ORIGINAL PROF-2165

MODY GENES;

LINKAGE ANALYSIS AND SUBGROUP DISCOVERY FROM TEXT DOCUMENTS

Ms. Attiya Kanwal, Ms. Sahar Fazal, Mr. Sohail Asghar, Mr. Muhammad Naeem.

ABSTRACT…..Objective: Genetic screening of Maturity Onset Diabetes of the Young (MODY) genes has not been performed in Pakistan so far; albeit MODY genes have been noticed in local population. A relevant research will help to establish a scheme for identification and treatment of MODY. **Data Source:** The data source for the subgroup discovery was retrieved from PubMed. **Study Design:** Family affected by MODY were contacted personaly for descriptive study. The family history was obtained from the representative members of the family and pedigree was drawn. **Setting:** The extensive clinical examination of both patients and their unaffected normal relatives was carried out by expert clinician. **Period:** Specific primers for region of interest in genomic DNA were designed at the IBGE Islamabad using Primer3 during last quarter of 2011. **Materials & Methods:** Mutation detection was performed followed by pattern discovery using subgroupd discovery technique. **Results:** Unidentified MODY genes facilitating the cause of a specific diabetes in European population may play a central role for diabetes characterized by autosomal dominant transmission in Pakistani population. Exclusion study indicates that there is no linkage to the known loci of MODY. Similarly genetic screening results suggest that no mutation is indicated in this examined family in MODY genes. **Conclusion:** There may be some environmental factors involved in causing this disease in this family; otherwise this disease is due to mutation in other reported MODY genes which are not screened in this study. Subgroup discovery results point out that all the reported MODY genes have association among themselves revealing 580 patterns.

Key words: Maturity-onset diabetes of the young, Diabetes Mellitus, Subgroup Discovery, Data mining.

Article Citation
Kanwal A, Fazal S, Asghar S, Naeem M. Mody genes; linkage analysis and subgroup discovery from text documents. Professional Med J 2013;20(4): 623-633.

INTRODUCTION

In humans, diseases are caused either by some microorganisms or due to some mutations in the chromosomes. The diseases caused by microorganisms are not inherited (except some viral diseases). But the diseases caused by the mutations at gene level are normally heritable and may be lethal. Genetic disorder arises when one or both copies of a specific gene undergo an alteration known as mutation. Defective genes may also be inherited from the parents. Currently about 4,000 genetic disorders are known, with more being discovered.

MODY relies on the elderly classification of diabetes into juvenile-onset and maturity-onset diabetes. An etiology-based classification for diabetes has been revised and introduced by both the American Diabetes Association (ADA) and World Health Organization (WHO). The group of "Genetic defect in B-cell function" has now included MODY as its part with its sub classification according to the gene involved[1].

Tattersall et al., first described MODY in 19742. Tattersal (1975), after taking into account a group of young people with diabetes who were treated without insulin 2 years or more after diagnosis[2]. Since the 1970s there has been great interest in MODY as it is a genetic form of diabetes. MODY is an ancestral form of early-onset type[2] diabetes. It is a monogenic form of diabetes mellitus inherited in autosomal dominant mode[3,4]. It is primarily an outcome of impaired B-Cells of pancreas[3,4]. MODY is not a single entity but represents genetic, metabolic, and clinical heterogeneity[5].

MODY generally develops in middle age, and mainly coupled with primarily scantiness of insulin secretion[6].

BACKGROUND

Various types of MODY have been discovered in European populations albeit still MODY is misdiagnosed as T2DM. MODY genes are named as MODY 1 to MODY 11 according to the year they got

recognized in MODY patients.

Fajans performed the linkage analysis of a family that consists of 360 members spanning 6 generations and 74 members with diabetes, including those with MODY[7]. This family had been studied prospectively since 1958. Linkage studies showed that the gene responsible for MODY in this family is tightly linked to 20q12-q13.1. This gene which maps to this chromosome is known as HNF1A. However, Fajans did not screen the HNF1A for detection of new mutation[7].

Similarly Froguel et al., also performed linkage analysis on 16 French families with MODY observing the linkage of the disease with GCK[8]. There was statistically significant evidence of genetic heterogeneity, with an estimated 45 to 95% of the 16 families showing linkage to glucokinase. Because glucokinase is a key enzyme of blood glucose homeostasis, the results suggested a pathogenetic connection.

A handful of researches are accessible in the literature for discovering interesting relation with respect to a property of interest or a specific target variable in a data. This aspect of data mining has received a lot of deliberation among the researchers in recent times. A succinct analysis of some topical revelation associated with subgroup discovery is presented here. Numerous algorithms have been developed for subgroup discovery. These algorithms can be categorized by making a distinction between extension of classification rule algorithm and extension of association rule algorithm.

METHODOLOGY

Figure 1 represents the conceptual model of proposed methodology for the first phase of this research study.

PEDIGREE DRAWING

Pedigree drawing was performed using the Cyrillic (v2.10) program for genetic implication by the standard method[9]. The exact genealogic relationships for all the affected individuals were obtained through the extensive personal interviews of elders of the families. Males were symbolized by squares and females by circle. The normal individuals were designated with unfilled symbols while the affected individuals by filled symbols. The individuals who are dead are designated by "/" symbol. Each generation was indicated by Roman numeral. The individuals within a generation were designated by Arabic numerals.

COLLECTION OF BLOOD SAMPLES

Blood samples were drawn with informed consent from all members of the family by 5 c.c. syringes in 10ml vaccutainer tubes (Becton Dickinson, Franklin Lakes, and NJ.) containing acid citrate dextrose (ACD) or heparin. The blood was stored at 4oC until DNA extraction.

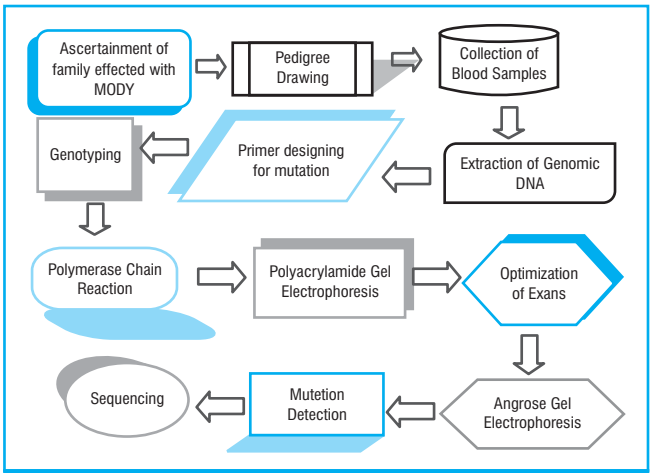

Fig. 1: Proposed methodology for linkage analysis/genetic screening

EXTRACTION OF GENOMIC DNA

Genomic DNA was extracted from blood samples using Phenol chloroform method or organic method. This method of DNA extraction from peripheral blood lymphocytes involves 3 days procedure.

PRIMER DESIGNING FOR MUTATION DETECTION

A primer is a short synthetic oligonucleotide which is

used in many molecular techniques from PCR to DNA sequencing. Specific primers for region of interest in genomic DNA were designed at the Institute of Biomedical and Genetic engineering (IBGE), Islamabad using Bioinformatics Primer Designing Tool Primer3 (http://primer3.sourceforge.net/). These primers were then synthesized commercially (Operon, Germany).

Following three genes were genetically screened for mutation detection:

1. HNF4α (MODY 1)
2. GCK (MODY 2)
3. HNF1α (MODY 3)

LINKAGE ANALYSIS

Linkage Analysis is a powerful method of gene mapping to determine if two or more genetic traits – i.e. a marker locus and a disease trait are co segregating within a pedigree. Linkage analysis, is used when the location of a gene is known, although the gene itself and its function is not.

GENOTYPING

To determine the linkage or exclusion study of the family to six known loci for MODY, a minimum of two microsatellite markers each of the candidate regions of these loci for MODY were genotyped in all the available individuals of this family. Microsatellite Markers used for genotyping this family were selected from Marshfield Map of STRPs set Version 8 available at IBGE.

POLYMERASE CHAIN REACTION

Standard PCR reactions will be performed using Bioline or Promega products to amplify the DNA. PCR reactions were carried out on ~40ng/ul of DNA in different reaction volumes. A typical reaction mixture consisted of 1 ul 10X PCR buffer [100mM Trizma-HCL, pH 8.3; 500mM KCL; 15mM MgCl2 ; 0.01% (w/v) gelatin)(Bioline)], 2mM dNTPs (Promega), 20uM Forward and reverse primers (Operon, Germany), 5U/ul ul of Taq DNA polymerase (Institute of Biomedical and genetic engineering, Islamabad, Pakistan, Fermentas) and autoclaved deionized water. Each PCR reaction will be consistently performed using the GeneAmp PCR system 9700 thermocycler (Applied Biosystems, foster city, California, USA).

Typical parameters for cycling consisted of an initial denaturation step for 2 min. at 94oC, followed by 35 cycles comprising of denaturation at 94oC for 45 seconds, an annealing step for again 45 seconds and extension at 72oC for 45 seconds with a final extension step at 72oC for 10 minutes. Annealing temperature will be initially calculated by estimating the Tm using the formula: Tm = [4(G+C) + 2(A+T)], and subtracting 5oC. The specific annealing temperature for each pair of primers will be then optimised to give a maximum yield of the required PCR product and minimise the output of non-specific products.

POLYACRYLAMIDE GEL ELECTROPHORESIS

Denaturing polyacrylamide gel electrophoresis technique was employed to distinguish the different alleles of the polymorphic microsatellite markers used in linkage/homozygosity analysis. The standard apparatus used (Bio-Rad Sequi-Gen Cell) measured 38 * 50 cm and utilized thick spacers. Sterilize Biorad SequiGen glass plate with 70% ethanol. They were then assembled according to the manufacturers' instructions and placed in a gel casting tray. The back plate (Integral plate chamber, IPC) was siliconized prior to assembly with Sigmacote (Sigma, RPMI – 1640 MEDIUM) according to the manufacturer's instructions. The concentration of acrylamide required for maximum resolution depends upon the size of the DNA fragments being resolved. In microsatellite analysis, the PCR products are usually in the range of 80 to 400bp and thus 10% polyacrylamide gels will be routinely used.

MUTATION DETECTION

Rapid detection of new mutations and substitutions in large number of samples is quite important for clinical diagnostics, population genetics, and molecular epidemiology.

To identify genes involved in susceptibility to MODY and to discover new genes and mutations contributing to MODY, following three genes for three different types of MODY will be screened for mutation detection:

1. HNF4α (MODY 1)
2. GCK (MODY 2)
3. HNF1α (MODY 3)

Mutation detection in any of the genes will be done through Single Strand Conformation Polymorphism technique.

OPTIMIZATION OF EXONS

Prior to SSCP we must set the condition and annealing temperature of each exon. Optimization of exons was done by polymerase chain reaction. Amplification was done in a final volume of 15μl containing 40ng genomic DNA, 1X PCR buffer (Bioline), 0.45mM MgCl2, 200μM dNTPs (Promega) 1μM each forward and reverse primer, 1U Taq DNA polymerase. Amplification was done with initial denaturizing at 94°C for 2 min, 35 cycles each consisting of denaturation at 94°C for 45 s, annealing for 45 s, extension at 72°C for 45 s and final extension for 10 min at 72°C. Forward and reverse primers for each exon with specific sequence were used in amplification. Amplified products were run on 2% (w/v) agarose gel containing 0.5μg/ml ethidium bromide at constant power supply of 200 volts for 30 minutes.

AGROSE GEL CASTING

The 2 percent agarose gel was prepared by adding 6 gms of agarose in 300 ml 1X TBE buffer (Tris Borate EDTA). The mixture was heated in a microwave for approximately 2-3 min. The mixture was swirled timely using hot gloves until all the agarose were dissolved, and then 6 μl of ethidium bromide was added in it. The agarose solution was then placed in water bath at 55˚C with constant shaking for 20 mins. Gel was poured in a gel-casting tray containing 3 combs. No air bubble should be trapped during gel pouring. The gel was then allowed to polymerize at room temperature for 30 min and stored in a gel tray containing deionized water.

GEL LOADING

Five μl of Orange G dye was added to 5 μl Amplified DNA sample (PCR product) and this mixture was loaded in wells using a micropipette. The 100 bp ladder (Fermentas) was also loaded in a well for size reference.

LADDER

The 100 bp ladder (Fermentas, Gene Rule 100 bp ladder, ready to use) used for approximate quantification and sizing of wide range of double standard DNA fragment on gel. The ladder composed of fourteen chromatography purified individual DNA fragments including two reference bands of 1000 and 500 bp for easy orientation. The ladder was premixed with the 6X DNA loading dye for direct loading on gel.

GEL ELECTROPHORESIS

After the gel was loaded, it was placed in a gel tank i.e. Maxicell electrophretic gel system filled with 1X TBE buffer and the negative and positive indicator on the cover and apparatus chamber were properly oriented. Both electrodes were carefully plugged into their respective input plugs into the power pac 3000 (BIO RAD) and gel was then allowed to run at 200V for 30 mins. The current was turned off according to the distance travelled by the tracking dye on gel. For best results, the tracking dyes were allowed to migrate 3.5 to 4 centimeters from the wells towards the end of the gel for adequate separation of DNA bands.

GEL DOCUMENTATION

In order to visualize the DNA, the gel was placed on a

transilluminator (Syngene, Cambridge, UK), (UV light of wavelength 254 nm). Once the gel was placed on a UV transilluminator, the sample molecule, e.g., DNA fluoresces as bright bands. Upon UV transillumination, the PCR product was visible on the photograph as bright DNA bands, and size was depicted by 100bp ladder, which was loaded for size reference.

GEL INTERPRETATION

Now DNA samples were typically run alongside a DNA ladder, this had many different molecules of defined sizes. Now, the size of our PCR product was estimated according to the fact that the band in the ladder that had migrated to the same distance (or thereabouts) as DNA sample, it is possible to equate the sample band size to the one in the ladder.

SINGLE STRAND CONFORMATION POLYMORPHISM

A single nucleotide change in a particular sequence, as seen in a double-stranded DNA, cannot be distinguished by electrophoresis, because the physical properties of the double strands are almost identical for both alleles. After denaturation by formamide, single-stranded DNA undergoes a 3-dimensional folding and may assume a unique conformational state based on its DNA sequence. The difference in shape between two single-stranded DNA strands with different sequences can cause them to migrate differently on an electrophoresis gel, even though the number of nucleotides is the same, which is, however, a shortfall of SSCP.

CLASSIFICATION RULE LEARNING :

ASSOCIATION RULE LEARNING

Association rule learning is machine learning approach in which interesting relations between variables are discovered. APRIORI is the best known algorithm to mine association rules[10]. It uses a breadth-first search strategy to count the support of item sets and uses a candidate generation function which exploits the downward closure property of support.

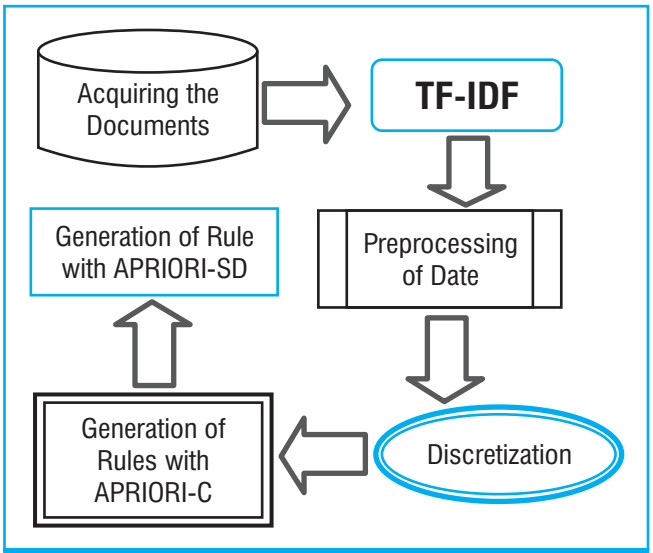

Fig.2: Proposed methodology of subgroup discovery

Data mining association rules has two main phases (i) To find all frequent item sets that have support above the predetermined minimum threshold. (ii) Using these frequent item sets, generate strong association rules that have confidence above the predetermined threshold. An association rule is an implication of the form X Y. X is an antecedent and Y is the consequent. The quality of an association rule can be determined by its confidence and support. An association rule X Y holds with support s, where s is the percentage of transaction in D that contain all the item sets that are union of sets X and Y. This is taken to be the probability, p (X U Y) Kavsek and Lavrac (2006).

An association rule X Y has confidence c in the transaction D, where c is the percentage of transactions in D containing X that also contain Y. This is taken to be the condition probability, p (*Y*I*X*) Kavsek and Lavrac (2006).

That is:

$s(X \quad Y) = p(X \cup Y)$

$c(X \quad Y) = p(Y \mid X)$

$= \frac{s(X \cup Y)}{s(X)}$

APRIORI-C

APRIORI-C is an extension of APRIORI algorithm for the purpose of classification by decreasing time complexity by feature subset selection, and improving the understandability of results by rule post processing. It is developed by implementing the following steps:

- Discretization of continuous attributes
- Binarization of discrete attributes
- Data preprocessing

OPTIMIZATION OF APRIORI-C

Optimization of APRIORI-C is carried out by taking in consideration only those rules which are with the single target item at the right hand side. In this way memory consumption can be decreased.

POST-PROCESSING BY RULE SUBSET SELECTION

The set of induced rules can be post-processed by rule ordering and best rule subset selection. In order to avoid the problems of rule redundancy, incapability of classification and poor accuracy, APRIORI-C selects the best rules by chosing the rules with highest support. Then it eliminates all the records in the dataset that are covered by this rule. The remaining rules are sorted according to the support. This process is repeated until B best rules are selected.

In APRIORI-C, to select the B best rules, the record weighting techniques is used. This technique of "example weighting" that was implemented in APRIORI-C is similar to the "use B best rules." The difference is that the covered records are not eliminated but instead their weights are decreased and the covered records are eliminated when their weights fall below a given threshold. However this procedure did not perform well in the experimentation of APRIORI-C. APRIORI-SD uses the example weighting schemes for the task of subgroup discovery[11]. This weighting scheme is used in the post processing step of selecting best rules.

WEIGHTING SCHEMES

The weighting schemes of the APRIORI-SD algorithm works in a way that the positive examples are not deleted when the currently "Best rule" is selected in the post processing step. Despite it, a rule is selected each time, the algorithm stores with each example a "count" variable that tells how many rules the examples has been covered so far. The weights of all examples are initialized to "1".

The weights of the positive examples that are covered by the selected rule are decreased by the formula $w(e_j,i)=1/i+4$[12,13]. Initial weights of all examples are same $w(e_j,o)=1$. While in the following iterations, the contributions of examples are inversely proportional to their coverage by already selected rules.

RULE EVALUATION MEASURE: WEIGHTED RELATIVE ACCURACY

Weighted relative accuracy measure (WRAcc) is a new rule evaluation measure. In APRIORI-SD WRAcc is used to evaluate the quality of induced rules instead of support in post processing step of best rule selection.

The following notations are used in our implemented algorithm:
n(X): Number of examples
covered by the rule X Y
n(Y): Number of examples of class Y
n(X.Y): Number of true position examples
For the corresponding probabilities notations p(X), p(X.Y) are used. In terms of association rule learning, the accuracy of the rule is defined as Acc(X Y)=Conf (X Y)=p(YIX)=p(X.Y)/p(X)

Weighted Relative Accuracy is defined as follows[12,13]:
WRAcc(X Y)
=p(X)
.(p(YIX)
-p(Y))

This definition of weighted Relative Accuracy has two parts: one is p(X) and the second is (p(YIX)-p(Y)). The first term is the generality and the second one is the relative accuracy. This accuracy is gained by the rule (X Y) relative to the fixed rule true Y that shows that all instances belong to class Y. The rule (X Y) is of our interest only in the case when it improves upon the default accuracy. Here the term generality refers to the "weight". Weighted relative accuracy trades off the generality of the rule p(X) and the relative accuracy (p(YIX)-p(Y)).

MODIFIED WRACC WITH EXAMPLES WEIGHTS

To handle the example weights by considering different parts of the example space in the post processing step of best rule selection, APRIORI-SD uses WRAcc as a quality measure with certain modifications. It is defined as[12,13]:

wWRAcc(X Y)
=n'(X)/N' (n'(XY)/n'(X) - n'(Y)/N')

The improved wWRAcc' that is used in APRIORI-SD in the post processing step of best rule selection is obtained by making the following adaptation in the above definition of wWRAcc.
n'(Y)/N' wWRAcc is replaced with n(Y)/N
it will force the wWRAcc' to reflect the improvement of the rule's accuracy wit respect to the accuracy of the default rule *true* *Y*.

So the improved definition of weighted relative accuracy measure is as follows Kavsek and Lavrac (2006):
wWRAcc$^{(X \quad Y)}$
=n'(X)/N' (n'(XY)/n'(X) - n'(Y)/N)
We experimentally evaluate our approach of APRIORI-SD on the data set of 11 reported MODY genes.

EVALUATION & ANALYSIS OF RESULTS

In the proposed preprocessing methodology related to wet lab at IBGE Islamabad, we performed the PCR and SSCP techniques on a large inbred Pakistani family. The subgroup discovery algorithm was then evaluated for the set of documents related to MODY genes. The experiment was performed by taking the blood samples of all the members of the affected family. Genomic DNA was extracted from these blood samples. Specific primers for region of interest in genomic DNA were designed using Bioinformatics Primer Designing Tool Primer3 and were then synthesized commercially. The wet lab work constructed in IBGE Islamabad is further divided into two parts. The first part is to determine the linkage to the known loci of MODY and the second one is mutation detection. A minimum of two microsatellite markers each of the candidate regions of these loci for MODY were genotyped in all the available individuals of this family.

Each lane represents a family member as shown by Table-I.

Lane 1: II: 6	Lane 2: III: 5
Lane 3: III: 2	Lane 4: IV: 3
Lane 5: IV: 1	Lane 6: IV: 8
Lane 7: IV: 6	Lane 8: IV: 7
Lane 9: IV: 9	Lane 10: IV: 7
Lane 11: IV: 11	Lane 12: V: 1
Lane 13: IV: 5	Lane 14: IV: 4
Lane 15: IV: 2	Lane 16: V: 2
Lane 17: V: 4	Lane 18: IV: 2
Lane 19: IV: 3	Lane 20: III: 7
Lane 21: IV: 12	Lane 22: V: 5

Table-I. Representation of lane.

These gel photographs are then analyzed to determine the different alleles of the polymorphic microsatellite markers used in linkage/homozygosity analysis.

The following findings come forward from this linkage analysis and exclusion study:

1. The analysis of the gel pictures indicate that there is no linkage to the known loci of MODY which are used for linkage in this study.
2. The unidentified MODY genes that facilitate to cause MODY in European population may play a central role for diabetes characterized by autosomal dominant transmission in Pakistani population.
3. The examined Pakistani family which is clinically diagnosed with MODY may be due to mutation in other MODY genes except the genes that are screened in the present research study for this family.

To detect the genes that are involved in susceptibility to MODY and to discover new genes and mutations contributing to MODY, the first three MODY genes, HNF4A, GCK and HNF1A were screened by SSCP technique. Prior to SSCP the coding regions (Exons) of all the genes are optimized by setting the annealing temperature of each exon through PCR. The amplified DNA sample (PCR product) along with Orange G dye is then loaded in the prepared 2% Agrose Gel. After loading the product the gel was placed in electrophretic gel tank for SSCP. After the run is completed, the gel was stained with ethidium Bromide and visualised in UV translluminator.

In SSCP technique there occurs a conformational difference in single stranded DNA sequence of identical length in case of mutation. There is no double stranded DNA. It means that the genes that are genetically screened have no mutated sequence. The findings that come to our knowledge from SSCP technique are that:

- There may be a chance that the examined family which is clinically diagnosed with MODY may have this type of diabetes due to mutation in other MODY genes except the genes that are screened in the present research study for this family.
- There may be some environmental factors that are involved in causing this disease in this family.

PRESENTATION OF FINDINGS, HARDWARE RESULTS

The second phase of this research study is to find interesting pattern by subgroup discovery algorithm with reverence to definite assests. For this purpose a well known subgroup discovery algorithm APRIROI-SD that was already proposed by Lavrac in 2006 was implemented in C#. We then evaluated the APRIORI-SD for the data set of MODY genes.

DOCUMENT RETRIEVAL

The data source for the APRIORI-SD system is PubMed database. Citations may include links to full-text content from PubMed Central and publisher web sites. We need to download all the documents containing the genes name. The abstracts of all the papers are downloaded from Pubmed database by sending the gene name as a query.

SET OF MODY GENES

{BLK, CEL, GCK, HNF1A, HNF4A, IPF1PDX1, KLF11, NEUROD1, PAX4, TCF2HNF1B, INS}

Around 10,000 papers were found in PUBMED that are relevant to these genes. The papers related to one gene are saved in one doc file. In this way we get 11 doc files for these 11 genes. These all doc files are then saved in one doc file.

TF-IDF

The next step is the pre-processing of documents. Let D be a set of text documents of MODY genes represented as D = {d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11}. The text document set D is converted from unstructured format into some common representation using the text processing techniques. The input data set D is preprocessed by removing the stop words. The Term Frequency Inverse Document Frequency (TF-IDF) algorithm is then applied on this

data set to find the frequency and weight of each word or term in the document. The results are stored in Microsoft Excel worksheet where rows represent the terms and columns represent the term frequency and term weight of each term. Total 56350 terms are extracted from the input data set D.

GENERATION OF RULE WITH APRIORI-S

The rules that are obtained by APRIORI-C are ordered by weighted relative accuracy measure. Only first 66 rules are extracted out of 580 rules as these 66 rules cover the relation between all the 11 MODY genes with each other. Some genes occur in more than one rules but with different frequencies, like association rule INS CEL occur both in rule 1 and rule 2 with frequencies INS (> 126) and CEL (66 – 126), IND (>126) and CEL (43 - 65). These 66 rules are further organized into 11 different subgroups based on 11 MODY genes.

RESULT ANALYSIS

The results obtained from the APRIORI-SD are interpreted from the literature. The association among all the genes in the subgroups is analyzed. Some association occur more than one times but with different frequencies, like association rule INS CEL occur both in rule 1 and rule 2 with frequencies INS (> 126) and CEL (66 – 126), IND (>126) and CEL (43 - 65). These rules are considered only once while performing analysis of obtained results. Although all these genes are responsible for the same type of disease MODY, hence it is obvious that they must have some kind of association with each other. And all of these 11 reported genes are involved in the development of B-cells present in the pancreas. A vivid association in the reported MODY genes has not been described yet. Even in the concerned metabolic pathway, only first six MODY genes have been mentioned. The discovery of subgroups with their underlying association rules can help in discovery of regulation among 11 MODY genes resulting in developing adaptations in the metabolic pathway of the MODY.

CONCLUSION & FUTURE WORK

In contrast to European population, Pakistani population has stable communities, large sib-ships and a high consanguinity rate. Marriages within families are strongly favored due to the existence of various linguistic and racial levels, cultural and economic differences.

In this research, a large inbred Pakistani family was ascertained from Punjab, Pakistan. This family spans four generations and 22 individuals affected with disease named Maturity-onset diabetes of the young (MODY). The objective of this research was to identify new genes and mutation that are responsible for the cause of this disease. We reasoned that this large family may provide an excellent opportunity to localize the new genes or mutation of MODY to get to know the casual patho-mechanisms of this disease. For this purpose we use the exclusion study and genetic screening techniques. In order to infer the correct lineage and inheritance pattern, an extended pedigree was drawn with the help of senior family members. As a first step the family was evaluated for the possibility that the phenotype is linked to all known loci for MODY. Therefore the linkage of the most common MODY genes, HNF4α (MODY 1), GCK (MODY 2), HNF1α (MODY 3) to the known loci of MODY by using Microsatellite Markers was done.

Our results of exclusion study shows that there is no linkage of these known loci to this family that is clinically diagnosed with MODY. Our second objective after performing exclusion study or linkage analysis was to discover new genes or mutation other than known eleven genes or mutations which have been reported so far. For this purpose we employ the Single Strand Conformation polymorphism. Our findings highlighted that those unidentified MODY genes that facilitate to cause this form of diabetes in European population may play a central role for diabetes characterized by autosomal dominant transmission in Pakistani population. The results of linkage

analysis/Genetic screening show that those unidentified MODY genes that facilitate to cause this form of diabetes in European population may play a central role for diabetes characterized by autosomal dominant transmission in Pakistani population.

This study concludes that the examined family which is clinically diagnosed with MODY may be due to mutation in other MODY genes except the genes that are screened in the present research study for this family. As the results of exclusion study indicates that there is no linkage to the known loci of MODY which are used for linkage in this study. Similarly genetic screening results show that no mutation is indicated in this examined family in MODY genes. There may be some environmental factors that are involved in causing this disease in this family or may be this disease is due to mutation in other reported MODY genes that are not screened in this study.

In future we can do research study by using the same techniques or methodology to the rest of the MODY genes on different families that are diagnosed as MODY patients.

In the second phase of this research we implemented a subgroup discovery algorithm for the patterns discovery among MODY genes. For this purpose, we study the research work of Lavrac, who proposed APRIORI-SD subgroup discovery algorithm. We experimentally evaluate our approach on data sets of MODY genes. The total 580 rules are generated by APRIORI-C with the minimum support set to 0.42% and no confidence threshold. These obtained rules are then ordered according to weighted relative accuracy by applying APRIORI-SD approach. The results obtained from the APRIORI-SD are interpreted from the literature. The associations among all the genes in the subgroups are analyzed. The obtained results demonstrate that APRIORI-SD is an aggressive loom towards the task of subgroup discovery. The generated rules are also analyzed through literature review. And it is revealed from this analysis that all the obtained are true and show association with each other.

This research will help to modify the metabolic pathway of maturity onset diabetes of the young as the current pathway show the linkage between just first six MODY genes. Pattern discovery is an emerging field of Bioinformatics and there remains much to be learned from the association among different biological objects. Hence, future work needs to be done to compare the APRIORI-SD with other subgroup discovery algorithm on the same data set or the data set related to different biological objects. Additional work can also be performed by applying this technique on the KEGG pathway of Metabolic Pathway of MODY.

REFERENCES

1. Vaxillaire M and Froguel P. **Monogenic Diabetes in the Young, Pharmacogenetics and Relevance to Multifactorial Forms of Type 2 Diabetes.** Endocrine Reviews 2008; 29(3):254–264.

2. Tattersall RB, Fajans SS, and Arbor A. **A difference between the inheritance of classical juvenile-onset and maturity-onset type diabetes of young people.** Diabetes, 1975; 24(1), 44-53.

3. Gloyn AL, Ellard S, Shepherd M, Howell RT, Parry EM, Jefferson A. and Hattersley AT. **Maturity-onset diabetes of the young caused by a balanced translocation where the 20q12 break point results in disruption upstream of the coding region of hepatocyte nuclear factor-4α (HNF4A) gene.** Diabetes, 2002; 51(7), 2329-2333.

4. Owen K, and Hattersley AT. **Maturity-onset diabetes of the young: from clinical description to molecular genetic characterization.** Best practice & research Clinical endocrinology and metabolism, 2001; 15(3), 309-323.

5. Costa A, Bescos M, Velho G, Chevre J, Vidal J, Sesmilo G, and Conget I. **Genetic and clinical characterisation of maturity-onset diabetes of the young in Spanish families.** European journal of endocrinology, 2000;

142(4), 380-386

6. Vaxillaire M and Froguel P. **Genetic basis of maturity-onset diabetes of the young.** Endocrinol Metab Clin North Am 2006; 35:371–384.

7. Fajans SS. **Maturity-onset diabetes of the young (MODY).** Diabetes Metab. 1989; 5: 579-606.

8. Froguel PH, Vaxillaire M, Sun F, Velho G, Zouali H, Butel MO, and Cohen D. **Close linkage of glucokinase locus on chromosome 7p to early-onset non-insulin-dependent diabetes mellitus.** Nature, 1992; 356(6365), 162-164.

9. Bennett RL, Steinhaus KA, Uhrich SB, O'Sullivan CK, Resta RG, Lochner-Doyle D, and Hamanishi J. **Recommendations for standardized human pedigree nomenclature.** Journal of Genetic Counseling, 1995; 4(4), 267-279.

10. Agrawal R, Imieliński T, and Swami A. **Mining association rules between sets of items in large databases.** In ACM SIGMOD Record 1993; (Vol. 22, No. 2, pp. 207-216). ACM.

11. Lavra N, Cestnik B, Gamberger D, and Flach P. **Decision support through subgroup discovery: Three case studies and the lessons learned.** Machine Learning, 2004; 57(1), 115-143.

12. Kavsek B and Lavrac N. **APRIORI-SD: adapting association rule learning to subgroup discovery.** Appl Artif Intell 2006; 20:543–583.

13. Mueller M, Rosales R, Steck H, Krishnan S, Rao B, and Kramer S. **Subgroup discovery for test selection: a novel approach and its application to breast cancer diagnosis.** 2009; Advances in Intelligent Data Analy.

AUTHOR(S):

1. **MS. ATTIYA KANWAL**
   Department of Bioinformatics
   Muhammad Ali Jinnah University, Islamabad, Pakistan.
2. **MS. SAHAR FAZAL**
   Department of Bioinformatics
   Muhammad Ali Jinnah University, Islamabad, Pakistan.
3. **MR. SOHAIL ASGHAR**
   Director/Assoc. Prof. Univ.
   Institute of IT PMAS-Arid Agriculture University,
   Rawalpindi, Pakistan.
4. ***Mr. Muhammad Naeem***
   Department of Computer Sciences,
   Muhammad Ali Jinnah University, Islamabad, Pakistan.

Correspondence Address:
MS. ATTIYA KANWAL
Department of Bioinformatics
Muhammad Ali Jinnah University, Islamabad, Pakistan.
attiya_kanwal@yahoo.com
sohail.asg@gmail.com

Article received on: 20/12/2012
Accepted for Publication: 28/03/2013
Received after proof reading: 05/06/2013

It is better to fail in originality than to succeed in imitation.

Herman Melville