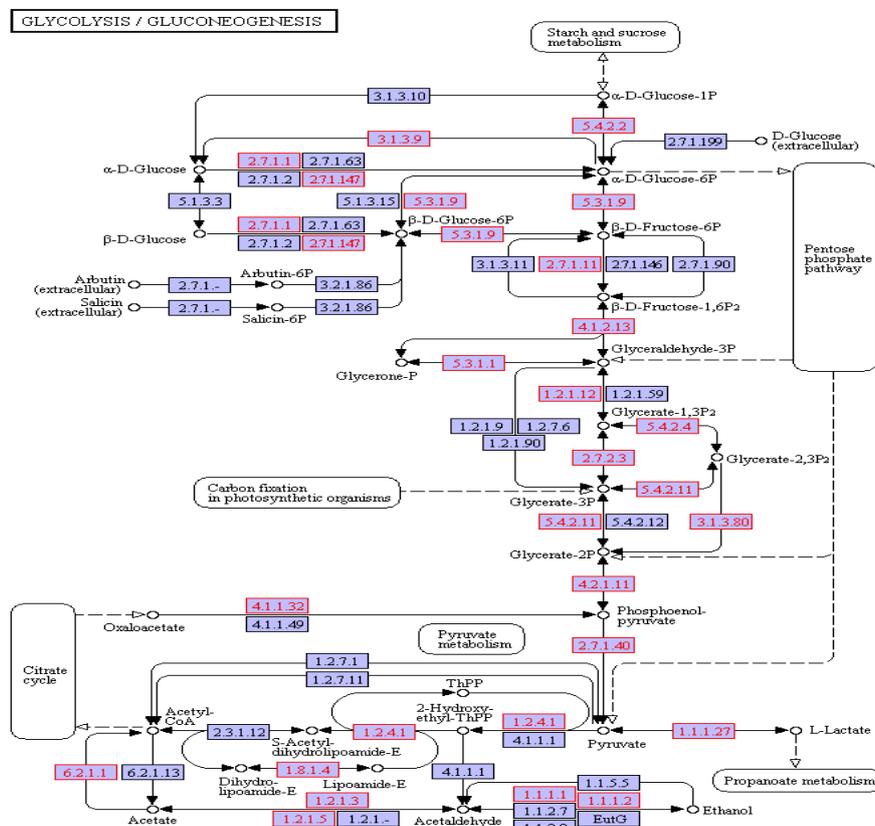


Figure 4: Glycolysis cycle of Healthy subject.

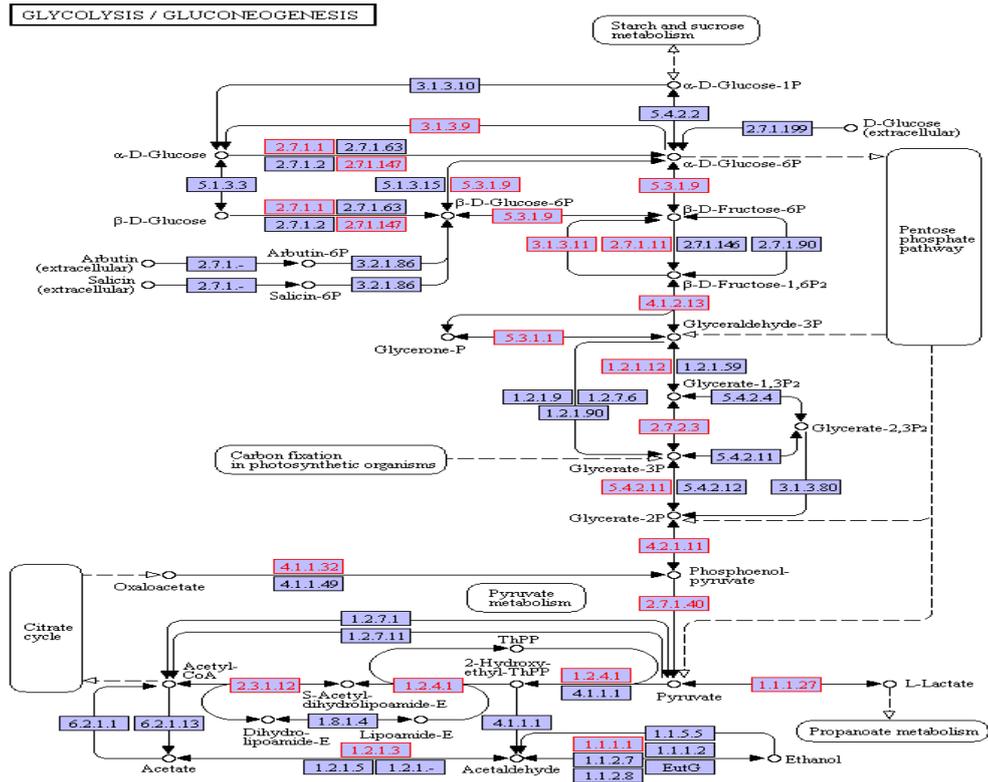


Figure 5: Glycolysis Cycle of Prediabetic Subject.

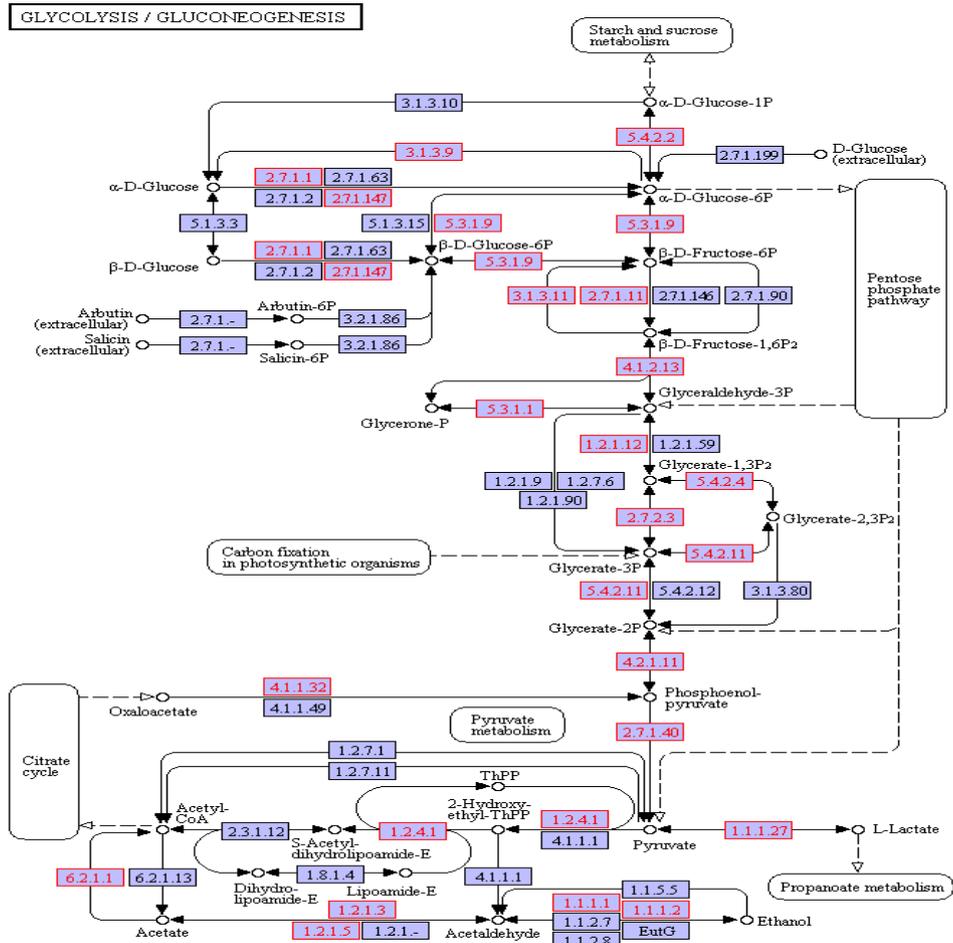
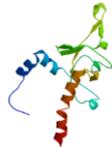
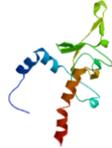


Figure 6: Glycolysis Cycle of diabetic Subject.

**Table 1: Showing the Q-Mean and Structure prediction of Healthy, Prediabetic and Diabetic Subject.**

Sr. No.	Description	E-Value	QMEAN	Protein Structure	Uniprot Accession
1.	Diabetic Patient	1.9E-30	-2.78		A0A024RBF6 (A0A024RBF6_HUMAN)
2.	Pre-Diabetic Patient	2.5E-29	-2.79		A0A024RBF6 (A0A024RBF6_HUMAN)
3.	Healthy Patient	8.2E-30	-1.02		A0A024RBF6 (A0A024RBF6_HUMAN)

**Table 2: Showing the list of EC (enzyme classification) number and class effecting in glycolysis pathway, and on/off in all 3 subjects and list of genes involved in it.**

Entry	Name	Class	Substrate	Product	Comment	Pathways	Genes
1.1.1.2	Alcohol dehydrogenase (NADP)	Oxidoreductase	Alcohol, NADP+	Aldehyde, NADPH	Off in PDS	Glycolysis, gluconeogenesis	HAS, PTS, PPr, GGO, PON, NLE, MCC, MCF, CASB, RRO
3.13.80	MIPP1	Hydrolases	2,3-bisphospho-D-Glycerate	2-phospho-D-glycerate	Off in PDS and DS	Glycolysis, gluconeogenesis	HAS, PTS, PPr, GGO, PON, NLE, MCC, MCF, CASB, RRO
5.4.2.11	Phosphoglycerate Mutase	Isomerases	2-phospho-D-Glycerate	3-phospho-D-Glycerate	Off in PDS	Glycolysis, gluconeogenesis	HAS, PTS, PPr, GGO, PON, NLE, MCC, MCF, CASB, RRO
2.3.1.12	Dihydrolipoyllysine-residue acetyltransferase	Transferase	Acetyl CoA	CoA	Off in HS and DS	Glycolysis, gluconeogenesis	HAS, PTS, PPr, GGO, PON, NLE, MCC, MCF, CASB, RRO
1.2.1.5	Aldehyde Dehydrogenase	Oxidoreductase	Aldehyde, NAD+, NADP+	Carboxylate, NADH and NADPH	Off in PDS	Glycolysis, gluconeogenesis	HAS, PTS, PPr, GGO, PON, NLE, MCC, MCF, CASB, RRO
5.4.2.2	Phosphoglucose mutase	Isomerase	Alpha-D-Glucose-Phosphate	D-Glucose 6-Phosphate	Off in PDS	Glycolysis, gluconeogenesis	HAS, PTS, PPr, GGO, PON, NLE, MCC, MCF, CASB, RRO
3.1.3.11	Fructose Diphosphate	Hydrolase	D-Fructose-1,6-Bisphosphate	D-Fructose 6 phosphate	Off in HS	Glycolysis, gluconeogenesis	HAS, PTS, PPr, GGO, PON, NLE, MCC, MCF, CASB, RRO

6.2.1.1	Acetate CoA ligase, synthetase	Ligases (forming C-S bond)	ATP, Acetate, CoA	AMP, Acetyl-CoA	Off in PDS	Glycolysis, gluconeogenesis	HAS, PTS, PPr, GGO, PON, NLE, MCC, MCF, CASB, RRO
1.8.1.4	Dihydropyridyl dehydrogenase	Oxidoreductase	Protein N6-(dihydropyridyl) lysine	Protein N6-lipoyllysine, NADH	Off in PDS and DS	Glycolysis, gluconeogenesis	HAS, PTS, PPr, GGO, PON, NLE, MCC, MCF, CASB, RRO

## RESULTS AND CONCLUSION

Genomic studies enlighten the effect used using different methods, algorithm, and programs in bioinformatics led to analyse the sequence, structure and predicting them. In these study, implementing Bioinformatics and their methods, various tools led to conclude the similarity, structure robustness and there activity in causing disease of Diabetes Mellitus Type II [Table 1]. Q-Mean shows the quality score of the said structures from DS to HS, ranging from -1.02 to -2.79, which concludes the quality of the structure between experimental Structure and target database structure, as nearer to Zero (0) (Graph:1) will have optimum and good quality score from the perspective of functionality, mobility and energy level of subjects taken. After performing the swiss model score generated led to conclude the average quality of the structure as -4.0 or below led to conclude low quality of the structure, which is of no use for further analysis.

Phylogenetic tree constructed in Pedigree analysis [fig. 1, fig. 2, and fig. 3] shows the implemented algorithm UPGMA, MP, ML prove the correlation between DS, HS and PDS.

Which will led to practitioners to make watch on Pre diabetic patients to control to convert into healthy and not moving towards diabetic one. One more aspect can be implemented in future about to correlated the method for functional finding in the sequences. Also to target the SNPs and Pseudogene involved in it. Here in this work target to highlight the genetic variation in metabolic pathway too, where fructose diphosphate found off (not expressed) EC- 3.1.1.11 in PDS and DS compared with healthy subject.

### Ethical Matters

Samples used in this work is approved by ethical committee, and informed consent are noted from the patients.

### REFERENCES

1. Talib Yusuf et al, RJLBPCS "Study of Genome Wide Association (GWAS) to analyze SNPs and Pseudogenes in Type 2Diabetes", 2018.
2. Robert Kelley<sup>1</sup> and Eric C. Rouchka<sup>1</sup> TR-ULBL- Bioinformatics Techniques Used in Diabetes Research, 2007.

3. H.-Y. LV, W.-G. Zhou, and C.-G. Zhou, "A Discrete Particle Swarm Optimization Algorithm for Phylogenetic Tree Reconstruction," 2004; 2650-2654.
4. J. Pevsner, "Molecular Phylogeny and Evolution," 1 ed John Wiley and Sons, Inc., 2003; 368-388.
5. Sneath PHA and Sokal RR, "Numerical Taxonomy," in *Numerical Taxonomy* San Francisco, CA: Freeman, 1973; 230-234.
6. N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Mol. Biol. Evol.*, July 1987; 4(4): 406-425.
7. W. M. Fitch and E. Margoliash, "Construction of phylogenetic trees," *Science*, Jan. 1967; 155: 760, 279-284.
8. W. Fitch, "Toward defining the course of evolution: minimum change for a specified tree topology.," *Syst. Zool.*, Jan.1971; 20: 406-416.
9. J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach," *J. Mol. Evol.*, 1981; 17(6): 368-376.
10. Uma Makhswari M and Sudarsanam D A review on Bioinformatics for diabetes mellitus, 2012.
11. Suparna Deepak, Rosalind Marita Genomics of Type 2 Diabetes, 2005.
12. H.-Y. LV, W.-G. Zhou, and C.-G. Zhou, "A Discrete Particle Swarm Optimization Algorithm for Phylogenetic Tree Reconstruction," 2004; 2650-2654.
13. J. Pevsner, "Molecular Phylogeny and Evolution," 1 ed John Wiley and Sons, Inc., 2003; 368-388.
14. Sneath PHA and Sokal RR, "Numerical Taxonomy," in *Numerical Taxonomy* San Francisco, CA: Freeman, 1973; 230-234.
15. N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees," *Mol. Biol. Evol.*, July 1987; 4(4): 406-425.
16. W. M. Fitch and E. Margoliash, "Construction of phylogenetic trees," *Science*, Jan. 1967; 155(760): 279-284.
17. W. Fitch, "Toward defining the course of evolution: minimum change for a specified tree topology.," *Syst. Zool.*, Jan. 1971; 20: 406-416.
18. J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach," *J. Mol. Evol.*, 1981; 17(6): 368-376.

19. Durbin R., S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Reprint ed. Cambridge: Cambridge University Press, 1999.
20. B. G. Hall, *Phylogenetic Trees Made Easy: A How-to Manual*, 3 ed Sinauer Associates, Inc., 2007.
  - a. N. Egan and K. A. Crandall, "Theory of Phylogenetic Estimation," in *Evolutionary Genetics: Concepts and Case Studies*, 1 ed Oxford University Press, USA, 2006; 426-436.
21. M. Dorigo and G. Di Caro, "The Ant Colony Optimization Meta-Heuristic," in *New Ideas in Optimization*. Corne D., M. Dorigo, and F. Glover, Eds. London: McGraw-Hill, 1999; 11-32.
22. L. Nakhleh, G. Jin, F. Zhao, and J. Mellor-Crummey, "Reconstructing phylogenetic networks using maximum parsimony," *Proc. IEEE Comput. Syst. Bioinform. Conf.*, 2005; 93-102.
23. M. Perretto and H. S. Lopes, "Reconstruction of phylogenetic trees using the ant colony optimization algorithm," *Genet. Mol. Res.*, 2005; 4(3): 581-589.
24. C. Kumnorkaew, K. Ku, and P. Ruenglerpanyakul, "Application of ant colony optimization to evolutionary tree construction," 2004.
25. S. Ando and H. Iba, "Ant algorithm for construction of evolutionary tree," Honolulu, HI, USA, 2007; 1552-1557.
26. M. Dorigo and T. Stutzle, *Ant Colony Optimization* Mit Press, 2004. J. Guo, L. Chen, L. Qin, and C. Wang, "A self-adaptive ant colony algorithm for phylogenetic tree construction," 01 ed 2006; 78.
27. J. Kennedy and R. C. Eberhart, "Particle swarm optimization,"
28. Piscataway, NJ, USA: IEEE Service Center, 1995; 1942 – 1948.