ISSN 22773061

Identification of Biomarkers for Obesity associated with Diabetes using Sequence Mining Techniques

Lalitha Saroja Thota, Allam Appa Rao
Research Scholar, Department of Computer Science, Acharya Nagarjuna Univeristy, Guntur, India
lalithasarojathota@gmail.com
Director, CRRao AIMSCS, University of Hyderabad campus, Hyderabad, India
apparaoallam@gmail.com

ABSTRACT

The advancements in the field of information technology are moving ahead in the discipline of medicine empowering the researchers with superior tools. By taking the advantage of Information Technology, today's researcher successfully navigate the flood of data and many diabetic complications can be overcome. Biomarker plays very major role in disease detection at early stages of its stages and also helpful in knowing the state of treatment and how body is acting or responding to the medication. The dramatic rise in obesity-associated diabetes resulted in an alarming increase in the incidence and prevalence of obesity an important complication of diabetes. The twin epidemic of diabetes and obesity pose daunting challenges worldwide. Differences among individuals in their susceptibility to both these conditions probably reflect their genetic constitutions. Predicting obesity associated diabetes is both useful and important because the number of obese patients is increasing while its main cause cannot yet be defined. Bioinformatics, a truly multidisciplinary science, aims to bring the benefits of computer technologies to bear in understanding the biology of life itself. The dramatic improvements in genomic and bioinformatic resources are accelerating the pace of gene discovery for many medical diseases. It is tempting to speculate the key susceptible genes/proteins biomarker that bridges diabetes mellitus and obesity. The emergence of post-genomic technologies has led to the development of strategies aimed at identifying specific and sensitive biomarkers from the thousands of molecules present in a tissue or biological fluid. In this regard, we evaluated the role of several genes/proteins that are believed to be involved in the evolution of obesity associated diabetes by employing a sequence mining technique, multiple sequence alignment using ClustalW tool and constructed a phylogram tree using functional protein sequences extracted from NCBI. Phylogram was constructed using Neighbor-Joining Algorithm a bioinformatic tool. Our bioinformatic analysis reports a biomarker, resistin gene as ominous link with obesity associated diabetes. This bioinformatic study will be useful for future studies towards therapeutic inventions of obesity associated type 2 diabetes.

Indexing terms/Keywords

Bioinformatics, biomarker, resistin, obesity, diabetes, data mining, sequence mining, multiple sequence alignment, ClustalW tool, phylogram, protein sequences, NCBI

Academic Discipline And Sub-Disciplines

Discipline : Computer Science Engineering, Sub-Discipline : Bioinformatics and Data Mining,

SUBJECT CLASSIFICATION

To find the role of several genes/proteins biomarkers that are believed to be involved in the evolution of obesity associated diabetes

TYPE (METHOD/APPROACH)

By employing a sequence mining technique, multiple sequence alignment using ClustalW tool and construct a phylogram tree using functional protein sequences involved in obesity associated diabetes extracted from NCBI

INTRODUCTION

In recent years, the escalating worldwide prevalence of obesity is considered as one of the most serious issues. This is because obesity is significantly associated with diabetes, heart disease, cancer, high blood pressure, and high cholesterol [1][2]. The medical complications of obesity are presented in Fig 1. Overweight and obesity are defined as abnormal or excessive fat accumulation that presents a risk to health. Obesity can be classified as one of the dangerous illness in the world as the number of the obese person keep on increasing.

Insulin is one of the most important hormones in the body. It aids the body in converting sugar, starches and other food items into the energy needed for daily life. However, if the body does not produce or properly use insulin, the redundant amount of sugar will be driven out by urination. This disease is referred to diabetes. The cause of diabetes is a mystery, although obesity and lack of exercise appear to possibly play significant roles

According to the WHO, 65% of the world's population lives in a country where overweight and obesity kills more people than underweight". Moreover it tells 44% of diabetes, 23% of ischaemic heart disease and more than 7% of certain cancers, globally, are attributable to obesity [3]. Diabetes mellitus has been on the rise across the world affecting over 150 million people. Over 20% of diabetics in the world are Indians. At present, it is higher in developed than in developing countries. The number of adults with diabetes in the world will rise from 135 million in 1995 to 300 million in the year 2025. The major part of this numerical increase will occur in developing countries. By the year 2025, greater than 75% of people with diabetes will reside in developing countries. The countries with the largest number of people with diabetes are, and will be in the year 2025, India, China and U.S [4].

It has been presumed from genetic studies that there could be subset of genes whose expression changes with obesity and those genes whose expression further changes in the progression to type 2 diabetes. However, the molecular basis that links obesity and diabetes is still largely unknown.

The ultimate goals for research focused on complex human diseases are to either prevent or to cure the diseases. These are ambitious goals that will be greatly facilitated by the identification of new biomarkers that can serve as novel diagnostic or prognostic indicators of disease course, that can be used as surrogate disease markers to track the efficacy of novel treatment strategies, or that may provide new targets for the treatment of the diseases.

For detecting a disease biomarker number of tests should be required from the patient. But using sequential data mining technique the number of test can be reduced. This reduced test plays an important role in time and performance. The emergence of post-genomic technologies has led to the development of strategies aimed at identifying specific and sensitive biomarkers from the thousands of molecules present in a tissue or biological fluid. Bioinformatics, a truly multidisciplinary science, aims to bring the benefits of computer technologies to bear in understanding the biology of life itself. The dramatic improvements in genomic and bioinformatic resources are accelerating the pace of gene discovery for many medical diseases like obesity and diabetes. It is tempting to speculate the key susceptible genes/proteins biomarker that bridges diabetes mellitus and obesity.

Bioinformatics has been in the focus since recent years for unraveling the structure and function of complex biological mechanisms. The analysis of primary gene products has further been considered as diagnostic and screening tool for disease recognition. Such strategies aim at investigating all gene products simultaneously in order to get a better overview about disease mechanisms and to find suitable therapeutic targets. This paper will therefore focus on potential implications of bioinformatics as a tool to identify novel metabolic patterns or biomarkers associated with obesity disease status. We will exemplify the potential of this method using the association between specific fats and development of obesity associated diabetes as a test case. In the present *in silico* study we have employed clustalW online bioinformatics tool for the analysis of seventeen genes, which are excepted to be play major role in obesity and diabetes, we sought to identify the common central gene/protein, a biomarker that connects both the metabolic disorders such as obesity and diabetes.

BACKGROUND

The use of the term "biomarker" has been dated back to as early as 1980. In 1998, the National Institutes of Health Biomarkers Definitions Working Group defined a biomarker as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention [5]. A Biomarker is a substance used as an indicator of a biologic state. The uses of biomarkers are shown in Fig 2. In genetics, a biomarker (identified as genetic marker) is a DNA sequence that causes disease or is associated with susceptibility to disease. They can be used to create genetic maps of whatever organism is being studied.

Biomarker discovery is a medical term describing the process by which biomarkers are discovered. Many commonly used blood tests in medicine are biomarkers. There is interest in biomarker discovery on the part of the pharmaceutical industry; blood-test or other biomarkers could serve as intermediate markers of disease in clinical trials, and as possible drug targets. The recent interest in biomarker discovery is spurred by new molecular biologic techniques, which promise to find the relevant markers rapidly without detailed insight into the mechanisms of a disease. By screening many possible biomolecules at a time, a parallel approach can be attempted the genomics and proteomics are some of the bioinformatic technologies used in this process. Secretomics has also emerged as an important technology in the high-throughput search for biomarkers, [6] however, significant technical difficulties remain.

An information-theoretic framework for biomarker discovery, integrating biofluid and tissue information, has been introduced; this approach takes advantage of functional synergy between certain biofluids and tissues, with the potential for clinically significant findings (not possible if tissues and biofluids were considered separately). [7] By conceptualizing tissue biofluids as information channels, significant biofluid proxies were identified and then used for guided development of clinical diagnostics. Candidate biomarkers were then predicted, based on information-transfer criteria across the tissue-biofluid channels. Significant biofluid-tissue relationships can be used to prioritize the clinical validation of biomarkers. Identification of biomarker can be metabolomics approach or lipidomics approach. The term metabolomics has been recently introduced to address the global analysis of all metabolites in a biological sample. Lipidomics refers to the analysis of lipids.

Biological scientists use these approaches for biomarker discovery by conducting various experiments involving huge funds and time. The molecular biomarkers have been defined as biomarkers that can be discovered using basic and acceptable bioinformatic platforms such as genomics and proteomics.The approach use computational power of information technology to produce benefical results with minimum cost and quick time. The biomarker identification and detection is shown in Fig 3.

Data mining techniques are used for variety of applications. In health care industry, data mining plays an important role for predicting diseases. Vijiyarani and Sudha [8] in a survey paper summarized the different algorithm of data mining used in the field of medical prediction for heart disease, breast cancer and diabetes.

Sequence mining is a topic of data mining concerned with finding statistically relevant patterns between data. The concept of sequence Data Mining and discovering sequential patterns was first introduced by Rakesh Agrawal and Ramakrishnan Srikant in the year 1995 [9]. Sequence mining is a special case of structured data mining. In general, sequence mining problems can be classified as string mining which is typically based on string processing algorithms. String mining typically deals with a limited alphabet for items that appear in a sequence, but the sequence itself may be typically very long.

Many interesting real-life mining applications rely on modeling data as Sequences. In computational biology, DNA, RNA and protein data are all best modeled as sequences. Protein sequences where each element is an amino acid that can take one of 20 possible values, or a Gene sequence where each element can take one of four possible values of nucleotide bases 'A', 'G', 'C' and 'T' [10]. In bioinformatic applications, analysis of the arrangement of the alphabet in strings can be used to examine gene and protein sequences to determine their properties. Knowing the sequence of letters of a DNA a protein is not an ultimate goal in itself. Rather, the major task is to understand the sequence, in terms of its structure and biological function. This is typically achieved first by identifying individual regions or structural units within each sequence and then assigning a function to each structural unit. In many cases this requires comparing a given sequence with previously studied ones. The comparison between the strings becomes complicated when insertions, deletions and mutations occur in a string.

Sequential pattern mining is one of the most well-known methods and has many broad applications including web-log analysis, customer purchase behavior analysis and medical record analysis. In the medical field, sequential patterns of symptoms and diseases exhibited by patients identify strong symptom/disease correlations that can be a valuable source of information for medical diagnosis and preventive medicine.

A survey and taxonomy of the key algorithms for sequence comparison for bioinformatics is presented by Abouelhoda & Ghanem [11], which include:

- *Repeat-related problems*: that deal with operations on single sequences and can be based on exact string matching or approximate string matching methods for finding dispersed fixed length and maximal length repeats, finding tandem repeats, and finding unique subsequences and missing (un-spelled) subsequences.
- *Alignment problems:* that deal with comparison between strings by first aligning one or more sequences; examples of popular methods include BLAST for comparing a single sequence with multiple sequences in a database, and ClustalW for multiple alignments. Alignment algorithms can be based on either exact or approximate methods, and can also be classified as global alignments, semi-global alignments and local alignment.

Some of the well known sequence mining algorithms are GSP, SPADE, SPAM, FREE_SPAN, PREFIX SPAN, WAP MINE, ALIGNMENTS and SPIRIT. Chetna et al [12] investigates classifying study of sequential pattern-mining algorithms into two broad categories. First, on the basis of algorithms which are designed to increase efficiency of mining and second, on the basis of various extensions of sequential pattern mining designed for certain application. Rad et al [13] use sequence pattern mining algorithms in discovery biological data.

The diversity of the applications may not be possible to apply a single sequential pattern model to all these problems. Each application may require a unique model and solution. A number of research projects were established in recent years to develop meaningful sequential pattern models and efficient algorithms for mining these patterns. Mahdi and Gabor [14] theoretically had shown three types of sequential patterns and some properties of them. These models fall into three classes are called periodic pattern, statistically pattern, and approximate pattern.

Boghey and Singh [15] reviews the state-of-the-art progress on methods of identifying sequential pattern from data base and describe a variety of models for sequential pattern mining task by classifying into two wide categories. The first one is to sequential pattern mining algorithms, which discovered sequential pattern over the static database and other one is to mining of sequential pattern over the incremented or updated database..

Sunitha Sarawagi [16] presented sequence mining and applications by reviewing techniques ranging from item set counting, MDL based discretization and Markov modelling to perform various supervised and unsupervised pattern discovery task on sequence with a case study on DNA sequence mining.

The field of sequence mining is still being actively explored spurred by emerging applications in the information extraction, bio-informatics and sensor networks. We can hope to witness more exciting research in the techniques and application of sequence mining in the coming years.

RELATED WORK

Obesity is closely related with type II diabetes, fatty liver, cardiovascular and cerebrovascular diseases, hypertension, dyslipidemia and other chronic diseases [17]. The strategies to solve obesity epidemic range from educating people about nutrition to enabling possibilities for physical exercise. Valentin and Howard [18] compare two methods for engaging individuals in exercise based on passive versus active-encouragement. In this section some research works on obesity & diabetes epidamic are presented.

Diabetes Mellitus continues to be a devastating and daunting health scourge spreading across geographical and genetic boundaries. The growing incidence of type 2 diabetes with increasing obesity reflects that obesity is an emerging risk factor for the progression of insulin resistance and subsequently to overt type 2 diabetes. Both in normoglycemic and hyperglycemic states, obese people exhibit a higher degree of hyper insulinemia that correlates with the degree of insulin resistance, in order to maintain normal glucose tolerance [19]. Following attainment of certain point, the progressive deterioration of the metabolic milieu leads to eventual failure of hyperinsulinemia to compensate fully for the insulin resistance and thereby produces impaired glucose tolerance that progress to overt diabetes [20, 21].

It is well known that body fat distribution and obesity are important risk factors for type 2 diabetes. Prediction of type 2 diabetes using a combination of anthropometric measures remains a controversial issue. Lee et al [22] study to predict the fasting plasma glucose (FPG) status that is used in the diagnosis of type 2 diabetes by a combination of various measures among Korean adults.

Khoo et al [23] provide an overview of the main physiological mechanisms associated with obesity and sleep-disordered breathing that are believed to result in metabolic and autonomic dysfunction, and review the models and modeling approaches that are relevant in characterizing the interplay among the multiple factors that underlie the development of the metabolic syndrome.

Obesity is associated with the rise of noncommunicable diseases worldwide. The pathophysiology behind this disease involves the increase of adipose tissue, being inversely related to adiponectin, but directly related to insulin resistance and metabolic syndrome (MetS). Klünder-Klünder et al [24] made a study aimed to determine the relationship between adiponectin levels with each component of MetS in eutrophic and obese Mexican children and found that adiponectin concentrations and MetS components have an inversely proportional relationship, which supports the idea that this hormone could be a biomarker for identifying individuals with risk of developing MetS.

Most common complex traits such as obesity, hypertension, diabetes, and cancers are known to be associated with multiple genes, environmental factors, and epistasis. Recently, the development of advanced genotyping technologies allows us to perform the genome-wide association studies (GWAS). For detecting the effects of multiple genes on complex traits, many approaches have been proposed for GWAS. Multifactor dimensionality reduction (MDR) proposed by Ritchie et al. [25] is one of the powerful methods for detecting epistasis, which detects high order interactions among genes. Sungyoung Lee et al [26] propose an efficient strategy to perform MDR analysis for GWAS data. Genome-wide association studies (GWAS) provide a new and powerful approach to investigate the effect of inherited genetic variation on risks of complex diseases like obesity. With recent advances in genotyping technology, genome-wide association studies are now becoming a reality.

Kanchana Narayanan and Jing Li [27] implemented a web application tool named MAVEN for Management, Analysis, Visualization and results sharing of GWA data using cutting edge technologies. Recent studies have revealed that human obesity has a microbial component: However, limited knowledge of the microbial interactions in the gut hinders our ability to design future experiments or effective treatments. New technologies like bioinformatics (e.g. high-throughput sequencing, 16S rRNA surveys) allow us to deeply sample the genetic content of a microbial environment in order to estimate its overall composition and functional capacity.

James Robert White and Mihai Pop [28] use 16S rRNA time-series sequence data from obese individuals on a one-year diet and employ mathematical modeling to study microbial population dynamics in the human gut. The model indicates several interspecific interactions in this microbial community and the impact of prebiotic and probiotic therapies for obesity through simulation.

Despite multiple efforts are being made to dampen obesity impact on the quality of life of affected patients, there remains a lot of complexity exists in the pathogenesis of obesity mediated type 2 diabetes. By virtue of endocrinal role of adipose tissue, it is known to produce a vast array of adipocyte derived factors such as tumor necrosis factor alpha, interleukin-6, leptin, adiponectin and resistin. Since many of these adipokines profoundly influence insulin sensitivity and glucose metabolism, they form a fundamental bridge between increased adiposity and impaired insulin sensitivity [29]. Although adipocytes are critical in obesity, their role in diabetes has been recognized.

Xue et al [30] has investigated the relationship between alpha 1-antitrypsin (A1AT), adiponectin, leptin, blood glucose, and insulin protein levels in human serum and obesity and found Alpha 1-antitrypsin correlates closely with obesity, and is related to other factors such as leptin, adiponectin, and insulin. Alpha 1-antitrypsin might be used as a clinical biomarker and be a potential target for treating obesity.

Recently Gerken T et al [31] performed bioinformatic analysis and reported that the variants in the fat mass and obesity associated gene are associated with increased body mass index in humans. Barcelo-Batllori S et al [32] utilizes the DIGE and Bioinformatic analysis for identification of potential drug targets of tungstate, DIGE analysis identified 20 proteins as tungstate obesity-direct targets, involved in: Krebs cycle, glycolysis, lipolysis and fatty acid oxidation, electron transport and redox. Protein oxidation was decreased by tungstate treatment, which confirmed a role in redox processes; however palmitate oxidation, as a measure of fatty acid beta-oxidation, was not altered by tungstate, thus questioning its putative function on fatty acid oxidation. Bioinformatic analyses using Ingenuity pathways highlighted peroxisome proliferator activated receptor coactivator 1 alpha (PGC-1 alpha) as a potential target. Elbers CC et al [33] identified five overlapping chromosomal regions for obesity and diabetes. These results illustrate the importance of proteomics and bioinformatics approaches for identify new therapeutic invention of obesity is a challenging subject.

Manisha Sankhla et al [34] made a study to explore the possible mechanism of obesity associated metabolic syndrome and found increased ex-pression of glucose-6-phosphate dehydroge- nase in obese subjects (more if it is associated with abdominal adiposity) might mediate the onset of obesity associated metabolic disorders by increasing oxidative stress.

METHODOLOGY

Bioinformatics is the application of computer technology to the management of biological information. Over the past 10 years, there has been a technical revolution in the life sciences leading to the emergence of a new discipline called bioinformatics [35]. Bioinformatics can be broadly defined as the creation and development of advanced information and computational techniques for problems in biology. Bioinformatics, a truly multidisciplinary science, aims to bring the benefits of computer technologies to bear in understanding the biology of life itself.

A biomarker, or biological marker, generally refers to a measured characteristic which may be used as an indicator of some biological state or condition. The term occasionally also refers to a substance whose presence indicates the existence of living organisms. Biomarkers are often measured and evaluated to examine normal biological processes, pathogenic processes, or pharmacologic response to a therapeutic intervention. Biomarkers are used in scientific field.

Genomic biomarkers are essential for understanding the underlying molecular basis of human diseases. Phan et al [36] describe a biomarker identification pipeline for cardiovascular disease, which includes: high-throughput genomic data acquisition, preprocessing and normalization of data, exploratory analysis, feature selection, classification, and interpretation and validation of candidate biomarkers.

The present research follows the above path and aims at finding the proteins responsible, a biomarker for obesity associated diabetes in two phases. The first phase of the research attempts to identify the candidate proteins/genes which are involved in these disorders through thorough literature search. The data pertaining to these proteins is extracted from the databases that are available online for free access. The functional protein sequences of these proteins in FASTA are extracted from (National Center for Biotechnology Information (NCBI), (http\\www. ncbi.nih.nlm.gov).

The sequences obtained are aligned with different alignment methods. Multiple alignment methods, a sequence miining technique try to align all of the sequences in a given query set. Molecular Biologists frequently compute multiple sequence alignments (MSA) to identify regions in protein families. Progressive alignment is a widely used approach to compute MSA. However, aligning a few hundred sequences by popular progressive alignment tools requires several hours on sequential computers. Due to the rapid growth of biological sequence databases biologists have to compute MSA in a far shorter time [37]. Classical MSA algorithms are designed to primarily capture conservations in sequences whereas couplings, or correlated mutations, are well known as an additional important aspect of sequence evolution. Hossain et al [38] present a novel approach ARMiCoRe to a classical bioinformatics problem of multiple sequence alignment (MSA) of gene and protein sequences.

The second phase of the research analyzes the data by employing Multiple Sequence Alignment using ClustalW online tool. These alignments produce a Phylogram tree along with the alignment scores.

The multiple sequence alignment compares many sequences in one go. The MSA Algorithm has three important steps.

- All pairs of sequences are aligned separately to calculate a Distance Matrix
- The guide tree is constructed from distance matrix
- The sequences are progressively aligned following the guide tree.

The ClustalW adds sequences one by one to the existing alignment to build a new alignment because of its progressive nature. Progressive in this context means, it starts with using pair wise method to determine the most related sequences and then progressively adding less related sequences initial alignment.

The flow chart of all steps involved in process of identification of biomarkers for obesity and diabetes using bioinformatic approach used in present research is shown in Fig 4. The figure also depicts proteomic approach of finding biomarkers.

RESULTS AND DISCUSSIONS

From thorough literature search, seventeen proteins (Table 1) were collected and constructed a phylogram as shown in Fig 5. From the close identification of the figure it has came to know that resistin is an important protein of obesity-associated diabetes which can be taken as a biomarker.

Numerous factors in obesity such as elevated free fatty acid levels, decreased adiponectin and increased adipocytokines are majorly responsible for evolution of insulin resistance [39]. Resistin is a one such novel putative adipocyte derived signaling molecule induced during adipogenesis [40]. It was discovered by virtue of its altered gene expression in mouse adipocytes in response to insulin sensitizers such as thiazolidinediones (TZD's) resistin was originally named for its resistance to insulin resistin circulates as trimer and hexamer with intertrimer disulfide bond and processing of these bonds may be crucial to resistin activation [40]. It is a peptide hormone that belongs to a family of tissue specific resistin like molecules [41]. Since the discovery of resistin, there remains a lot of ambiguity with regard to the functional significance of resistin. Plasma resistin levels are increased in ob/ob, db/db and diet induced obese mice [40]. Concomitantly resistin m-RNA levels in obese rodents are often found be decreased [42]. There is often a discrepancy between circulating protein levels of resistin and m-RNA content in adipocytes [43].

In animals, resistin has been shown to be secreted by adipocytes and to impair glucose tolerance and insulin action when infused into mice. A study has also reported increased resistin expression in human abdominal tissue. Several studies, however, have reported reduced resistin expression in human and rat obesity. Insulin, FFAs, and TNF-a have all been shown to inhibit resistin expression and all of these factors are elevated in obesity. Therefore, contrasting results obtained from both human and a rodent study made the role of resistin in obesity-induced diabetes is more and more controversial. The human resistin is a dimeric protein with 108 amino acids as compared to the murine resistin which comprises 114 amino acids. It raises blood glucose and insulin concentration and reduces hypoglycemic response to insulin infusion [44]. Thus it was proposed to be an important link between obesity and insulin resistance. But in human its physiological function is still debatable. This is also produced by peripheral monocytes and its level correlate with IL-6 concentration raising the possibilities that it is probably associated with inflammation induced insulin resistance.

Recently List Eo et al [45] performed proteomic analysis using MALDI-MS/MS and reported that 17 proteins out of 28 proteins are involved in the energy metabolism. Smith et al [46] study reported that a polymorphism in the promoter region was associated with resistin mRNA levels in abdominal subcutaneous fat. Associations between resistin polymorphisms and type 2 diabetes have been reported in few studies [47]. On the contrary, few other studies reported no such association between resistin polymorphisms and type 2 diabetes [48]. Variation in the resistin gene is associated with obesity and insulin related phenotypes in Finnish human population. The variation in the resistin gene is not directly involved in the beta cell dysfunction but it may play crucial role in the pathobiology of obesity and insulin resistance that resulted in type 2 diabetes [49]. Therefore, for the first time, this bioinformatics study reinforces the role of resistin in the pathophisiology of obesity mediated insulin resistance and type 2 diabetes.

CONCLUSIONS

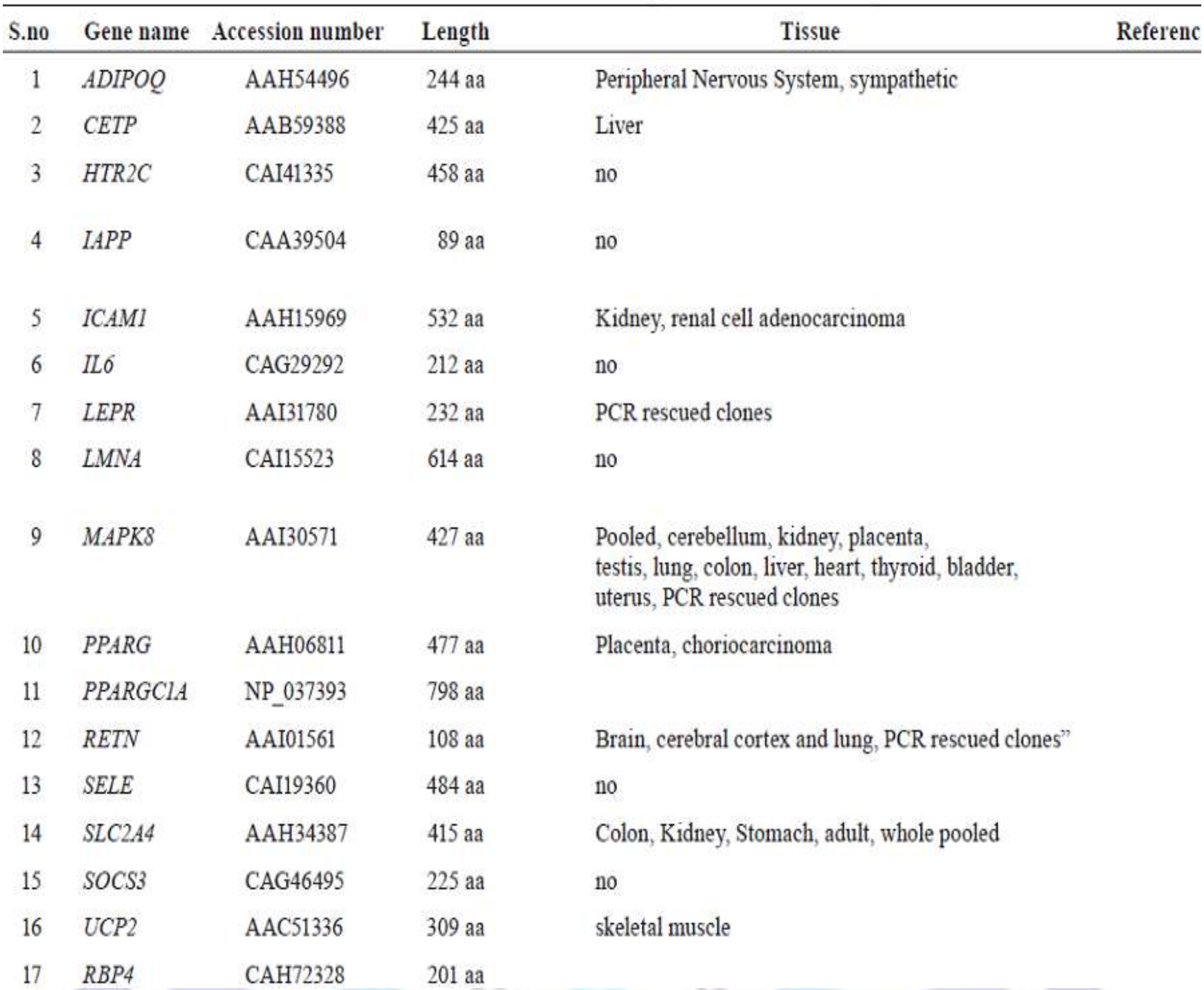
Technologies for high-throughout scanning of the human genome and its encoded proteins have rapidly developed to allow systematic analyses of human disease. Application of these bioinformatic technologies is becoming an increasingly effective approach for identifying the biological markers of genetically complex obesity and diabetic diseases. Our bioinformatic analysis reports a biomarker, resistin gene as ominous link with obesity associated diabetes.

Any rigid assessment of disease patterns will need support from well documented and curated databases. However, there are also severe practical and theoretical constraints known if applying bioinformatics as a tool for improved understanding and diagnostics of disease patterns Though lot of controversies exist with regard to the role of resistin in metabolic disorders such as obesity and diabetes mellitus, it's role is not completely excluded. Our Bioinormatics analysis once again heightens the possible role of Resistin gene that connects obesity and diabetes mellitus. In future studies like this may pave way for new therapeutic inventions of obesity associated diabetes.

Not all biomarkers should be used as surrogate endpoints to assess clinical outcomes. Biomarkers can be difficult to validate and require different levels of validation depending on their intended use. If a biomarker is to be used to measure the success of a therapeutic intervention, the biomarker should reflect a direct effect of that intervention.

There are many interesting issues that need to be studied further, Especially, the developments of specialized sequential pattern mining methods for particular applications, such as genomic and proteinomic sequence mining that may admit faults, such as allowing insertions, deletions, and mutations in DNA sequences and protein sequences, and handling industry/ engineering sequential process.

Table 1. Showing the genes/proteins that have been studied in the present study, which are believed to be involved in type2 diabetics and obesity



S.no	Gene name	Accession number	Length	Tissue	Referenc
1	*ADIPOQ*	AAH54496	244 aa	Peripheral Nervous System, sympathetic	
2	*CETP*	AAB59388	425 aa	Liver	
3	*HTR2C*	CAI41335	458 aa	no	
4	*IAPP*	CAA39504	89 aa	no	
5	*ICAM1*	AAH15969	532 aa	Kidney, renal cell adenocarcinoma	
6	*IL6*	CAG29292	212 aa	no	
7	*LEPR*	AAI31780	232 aa	PCR rescued clones	
8	*LMNA*	CAI15523	614 aa	no	
9	*MAPK8*	AAI30571	427 aa	Pooled, cerebellum, kidney, placenta, testis, lung, colon, liver, heart, thyroid, bladder, uterus, PCR rescued clones	
10	*PPARG*	AAH06811	477 aa	Placenta, choriocarcinoma	
11	*PPARGC1A*	NP_037393	798 aa		
12	*RETN*	AAI01561	108 aa	Brain, cerebral cortex and lung, PCR rescued clones"	
13	*SELE*	CAI19360	484 aa	no	
14	*SLC2A4*	AAH34387	415 aa	Colon, Kidney, Stomach, adult, whole pooled	
15	*SOCS3*	CAG46495	225 aa	no	
16	*UCP2*	AAC51336	309 aa	skeletal muscle	
17	*RBP4*	CAH72328	201 aa		

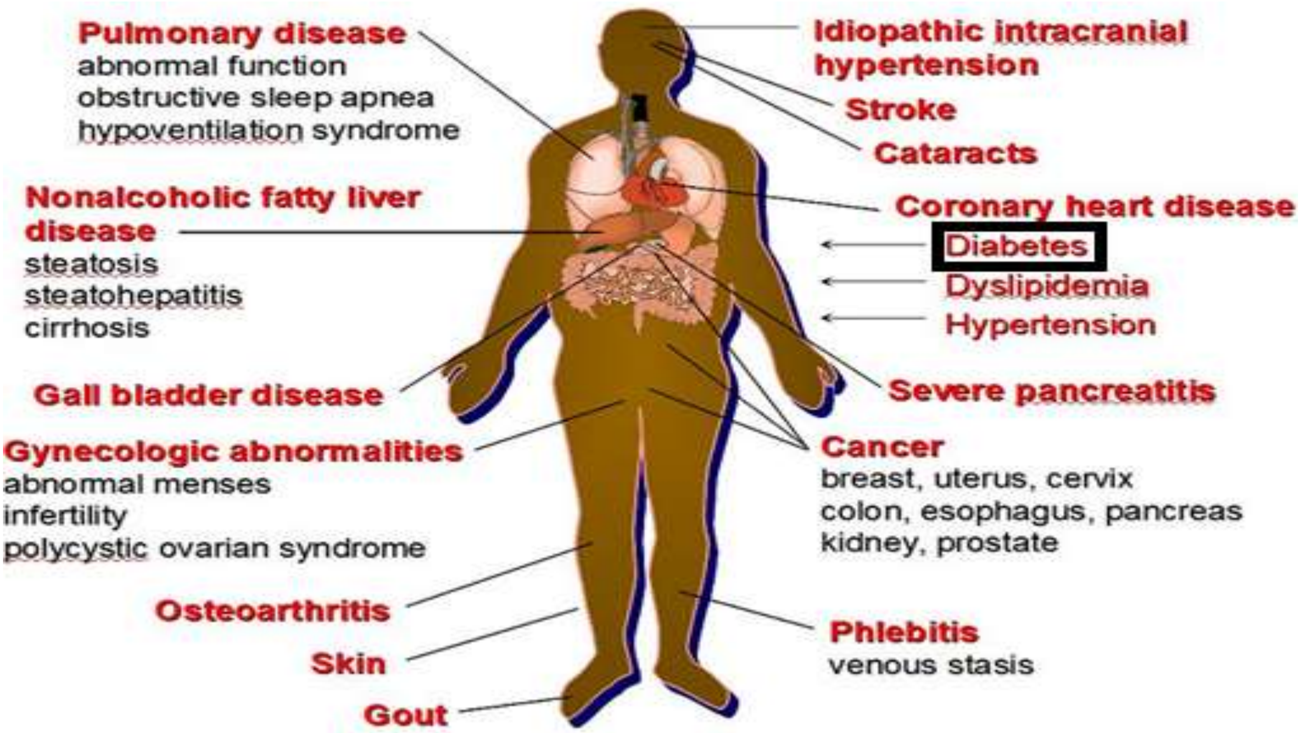


Fig1 :Medical Complications of Obesity

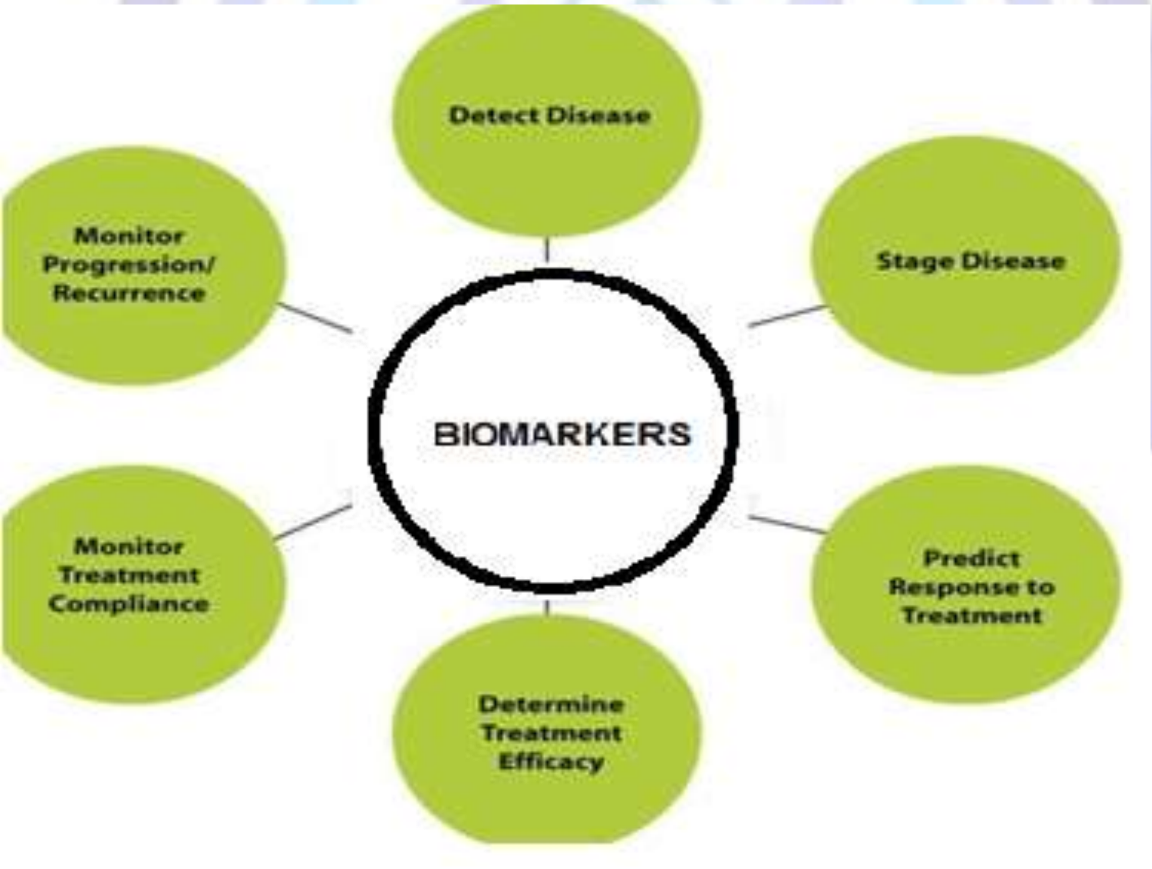

Detect Disease
Stage Disease
Predict Response to Treatment
Determine Treatment Efficacy
Monitor Treatment Compliance
Monitor Progression/ Recurrence
BIOMARKERS

Fig 2 : Biomarkers Use

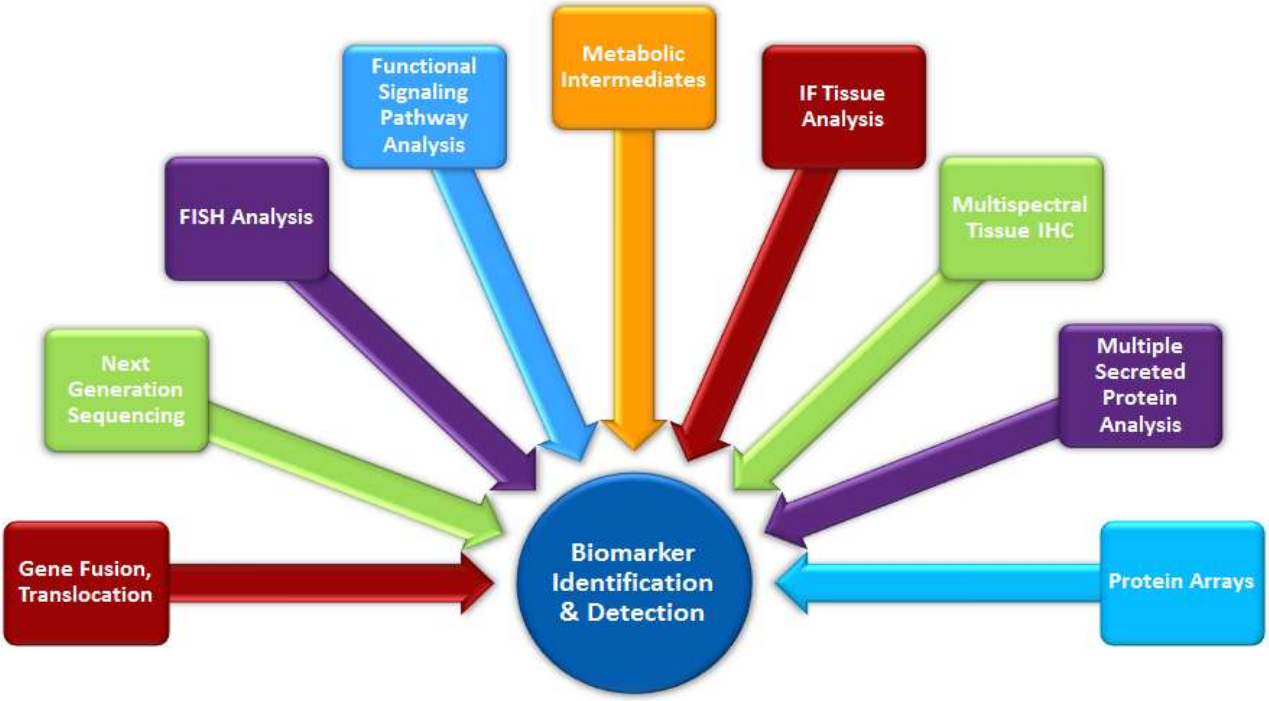

Functional Signaling Pathway Analysis
Metabolic Intermediates
IF Tissue Analysis
FISH Analysis
Multispectral Tissue IHC
Next Generation Sequencing
Multiple Secreted Protein Analysis
Gene Fusion, Translocation
Biomarker Identification & Detection
Protein Arrays

Fig 3 : Biomarker Identification and Detection

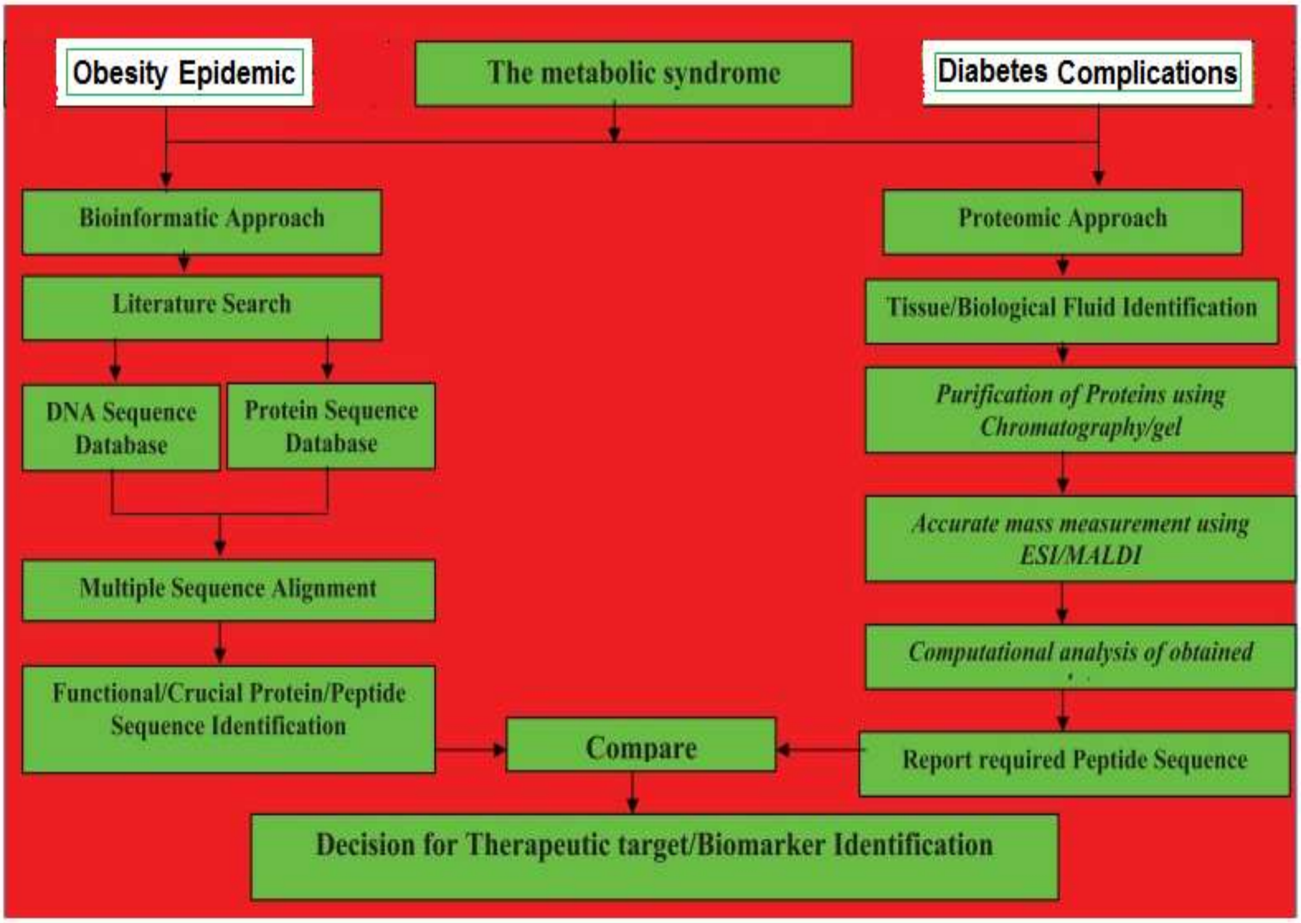

Obesity Epidemic
The metabolic syndrome
Diabetes Complications
Bioinformatic Approach
Proteomic Approach
Literature Search
Tissue/Biological Fluid Identification
DNA Sequence Database
Protein Sequence Database
Purification of Proteins using Chromatography/gel
Accurate mass measurement using ESI/MALDI
Multiple Sequence Alignment
Computational analysis of obtained
Functional/Crucial Protein/Peptide Sequence Identification
Compare
Report required Peptide Sequence
Decision for Therapeutic target/Biomarker Identification

Fig 4 : Schematic representation of obesity and diabetic complications with reference to bioinformatic and proteomic approaches for biomarker identification.

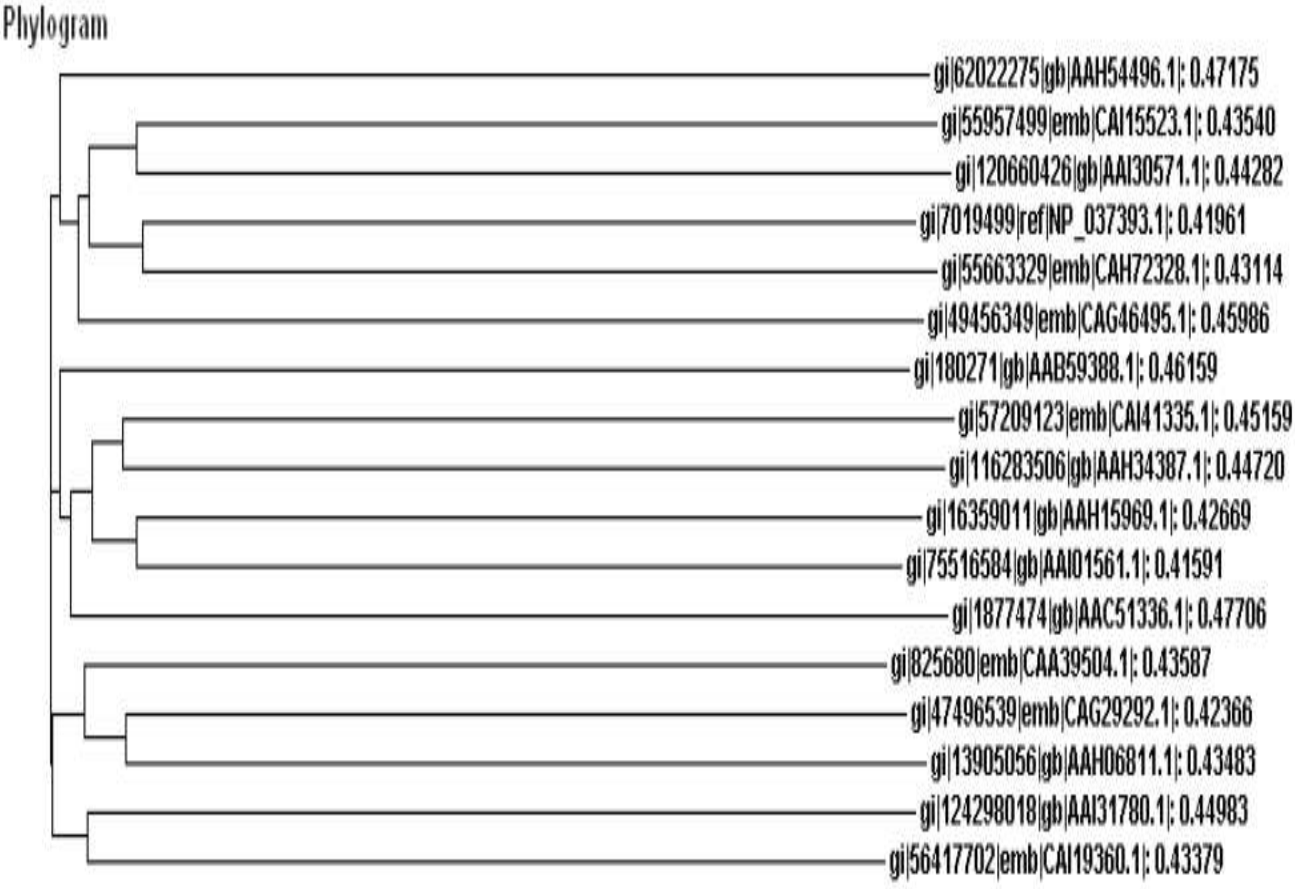

Phylogram
gi|62022275|gb|AAH54496.1|: 0.47175
gi|55957499|emb|CAI15523.1|: 0.43540
gi|120660426|gb|AAI30571.1|: 0.44282
gi|7019499|ref|NP_037393.1|: 0.41961
gi|55663329|emb|CAH72328.1|: 0.43114
gi|49456349|emb|CAG46495.1|: 0.45986
gi|180271|gb|AAB59388.1|: 0.46159
gi|57209123|emb|CAI41335.1|: 0.45159
gi|116283506|gb|AAH34387.1|: 0.44720
gi|16359011|gb|AAH15969.1|: 0.42669
gi|75516584|gb|AAI01561.1|: 0.41591
gi|1877474|gb|AAC51336.1|: 0.47706
gi|825680|emb|CAA39504.1|: 0.43587
gi|47496539|emb|CAG29292.1|: 0.42366
gi|13905056|gb|AAH06811.1|: 0.43483
gi|124298018|gb|AAI31780.1|: 0.44983
gi|56417702|emb|CAI19360.1|: 0.43379

Fig 5 : Phylogenetic tree that was constructed based on the alignment scores of all the protein sequences involved in of obesity associated with diabetes.

REFERENCES

[1] Ali H. Mokdad, Earl S. Ford, Barbara A. Bowman, William H. Dietz, Frank Vinicor, Virginia S. Bales, and James S. Marks, Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors, 2001, JAMA, Vol 289, No. 1, January 1 2003.

[2] George A. Bray, Pennington Biomedical Research Center, Medical Consequences of Obesity, J Clin Endocrinol Metab, 89(6):2583-2589, June 2004.

[3] World Health Organization, 10 facts on obesity, http://www.who.int/features/factfiles/ obesity/facts/ en/ index.html (accessed April 2012).

[4] King, H., Albert, RE., Herman, W.H., Global burden of diabetes, 1995-2025- prevalence, numerical estimates and projections; Diabetes care, 21: 1414-31, 1998

[5] http://en.wikipedia.org/wiki/Biomarker

[6] Hathout, Yetrib, Approaches to the study of the cell secretome, Expert Review of Proteomics 4 (2): 239–48, 2007

[7] Alterovitz, G; Xiang, M; Liu, J; Chang, A; Ramoni, MF, System-wide peripheral biomarker discovery using information theory, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing: 231–42, 2008.

[8] Vijiyarani and Sudha, Disease Prediction in Data Mining Technique – A Survey, International Journal of Computer Applications & Information Technology, Vol. II, Issue I, January 2013

[9] Agrawal R. And Srikant R. , Mining Sequential Patterns, In Proc. of the 11th Int'l Conference on Data Engineering, Taipei,Taiwan, March 1995

[10] Eskin, E., W. Lee, and S. J. Stolfo, Modeling system calls for intrusion detection with dynamic window sizes. Proceedings of DISCEX II, 2001

[11] Abouelhoda, M.; Ghanem, M., String Mining in Bioinformatics. In Gaber, M. M.Scientific Data Mining and Knowledge Discovery. Springer. 2010

[12] Chetna Chand, Amit Thakkar, Amit Ganatra ,Sequential Pattern Mining: Survey and Current Research Challenges, International Journal of Soft Computing and Engineering (IJSCE), Volume-2, Issue-1, March 2012

[13] Morteza Poyan Rad, Seyed Hossein Hosseiniasl and Ali Abbaszadeh Sori,Using Pattern Mining Algorithms in Discovery Biological Data, Research Journal of Applied Sciences, Engineering and Technology 4(20): 4039-4042, 2012

[14] Mahdi Esmaeili and Fazekas Gabor,Finding Sequential Patterns from Large Sequence Data, IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1, January 2010

[15] Boghey, Singh, Sequential Pattern Mining: A Survey on Approaches, International Conference on Communication Systems and Network Technologies (CSNT), IEEE, 2013

[16] Sunitha Sarawagi, Sequence data mining techniques and applications, 19th International Conference on Data Engineering, IEEE, 2003

[17] Alarm I, Lewis K, Stephens JW, et al. obesity, metabolic syndrome and sleep apnoea: all pro-inflammatory states [J]. Obesity reviews,8(2):119-127, 2007

[18] Valentin, Howard, Dealing with childhood obesity: Passive versus active activity monitoring approaches for engaging individuals in exercise, Biosignals and Biorobotics Conference (BRC), ISSNIP, IEEE 2013

[19] Bonadonna RC, Groop L, Kraemer N, Ferrannini E, et al. Obesity and insulin resistance in humans: a dose-response study. Metabolism, 39:452-9, 1990

[20] DeFronzo RA, Bonadonna RC, Ferrannini E. Pathogenesis of NIDDM. A balanced overview. Diabetes Care,15:318-68, 1992

[21] DeFronzo RA. Lilly Lecture. The triumvirate: beta cell, muscle, liver. A collusion responsible for NIDDM. Diabetes, 37:667-87, 1988

[22] Lee B, Bum Ju Lee, Ku B. ; Pham, Nam, Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes, IEEE Journal of Biomedical and Health Informatics Volume:PP , Issue: 99, 2013

[23] Khoo , Oliveira, Limei Cheng, Understanding the Metabolic Syndrome: A Modeling Perspective, IEEE Reviews in Biomedical Engineering (Volume:6) 2013

[24] Miguel Klünder-Klünder, Samuel Flores-Huerta, Rebeca García-Macedo, Jesús Peralta Romero and Miguel Cruz, Adiponectin in eutrophic and obese children as a biomarker to predict metabolic syndrome and each of its components, BMC Public Health 13:88, 2013

[25] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F.F. Parl, J. H. Moore, Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, Am.J. Hum. Genet, vol. 69, pp. 138-147, 2001.

[26] Sungyoung Lee, Sohee Oh, Min-Seok Kwon, Seungyeoun Lee, and Taesung Park, Two-way interaction analysis of Obesity Trait from Korean population Using Generalized MDR, IEEE International Conference on Bioinformatics and Biomedicine Workshops, 2010

[27] Kanchana Narayanan and Jing Li, Visualization and Functional Analysis of Genome-Wide Association Results, IEEE Ohio Collaborative Conference on Bioinformatics, 2009

[28] James Robert White and Mihai Pop, Microbial Dynamics of Human Obesity, IEEE 2009

[29] Fassshauer M, Paschke R. Regulation of adipocytokines and insulin resistance. Diabetologia, 46:1594-1603, 2003

[30] Xue GB, Zheng WL, Wang LH, Lu LY, Alpha 1-antitrypsin-A novel biomarker for obesity in humans, Saudi Med J. ;34(1):34-9, 2013

[31] Gerken T, Girard CA, Tung YC, Webby CJ, et al. The obesity-associated FTO gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. Science, 318:1469-72, 2007

[32] Barceló-Batllori S, Kalko SG, Esteban Y, Moreno S, et al. Integration of DIGE and bioinformatics analyses reveals a role of the anti-obesity agent tungstate in redox and energy homeostasis pathways in brown adipose tissue. Mol Cell Proteomics. 2007

[33] Clara C. Elbers, N. Charlotte Onland-Moret, Lude Franke, Anne G.Niehoff, Yvonne T. van der Schouw and Cisca Wijmenga., A strategy to search for common obesity and type 2 diabetes genes. Trends in Endocrinology & Metabolism, 18:19-26, 2007

[34] Manisha Sankhla, Keerti Mathur, Jai Singh Rathor, Is there any role of glucose-6-phosphate dehydrogenase in obesity induced metabolic disorder, Health 4, Vol.4, No.12A, 1530-1536, 2012

[35] Miskowski, J.A, Howard, D.R, Abler, M.L, Grunwald, S.K,Design and implementation of an interdepartmental bioinformatics program across life science curricula, Biochemistry and Molecular Biology Education 35 (1), pp. 9-15,2007

[36] Phan, J.H, Quo, C.F,Wang, M.D, Cardiovascular Genomics: A Biomarker Identification Pipeline, IEEE Transactions on Information Technology in Biomedicine Volume:16 , Issue: 5, 2012

[37] Oliver, T, Schmidt, B, Nathan, D, Clemens, R, Maskell, D,Multiple sequence alignment on an FPGA , Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS 2, pp. 326-330,2005.

[38] Hossain, K. ; Virginia Tech, Blacksburg ; Patnaik, D. ; Laxman, S. ; Jain, P, Improved Multiple Sequence Alignments Using Coupled Pattern Mining, IEEE/ACM Transactions on Computational Biology and Bioinformatics,Volume:PP , Issue: 99, 2013

[39] Sangeeta R Kashyap, Ralph A Defronzo. The insulin resistance syndrome: physiological considerations. Diabetes Vasc Dis Res;4:13-19, 2007

[40] Steppan CM, Bailey ST, Bhat S, Brown Ej, et al. The hormone resistin links obesity to diabetes. Nature; 409:307-312, 2001

[41] Steppan CM, Brown Ej, Wright CM, Bhat S, et al. A family of tissuespecific resistin-like molecules. PNAS.; 98: 502-506, 2001

[42] Rajala MW, Lin Y, Ranalletta M et al. Cell type-specific expression and co regulation of murine resistin and resistin-like molecule-alpha in adipose tissue. Mol Endocrinol; 16:1920-1930, 2002

[43] Lee JH, Bullen JW Jr, Stoyneva VL, Mantzoros CS. Circulating resistin in lean, obese and insulin-resistant mouse models: lack of association with insulinemia and glycemia. Am J Physiol Endocrinol Metab; 288(3):E625-632, 2004

[44] Ukkola O. Resistin-a mediator of obesity associated insulin resistance or an innocent bystander? Eur J Endocrinol; 147:571-574, 2002

[45] List EO, Berryman DE, Palmer AJ, Qiu L, et al. Analysis of mouse skin reveals proteins that are altered in a diet-induced diabetic state: a new method for detection of type 2 diabetes., Proteomics; 7:1140-9, 2007

[46] Smith SR, Bai F, Charbonneau C, Janderova L, Argyropoulos G. A promoter genotype and oxidative stress potentially link resistin to human insulin resistance. Diabetes; 52:1611-1618, 2003

[47] Tan MS, Chang SY, Chang DM, Tsai JC, Lee YJ. Association of resistin gene 3'-untranslated region +62G>A polymorphism with type 2 diabetes and hypertension in a Chinese population. J Clin Endocrinol Metab; 88:1258-1263, 2003

[48] Ochi M, Osawa H, Onuma H, et al. The absence of evidence for major effects of the frequent SNP +299G>A in the resistin gene on susceptibility to insulin resistance syndrome associated with Japanese type 2 diabetes. Diabetes Res Clin Pract; 61:191-198, 2003

[49] Conneely KN, K. Silander, L. J. Scott, K. L. Mohlke, et al. Variation in the resistin gene is associated with obesity and insulin-related phenotypes in Finnish subjects. Diabetologia, 47(10):1782-1788, 2004

Author' biography with Photo

Lalitha Saroja Thota received M.Tech in Software Engineering from Jawaharlal Nehru Technological University, Hyderabad, India in 2010 and M.Sc in Computer Science from Annamalai University, Chennai, Tamilnadu, India in 2008. She is pursuing Ph.D in Computer Science and Engineering from Acharya Nagarjuna Univeristy, Guntur, A.P, India. Her areas of interest are bioinformatics and data mining. She has more than 6 research papers in international journals and conferences to her credit.

Allam Appa Rao received Ph.D in Computer Engineering from Andhra University, Vishakapatnam, India in 1984. He has four decades of professional experience. His primary area of research is Bioinformatics. He has 50^{+} strong research team. 40 scholars were awarded Ph.D degrees under his guidance. He is currently Director, CRRao AIMSCS, University of Hyderabad campus, Hyderabad, A.P, India. He has more than 150 research papers in international journals and conferences to his credit.