REVIEW PROF-2166

SUBGROUP DISCOVERY OF THE MODY GENES;
FROM TEXT DOCUMENTS

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

ABSTRACT…..Background: The pandemic of metabolic disorders is accelerating in the urbanized world posing huge burden to health and economy. The key pioneer to most of the metabolic disorders is Diabetes Mellitus. A newly discovered form of diabetes is Maturity Onset Diabetes of the Young (MODY). MODY is a monogenic form of diabetes. It is inherited as autosomal dominant disorder. Till to date 11 different MODY genes have been reported. **Objective:** This study aims to discover subgroups from the biological text documents related to these genes in public domain database. **Data Source:** The data set was obtained from PubMed. **Period:** September-December, 2011. **Materials and Methodology:** APRIORI-SD subgroup discovery algorithm is used for the task of discovering subgroups. A well known association rule learning algorithm APRIORI is first modified into classification rule learning algorithm APRIORI-C. APRIORI-C algorithm generates the rule from the discretized dataset with the minimum support set to 0.42% with no confidence threshold. Total 580 rules are generated at the given support. APRIOIR-C is further modified by making adaptation into APRIORI-SD. **Results:** Experimental results demonstrate that APRIORI discovers the substantially smaller rule sets; each rule has higher support and significance. The rules that are obtained by APRIORI-C are ordered by weighted relative accuracy. **Conclusion:** Only first 66 rules are ordered as they cover the relation between all the 11 MODY genes with each other. These 66 rules are further organized into 11 different subgroups. The evaluation of obtained results from literature shows that APRIORI-SD is a competitive subgroup discovery algorithm. All the association among genes proved to be true.

Key words: Data mining, MODY, Subgroup Discovery.

Article Citation
Kanwal A, Fazal S, Asghar S, Naeem M. Subgroup discovery of the MODY genes from text documents. Professional Med J 2013;20(5): 644-652.

INTRODUCTION

Rule learning is a well known research methodology for discovering interesting relations between variables in large databases. It is commonly used in the perspective of classification rule learning and association rule learning. Classifications rule learning is a form of predictive induction whose aim is to build a set of rules to be used for classification[1,2]. On the other hand, association rule learning is a form of descriptive induction. It aims at discovering the individual rules which identify remarkable patterns in data[3]. Normally rule learning algorithms were proposed for classification rule but with the passage of time and due to certain requirements like subgroup discovery which is an intermediate form of descriptive and predictive rule learning and other approaches to non-classification induction, association rule learning recently gained much attention in the area of machine learning.

This paper considers the task of subgroup discovery which is defined as follows: To determine the subsets of the population whose class distribution is drastically dissimilar from the general distribution. It discovers motivating relationships in a given data set with reverence to definite assets which is of great significance to the target variable[4]. The patterns extorted are generally signified in the form of rules and called subgroups[5].

Rule learning may be an appropriate approach for solving the task of discovering subgroups but standard propositional and relational rule learning are not found to be apposite for this task. In 2003, Lavrac investigates how to acclamatize the rules which determine the subgroups being sufficiently distinctive for detecting most of the target population[6]. For this purpose, they presented a propositional approach to relational subgroup discovery (RSD)[6]. Relational subgroup discovery (RSD) algorithm is an upgrade

version of CN2-SD algorithm in which CN2 classification rule learner was adapted to subgroup discovery. In this paper, we adapted association rule learning (descriptive induction) for the task of subgroup discovery while following some of the guideline[7]. This subgroup discovery algorithm was developed by first making adaptation in association rule learning algorithm APRIORI into classification rule learner APRIORI-C. APRIORI-C is promoted by a novel post-processing method using example weighting incorporated into the covering algorithm and into the modified weighted relative accuracy measure of rule quality, probabilistic classification scheme. These modifications make the APRIORI-C appropriate for subgroup discovery.

This paper presents a subgroup discovery algorithm known as APRIORI-SD and its experimental evaluation on the selected data set related to Maturity-onset diabetes of the young (MODY). MODY is a monogenic form of diabetes mellitus inherited in autosomal dominant mode[8,9]. It is an ancestral form of early-onset type2 diabetes and is primarily an outcome of impaired Beta-Cells of pancreas[8,9]. The evaluation measures used in the experimental evaluation to verify the quality of discovered subgroups are coverage and support. APRIORI-SD produces substantially smaller rule sets with higher coverage and significance which are important for subgroup discovery.

METHODOLOGY

Association Rule Learning

In recent years, with modern technical progress in processing power and storage capability, scientists and researchers have been able to produce huge mass of biological data. Association rule mining has received lot of attention in the aspect of data mining. Association rule learning is one of the approaches of machine learning in which interesting relations between variables are discovered. APRIORI is the best known algorithm to mine association rules[10,11]. It uses a breadth-first search strategy to counting the support of item sets and uses a candidate generation function which exploits the downward closure property of support. 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 $X \longrightarrow Y$. 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 $X \longrightarrow Y$ 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 $p(X \cup Y)$[1].

An association rule $X \longrightarrow Y$ 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, $p(X \mid Y)$.

That is:

$$s(X \longrightarrow Y) = p(X \cup Y) \quad \text{(Eq 1)}$$

$$c(X \longrightarrow Y) = p(Y \mid X) = \frac{s(X \cup Y)}{s(X)} \quad \text{(Eq 2)}$$

APRIORI-C

This section presents the APRIORI-C algorithm which is developed by making adaptations in APRIORI algorithm for the purpose of classification[12]. APRIORI algorithm possess the main advantages of decreased memory consumption and time complexity is upgraded into APRIORI-C algorithm to deal the classification problems by further decreasing its time complexity by feature subset selection, and improving the understandability of results by rule post processing.

APRIORI-C is developed by implementing the following steps:

Discretization of continuous attributes

Many data mining algorithm including classification rules require discretization when predictor attributes are continuous. Discretization refers to the process of converting the values of a continuous variable in two or more bins, the boundaries of which are referred to as split points.

Binarization of discrete attributes

Transforming both continuous and discrete attributes into one or more binary attribute is known as binarization.

Data preprocessing

Data preprocessing is performed through feature subset selection. Data preprocessing by feature subset selection is one of the ways to reduce the dimensionality of the data.

Optimization of the APRIORI-C algorithm

APRIORI-C involve the optimization to be able to better adapt to classification purpose by taking in consideration only those rules that 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 the APRIORI-C selects the best rules by selecting the rules that have the 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 and the 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.

Aprioir-SD

The pseudo-code of the APRIORI-SD which was followed is given below:[24]

ALGORITHM APRIORI_SD

(Examples; Classes; minSup; minConf ; k)
Ruleset = APRIORI _ C(Examples; Classes; minSup; minConf) set all example weights of Examples to 1
Majority = the majority class in Examples
Resultset = {} repeat
BestRule = rule with the highest weighted relative accuracy value in Ruleset
Resultset = Resultset U BestRule
Ruleset = Ruleset \BestRule decrease the weights of examples covered by BestRule (using the example weighting scheme) remove from Examples the examples covered more than k-times
until Examples = {}or Ruleset = {}
return Resultset = Resultset U "true Majority"

Weighting Scheme

APRIORI-SD uses the example weighting schemes for the task of subgroup discovery. This weighting scheme is used in the post processing step of selecting best rules.

The post processing step of APRIORI-SD to select the best rules is as follows:

The algorithm first sort the rules in a way the rules having higher support to the rules that have less support (best to worst) in the terms of weighted

relative accuracy measure. Decrease the weight of the examples that are covered by this best rule until all the examples have been covered and there are no more rules left.

Weighting schemes to select the Best rule

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 of this, each time a rule is selected, the algorithm stores with each example a "count" variable that tells with 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 $w(e_j,i)=\frac{1}{i+1}$ (24).

Initial weights of all examples are same $w(e_j,0)=1$. while in the following iterations, the contributions of examples are inversely proportional to their coverage by already selected rules. This will decrease the weights of the examples that are covered by one or more selected rules and the uncovered examples get the more chance to be covered in the following iterations as their weights have not been decreased. When the weights of the covered examples will fall below a given threshold, these examples will be removed from the data set.

Rule Evaluation Measure: Weighted Relative Accuracy

Weighted relative accuracy measure (WRAcc) is a new rule evaluation measure that is recently proposed in iteilp99-lavrac-flach-zupan. It is equivalent to the novelty measure used in the descriptive induction and takes into account the improvement of the accuracy relative to the default rule in contrast to the widely accepted rule evaluation measures in terms of accuracy. 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
cvered by the rule X —› Y
n(Y): Number of examples of class Y
n(X.Y): Number of true positive examples

For the corresponding probabilities *ñ(X), ñ(X.Y)* notations are used. In terms of association rule learning, the accuracy of the rule is defined as

$$Acc(X \longrightarrow Y)=Conf(X \longrightarrow Y)=ñ(Y \mid X)=\frac{ñ(X.Y)}{ñ(X)} \quad \text{(Eq 3)}$$

Weighted Relative Accuracy is defined as follows[24]:

$$WRAcc(X \longrightarrow Y)=ñ(X).(ñ(Y|X)-ñ(Y)) \quad \text{(Eq 4)}$$

This definition of weighted Relative Accuracy has two parts one is p(X) and the second is (ñ(YIX)-ñ(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. It is easy to get the high relative accuracy for the rule with low generality. Here the term generality refers to the "weight". Weighted relative accuracy trades off the generality of the rule p(X) and the relative accuracy (ñ(YIX)-ñ(Y)). The probabilities can be estimated by the relative frequencies.

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. This modified weighted relative accuracy measure is defined as follows:

$$wWRAcc(X \longrightarrow Y) = \frac{n'(X)}{N'}\left(\frac{n'(XY)}{n'(X)}-\frac{n'(Y)}{N'}\right) \quad \text{(Eq 5)}$$

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.

$\frac{n'(Y)}{N'}$ in wWRAcc is replaced with $\frac{n'(Y)}{N'}$ it will force the wWRAcc' to reflect the improvement of the rule's accuracy with respect to the accuracy of the default rule
true—› Y

So the improved definition of weighted relative accuracy measure is as follows[24].
wWRAcc (X —› Y)

$$= \frac{n'(X)}{N'}\left(\frac{n'(XY)}{n'(X)} - \frac{n'(Y)}{N'}\right) \quad \text{(Eq6)}$$

Evaluation on Data Set of MODY Genes

We experimentally evaluate our approach of APRIORI-SD on the data set of 11 reported MODY genes. MODY is the term that relies on the old classification of diabetes into juvenile-onset and maturity-onset diabetes[18]. First described MODY in 1974, after taking into account a group of young people with diabetes who were treated without insulin 2 years or more after diagnosis. 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 type2 diabetes. It is a monogenic form of diabetes mellitus inherited in autosomal dominant mode[8,9]. It is primarily an outcome of impaired Beta-Cells of pancreas[8,9]. MODY is not a single entity but represents genetic, metabolic, and clinical heterogeneity[19]. It generally develops in middle age, and mainly coupled with primarily scantiness of insulin secretion[20].

Till 2009 eight discrete MODY genes have been acknowledged[21]. These are the genes including HNF4A, encoding hepatocyte nuclear factor 4 Alpha[22], GCK, encoding glucokinase[23] HNF1A, encoding hepatocyte nuclear factor 1 Alpha[24], IPF1, encoding insulin promoter factor 1[21], HN F1B, encoding hepatocyte nuclear factor 1 Beta[22], NEUROD1, encoding neurogenic differentiation 1[23], KLF11, encoding for kruppel-like factor 11[25] and CEL, encoding carboxyl-ester lipase[26]. Now four further genes have been exposed that cause MODY. These genes are PAX4, encoding, Paired Domain gene 4 (OMIM 612225), IPF1, encoding, Insulin Promoter Factor 1 (OMIM 606392), INS, encoding, Insulin (OMIM 176730) and BLK, encoding, Tyrosine kinase, B-Lymphocyte. (OMIM 191305).

The true relative prevalence of the eleven distinct MODY subtypes is unknown and varies substantially in different populations[26,27,28]. Mutations in the genes encoding HNF1A and GCK are by far the most prevalent. Mutations in GCK (MODY 2) account for 7-41%[26]. Whereas mutations in TCF1 (MODY 3) may account for 11-63%[28] of mutations in subjects with clinically diagnosed MODY. Mutations in HNF4A (MODY1) are less frequent and may account 2-5% of subjects with MODY[28]. The prevalence of MODY patients with mutations in TCF2 (MODY 5) is considered very rare. It may comprise up to approximately 2 %. Mutations in IPF1 (MODY 4), NEUROD1 (MODY 6), KLF11 (MODY 7), CEL (MODY 8) and PAX4 (MODY 9) are also very rare and have been identified only 1% McCarthy and[23]. The prevalence of MODY patients with mutations in INS (MODY 10) and BLK (MODY 11) are also considered only 1%[23].

Steps Followed for the Subgroup discovery:

Acquiring the documents

All the documents related to 11 reported MODY genes are acquired from the public domain database PUBMED. The documents are downloaded by sending the gene name as a query. 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.

Tf-idf

The frequency of each term present in all these 11 documents is computed by using Tf-idf algorithm. Term frequency-inverse document frequency weight is a statistical measure that evaluates the importance of word to a document in a collection. The importance is proportional to the number of times that word is present in a document and increase as the number increases. However it is offset by the frequency of the word in a collection.

Preprocessing of Data

The obtained results from Tf-idf algorithm are further processed so that they will be valid input for subgroup discovery algorithm. Firstly the dataset is reduced by removing all records where the frequency in only 1 column was greater than zero. The column that contains the terms would be removed as they are not relevant to our purpose.

Discretization

As the results of Tf-idf are in continuous form so we perform the discretization of this result set as APRIORI-SD handles only attributes in discrete form. The frequencies of all the terms in each document are grouped in such a way that each group contains at least 1000 values. In this way we get 20 groups. After grouping the reduced data set is discretized by replacing the value of each term as per the grouping.

Generation of Rules with APRIORI-C

APRIORI-C algorithm is used to generate the rule from the discretized dataset with the minimum support set to 0.42% and no confidence threshold. The total 580 rules are generated at the given support.

Generation of Rule with APRIORI-SD

The rules that are obtained by APRIORI-C are ordered by weighted relative accuracy. Only first 66 rules are extracted as they cover the relation between all the 11 MODY genes with each other.

RESULTS

Evaluation on selected data sets

We experimentally evaluate our approach on data sets of MODY genes. This data set includes the records of all the documents that are present in the PUBMED related to Maturity onset diabetes of the young. PubMed comprises more than 20 million citations for biomedical literature from MEDLINE, life science journals, and online books. The task of the challenge is to produce the subgroups to predict the association among different 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. Only first 66 rules are ordered as they cover the relation between all the 11 MODY genes with each other. These rules are further organized into 11 different subgroups.

Interpretation of Results

The results obtained from the APRIORI-SD are interpreted from the literature. The association among all the genes in the subgroups are 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 doing analysis of obtained results. Although all these genes are responsible for the same type of disease, MODY so 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. The clear evaluation of all of these rules are compiled into 22 tables. The reader is encouraged to contact author for these detailed information.

DISCUSSION

Maturity onset diabetes of the young is a monogenic form of diabetes mellitus inherited in autosomal dominant mode. It is primarily an outcome of impaired

Beta-Cells of pancreas. Till 2009 eight discrete MODY genes have been acknowledged. These are the genes including six transcription factors HNF4A, HNF1A, IPF1, HNF1B, PAX4 and NEUROD1. Now four further genes have been exposed that cause MODY. These include KLF11, CEL, INS, and BLK gene. All these MODY genes form a network and play an important role in B-cells function, survival and pancreatic development. The transcription factors mentioned above controls the expressions of one another as well as number of other B-cell genes.

Mutations in all these transcriptions factors escort to abnormal expression of genes that are concerned with pancreatic islet development and metabolism. The linkage between first six MODY genes is also shown in the metabolic pathway of Maturity onset diabetes of the young.

CONCLUSIONS

In the present research, we have modified the classification rule learning algorithm APRIORI-C to subgroup discovery that results in an appropriate subgroup discovery algorithm APRIORI-SD. The proposed algorithm is implemented in C#. The main amendment of the APRIORI C algorithm that formulates it suitable for the task of discovering subgroups is the involvement of example weighting schemes in the implementation of APRIORI-SD. The performance of these example weighting schemes was analyze by using the ROC, the scheme that is offered by Flach and Furukranz. The influence for variation of WRacc to wWRAcc` was also presented. The analysis of wWRAcc` by weighting just the covered positive examples and that covered the negative examples was also presented by using the ROC space. It is concluded that the first scheme is more focused, that selects smaller and highly accurate subgroups and the second scheme selects the rule which are larger and less accurate. The proposed approach was evaluated on the data sets of MODY genes that show 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.

REFERENCES

1. 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.

2. Lavrac N, Zelezny F, Flach PA: **RSD: relational subgroup discovery through first-order feature construction. In: Proceedings of the 12th international conference inductive logic programming.** 2003, vol 2583. Springer, LNCS, 149–165.

3. Lavrac N, Cestnik B, Gamberger D, Flach PA: **Decision support through subgroup discovery: three case studies and the lessons learned.** Mach Learn. 2004, 57(1–2):115–143

4. Anna L.Gloyn, Sian Ellard, Maggie Shepherd, Rodney T. Howell, Elizabeth M.Parry, Andrew Jefferson, Elaine R. Levy and Andrew T. Hattersley: **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-4A Gene.** DIABETES. 2002, 51: 2329-2333.

5. Owen K, Httersley A: **Maturity-onset Diabetes of the young: from clinical description to molecular genetic characterization.** In Best Practice and Research Clinical Endocrinology and Metabolism. 2001, 15: 309-323.

6. Agrawal, R., T. Imielinski, and R. Srikant: **Mining association rules between sets of items in large databases.** In Proceedings of ACM SIGMOD Conference on Management of Data. 1993, 207–216.

7. Agrawal, R. and R. Srikant: **Fast algorithms for mining association rules.** In Proceedings of the 20th International Conference on Very Large Databases. 1994, 487–499.

8. Jovanoski, V. and N. Lavrac˘: **Classification rule learning with APRIORI-C.** In Progress in Artificial Intelligence. Proceedings of the 10th Portuguese Conference on Artificial Intelligence. 2001, 44–51.

9. Rivest, R. L.0: **Learning decision lists.** Machine Learning. 1987, 2(3):229–246.

10. Provost, F. J. and T. Fawcett.: **Robust classification for imprecise environments.** Machine Learning. 2001, 42(3):203–231.

11. Flach, P. A.: **The geometry of ROC space: Understanding machine learning metrics through ROC isometrics.** In Proceedings of the 20th International Conference on Machine Learning. 2003, 194–201.

12. Fu¨rnkranz, J. and Flach P. A., **An analysis of rule evaluation metrics.** In Proceedings of the 20th International Conference on Machine Learning. 2003, 202–209.

13. Klo¨sgen, W.: EXPLORA: **A multipattern and multistrategy discovery assistant.** In Advance in Knowledge Discovery and Data Mining. 1996, 249–271.

14. Tattersall RB, Fajans SS: **A difference between the inheritance of classical juvenile-onset and maturity-onset type diabetes of young people.** Diabetes. 1975, 24:44-53.

15. A Costa, M Bescos, G Velho, J Chevre, J Vidal, G Sesmilo, C Bellanne-Chantelot, P Froguel, R Casamitjana, F Rivera- Fillat, R Gomis, and I Conget: **Genetic and clinical characterization of maturity-onset diabetes of the young in Spanish families.** European Journal of Endocrinology. 2000, 142: 380–386.

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

17. Maciej Borowiec, Chong W. Liew, Ryan Thompson, Watip Boonyasrisawat, Jiang Hu, Wojciech M. Mlynarski, Ilham El Khattabi, Sung-Hoon Kim, Lorella Marselli, Stephen S. Rich, Andrzej S. Krolewski, Susan Bonner-Weir, Arun Sharma, Michele Sale, Josyf C. Mychaleckyj, Rohit N. Kulkarni and Alessandro Doria: **Mutations at the BLK locus linked to maturity onset diabetes of the young and B-cell dysfunction.** PNAS. 2009, 106:

18. Yamagata K, Furuta H, Oda N, Kaisaki PJ, Menzel S, Cox NJ, et al.: **Mutations in the hepatocyte nuclear factor-4alpha gene in maturity-onset diabetes of the young (MODY1).** Nature. 1996, 384: 458–60.

19. Froguel P, Zouali H, Vionnet N, Velho G, Vaxillaire M, Sun F, Lesage S, Stoffel M, Takeda J, Passa P, et al.: **Familial hyperglycemia due to mutations in glucokinase. Definition of a subtype of diabetes mellitus.** N Engl J Med. 1993, 328: 697–702.

20. Yamagata K, Oda N, Kaisaki PJ, Menzel S, Furuta H, Vaxillaire M, Southam L, Cox RD, Lathrop GM, Boriraj VV, Chen X, Cox NJ, Oda Y, Yano H, Le Beau MM, Yamada S, Nishigori H, Takeda J, Fajans SS, Hattersley AT, Iwasaki N, Hansen T, Pedersen O, Polonsky KS, Bell GI, et al: **Mutations in the hepatocyte nuclear factor-1 gene in maturity-onset diabetes of the young (MODY3).** Nature. 1996, 384:455–458, 1996.

21. Stoffers DA, Ferrer J, Clarke WL, Habener JF: **Early-onset type-II diabetes mellitus (MODY4) linked to IPF1.** Nat Genet. 1997, 17:138–139.

22. Horikawa Y, et al.: **Mutation in hepatocyte nuclear factor-1 beta gene (TCF2) associated with MODY.** Nat Genet. 1997, 17:384–385.

23. McCarthy MI, Hattersley AT. **Novel Insights Arising From the Definition of Genes fro Monogenic and Type 2 Diabetes.** DIABETES. 2008;57: 2889-98.

24. Kavsek B and Lavra N. **Apriori-Sd: Adapting Association Rule Learning To Subgroup Discovery.** Applied Artificial Intelligence. 2006;20:543–583.

25. Neve B, Fernandez-Zapico ME, Ashkenazi-Katalan V, Dina C, Hamid YH, Joly E, et al: **Role of transcription factor KLF11 and its diabetes-associated gene**

variants in pancreatic beta cell function. Proc Natl Acad Sci U S A. 2005, 102: 4807–12.

26. Ræder H., Johansson S., Pål I.H, Haldorsen I. S., Mas E., Sbarra V., Nermoen I., Eide S. Å., Grevle L., Bjørkhaug L., Sagen J. V., Aksnes L., Søvik O., Lombardo D., Molven A. and Njølstad P. R., **Mutations in the CEL VNTR cause a syndrome of diabetes and pancreatic exocrine dysfunction.** Nat Genet. 2006, 38: 54–62.

27. Barrio R, Bellanne-Chantelot C, Moreno JC, Morel V, Calle H, Alonso M, Mustieles C: **Nine novel mutations in maturity-onset diabetes of the young (MODY) candidate genes in 22 Spanish families.** J Clin Endocrinol Metabolism. 2002, 87: 2532–2539.

28. Pruhova S, Ek J, Lebl J, Sumnik Z, Saudek F, Andel M, Pedersen O, Hansen T. **Genetic epidemiology of MODY in the Czech Republic: new mutations in the MODY genes HNF-4 alpha, GCK and HNF-1 alpha.** Diabetologia. 2003;46:291–295.

29. James D. Johnson: **Pancreatic Beta-cell Apoptosis in Maturity Onset Diabetes of the Young.** Canadian journal of Diabetes. 2007, 31(1):67-74.

AUTHOR(S):

1. **MISS. ATTIYA KANWAL**
   Department of Bioinformatics
   Muhammad Ali Jinnah University, Islamabad Pakistan.
2. **MISS. 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:
Dr. Sohail Asghar
Director/Assoc. Prof. Univ.
Institute of IT PMAS-Arid Agriculture University,
Rawalpindi Pakistan.
sohail.asghar@uaar.edu.pk

Article received on: 20/12/2012
Accepted for Publication: 25/04/2013
Received after proof reading: 21/09/2013

"Success is not final, failure is not fatal: it is the courage to continue that counts."

Winston Churchill