

IDENTIFYING GENETIC MUTATION RARE GENETIC DISORDER BY ANALYZING CHARACTERISTICS OF GENOTYPE-PHENOTYPE BY IMPLEMENTING APRIORI ALGORITHM

Bipin nair B ^{J1}, Ratheesh A ^{A2}, Koushik K S ^{S3}

Department of Computer Sciences

Amrita School of Arts and Sciences

Amrita Vishwa Vidyapeetham, Mysore Campus, Karnataka, India

Abstract: *An The purpose of the paper —Identifying genetic mutation rare genetic disorder by analyzing characteristics of genotype-phenotype by implementing data mining algorithm's laims at reducing the complexities involved in determining the mutation occurred in human DNA sequence. The Data mining algorithm is chosen from specific areas like association. The algorithms are grouped in such way that it can be adapt to user requirements. The visualization provided for the mined outputs are represented in graph .The available data mining tools require expert users to carry out experiments. The scope of this relies in the area of visualization of the mined data. The visualizations can be made interactive such that the outputs can be easily interpreted by the user with less effort.*

Keywords: *DNA; mutation; genotype; phenotype; genetic;*

I. Introduction

Data Mining is the process of analyzing data from different prospective and summarizing it into useful information .Data Mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it and summarize the relationship identified. Technically Data Mining is the process of finding correlations or patterns among dozens of fields in a large relational database. Data mining often involves the analysis of data stored in a data warehouse. One of the major data mining techniques is association. It is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items of genetics disease symptoms in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. Manually analyzing the genetic variation is very difficult. Since many genetics shares similar symptoms, it is very difficult to identify the genetic disorder accurately. To analyze genetic variation in medical labs it requires amino acid analyzer which is very expensive.

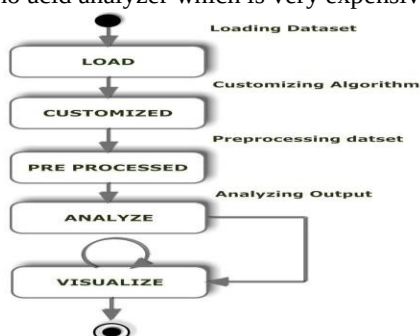


Figure 1 A sample flow diagram for working of the algorithm

II. Overview

This paper deals with the analysis of a mutation is a change of the nucleotide sequence of the genome of an organism, virus, or extra chromosomal genetic element. Mutations result from unrepaired damage to DNA or to RNA genomes (typically caused by radiation or chemical mutagens); Mutation can result in several different types of change in sequences. Manually analyzing the genetic variation is very difficult. Since many genetics shares similar symptoms, it is very difficult to identify the genetic disorder accurately. To analyze genetic variation in medical labs it requires amino acid analyzer which is very expensive. So, implementing the data mining algorithm in such a way that it can handle the genetic data sets and to visualize the result accurately in an understandable manner

III. Problem Statement

Identifying genetic mutation rare genetic disorder by analyzing characteristics of genotype-phenotype by implementing apriori algorithm. There are many data mining tools available today. There are excellent tools that are available. But they lack in some points. They are,

- Sometimes expert support is required to use these tools.
- The dataset should have a common format. If not, it will ask for the conversion to its own format. The naive user finds difficulty to do this step.
- They provide limited user interaction.
- The processed output of these tools is not completely understandable.
- Manually analyzing the genetic variation is very difficult.
- Since many genetics shares similar symptoms.
- It is very difficult to identify the genetic disorder accurately.
- To analyze genetic variation in medical labs it requires amino acid analyzer which is very expensive.

IV. Problem Formulation

The proposed paper helps the user to work with the major data mining algorithms in an easy way, and the user is provided with an interactive and effective visualization that helps to make a useful decision. It mainly overcomes the limitations of the existing work in terms of user involvement and visualization.

The Proposed work performs,

The algorithm handles text as well as numeric data.

- Developing an understanding of the application domain, relevant prior knowledge and the goals of the end-user.
- Selecting a dataset on which discovery is to be performed.
- The system will be user friendly; a user will know what is happening to their input data.
- Data Pre-processing: This stage includes operations for dimension reduction (such as feature selection and sampling); data cleansing (such as handling missing values, removal of noise or outliers); and data transformation (such as discretization of numerical attributes and attribute extraction).
- Prediction of genetic disorder undergoes various procedures such as
 - *Uploading dataset*: In this stage the dataset which is to be examined should be uploaded.
 - *Pre-Process*: In this stage elimination of the records whose attribute information are missed and other pre-processing is done.
 - *Analysis*: In this stage the outcomes are clearly analyzed to provide remedial measure and other suggestion like gene therapy required to the patient will be presented.
 - *Visualize*: In this stage visualization of the output is seen.
- The visualization will be easily understandable for a user.
- To analyze genetic variation in medical labs it doesn't requires amino acid analyzer which is very expensive
- Choosing the data mining algorithm. This stage includes selecting the specific method to be used for searching patterns.
- Employing the data mining algorithm.
- Generate the frequent pattern of symptoms.
- Selects algorithm to analyze genetic variation in DNA sequence.
- Evaluating and interpreting the mined patterns

V. Related Work

- Visualization Module
- Request Handler

The request handler consists of a controller class which controls the requests and responses. Controllers provide access to the application behavior which is typically defined by a service interface. Controllers interpret user input and transforms such inputs into a sensible model which will be represented to the user by the view. Controller is a single method that is responsible for handling a request and retrieving an appropriate model and view. The DB Controller controls the relational data store. The details of the user are stored in a local data base. Each Algorithm has tree formatter.

- Input
- Output
- Visualization

A) Apriori Algorithm

- It is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items of genetics disease symptoms in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.

• Apriori algorithm Pseudo code:

Input: A patient data Set Output: Frequent symptoms set Method:

Ck: Candidate itemset of size k

Lk : frequent itemset of size k

L1 = {frequent items};

for (k = 1; Lk !=MAX; k++) do begin

Ck+1 = candidates

generated from Lk;

for each transaction t in database do

increment the count of all candidates in Ck+1 that are contained in t

Lk+1 = candidates in Ck+1 with min_support

End

Return k Lk;

VI. Experiment Result

Figure 2 Taking input

Patient	Name	Age	Gender	Family	Blood c.	Multip.	Head a.	Fatigue	Enlarg.	Stomach	Joint p.	Diarr.	Stomach	Urinal.	Feeding	Stomach	Bleeds	Swells	Parotid	Urinary	Head a.	Head p.	Head c.	Diarr.
1.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	NO	YES	NO	YES	YES	NO	NO	NO
2.0	ANNA	22.0	Female	YES	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	NO	YES	NO	YES	YES	NO	NO	NO
3.0	ANNA	22.0	Female	YES	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	NO	YES	NO	YES	YES	NO	NO	NO
4.0	BERNARD	25.0	Male	NO	NO	NO	NO	YES	YES	NO	YES	NO	NO	NO	NO	YES	YES	NO	NO	YES	NO	YES	YES	YES
5.0	BERNARD	25.0	Male	NO	NO	NO	NO	YES	YES	NO	YES	NO	NO	NO	NO	YES	YES	NO	NO	YES	NO	YES	YES	YES
6.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
7.0	BERNARD	25.0	Male	NO	NO	NO	NO	YES	YES	NO	YES	NO	NO	NO	NO	YES	YES	NO	NO	YES	NO	YES	YES	YES
8.0	ANNA	22.0	Female	NO	NO	NO	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
9.0	ANNA	22.0	Female	NO	NO	NO	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
10.0	ANNA	22.0	Female	NO	NO	NO	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
11.0	ANNA	22.0	Female	YES	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
12.0	ANNA	22.0	Female	YES	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
13.0	BERNARD	25.0	Male	YES	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
14.0	BERNARD	25.0	Male	NO	NO	NO	NO	YES	YES	NO	YES	NO	NO	NO	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
15.0	BERNARD	25.0	Male	NO	NO	NO	NO	YES	YES	NO	YES	NO	NO	NO	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
16.0	BERNARD	25.0	Male	NO	NO	NO	NO	YES	YES	NO	YES	NO	NO	NO	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
17.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
18.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
19.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
20.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
21.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
22.0	BERNARD	25.0	Male	YES	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
23.0	BERNARD	25.0	Male	NO	NO	NO	NO	YES	YES	NO	YES	NO	NO	NO	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
24.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
25.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
26.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
27.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
28.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
29.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
30.0	BERNARD	25.0	Male	YES	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
31.0	BERNARD	25.0	Male	NO	NO	NO	NO	YES	YES	NO	YES	NO	NO	NO	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
32.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
33.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
34.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
35.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
36.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
37.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
38.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
39.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO
40.0	ANNA	22.0	Female	NO	YES	YES	NO	YES	NO	YES	YES	NO	YES	YES	NO	YES	YES	NO	NO	YES	YES	NO	NO	NO

Figure 3 preprocessed data set after avoiding anomalies

Row Count	1	ANNA	Gender: Female
Processed Data	11010101100101100		
Size Of Items (Supported)	6		
Row Count	2	ANNA	Multiple sclerosis
Processed Data	11101010001001100		
Size Of Items (Supported)	12		
Row Count	3	ANNA	Gender: Female
Processed Data	011100100010101001		
Size Of Items (Supported)	11		
Row Count	4	BERNARD	Gender: Male
Processed Data	0000101000010001011		
Size Of Items (Supported)	10		
Row Count	5	ANNA	Gender: Female
Processed Data	100010000100011100		
Size Of Items (Supported)	10		
Row Count	6	ANNA	Inflammatory bowel disease
Processed Data	0101001100111000110		
Size Of Items (Supported)	15		
Row Count	7	BERNARD	Gender: Male
Processed Data	01110100011001101		
Size Of Items (Supported)	12		

Figure 4 Frequent pattern generated using apriori

Row Count	Processed Data	Size Of Items (Supported)
1	11010101100101100	6
2	11101010001001100	12
3	011100100010101001	11
4	0000101000010001011	10
5	100010000100011100	10
6	0101001100111000110	15
7	01110100011001101	12

VII. Literature Survey

Implementation of the tree scanning method to detect associations between genetic haplotypes and quantitative traits, utilizing the evolutionary history of the haplotypes, in samples of unrelated individuals [17].

Sub-sumption,properties,class description are taken into account for data selection before data mining. Each of this is scenario is illustrated [18]

GWAF provides functions to visualize results. We evaluated GWAF using a simulated continuous trait and a binary trait [19].

To develop phylogenetic substitution models to test for associations between evolutionary rate of genotype and phenotype. In this work introduce a new method for rating hybrid rate matrices between genotype and phenotype [20]

The multiple-to-one association rules with stringent FDR level for aerobic, anaerobic, facultative, endospore and Gram-negative phenotype contain significantly larger numbers of COGs than those by pairwise methods [21]

VIII. Conclusion

The paper Analysis, The advantage of the proposed system is to improve the efficiency of the existing system by the reducing the complexities involved in determining the mutation occurred in human DNA sequence. This result has a profound impact in medical sciences. This work will help to identify common symptoms observed in same category of genetic disorder in an efficient way so that it will reduce the through put and in improve efficiency. The proposed idea will help physicians to diagnose the genetic disorder and develop strategies for its therapy.

This work is part of a research paper, so there were time constraints in the implementation of the paper. As the technologies being used for this were changed in order to come up with the best result, all that has been achieved within this time is the final paper.

References

- [1] Data Mining- Introductory and Advanced Topics
- [2] Data Mining: Concepts and Techniques by Jiawei Han
- [3] A Comparative Analysis of Data Mining Tools in Agent Based Systems by Sharon Christa, K. Lakshmi Madhuri, V. Suma
- [4] CSE5230 Tutorial The Na ve ayes lassifier
- [5] The Apriori Algorithm – a Tutorial by Markus Hegland
- [6] DATA MINING: Theory and Practices by Dr.Shyam Divakar and Dr.K.P.Soman.
- [7] Naive Bayes Classifier example - Eric Meisner lecture05-NaiveBayes-2up.pdfAn Improved k-Nearest Neighbor Classification Using Genetic Algorithm ,N. Suguna1, and Dr. K. Thanushkodi2
- [8] International Journal of Computer Trends and Technology (IJCTT) - volume4Issue4 –April 2013 ENHANCED DBSCAN ALGORITHM by Priyamvada Paliwal#1, Meghna Sharma.
- [9] An Introduction to Neural Networks by Vincent Cheung and Kevin Cannons.
- [10] Jiawei Han and Micheline Kamber, University of Illinois at urbana-champaign, Concepts and Techniques, Data mining, second edition(2006), [3]Barandela, R., Sánchez, J.S., García, V., Rangel, E. Strategies for Learning in Class Imbalance Problems. Pattern Recognition 2003, 36(3), pp.849-851
- [11] K.P Soman, Shyam Diwakar, V.Ajay, data mining theory and practice, Amrita Vishwa Vidyapeetham, (2010).
- [12] Data Mining: Theory and Practices by Dr.Shyam Divakar and Dr.K.P.Soman.
- [13] Apriori Algorithm Review for Finals Presentation by SE 157B, Spring Semester 2007 Professor Lee By Gaurang Negandhi.
- [14] PDF-”Using TF-IDF to Determine Word Relevance in Document Queries” by Juan Ramos, Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855 [15] Figure 1 – Schematic drawing made through Rational Rose.
- [16] Figure 2 to 5- Resulting Snapshots obtained by implementing algorithms Naïve Bayes and KNN.
- [17] David Posada,Taylor J. Maxwelland Alan R. TempletonTreeScan: a bioinformatic application to search forgeno-type/phenotype associations using haplotype trees, Variagenics, Inc., 60 Hampshire Street, Cambridge, MA 02139, USA and Department of Biology, WashingtonUniversity, St Louis, MO 63130-4899, USAReceived on November 5, 2004; revised on January 10, 2005; accepted on January 25, 2005Advance Access publication January 28, 2005, Vol. 21 no. 9 2005, pages 2130–2132 doi:10.1093/bioinformatics/bti293
- [18] [18] AdrienCoulet, MalikaSmail-Tabbone, Pascale Benlian, AmedeoNapoliand Marie-Dominique Devignes ,l Ontolo-gy-guided data preparation for discovering genotype-phenotype relationships, Address:KIKA Medical, Paris, F-75012, France,LORIA (UMR 7503 CNRS-INPL-INRIA-Nancy2-UHP),Vandoeuvrelès-Nancy, F- 54506, France andUniversi-té Pierre et Marie Curie - Paris6, INSERM UMRS 538 Biochimie-BiologieMoléculaire,Paris, F-75571, FranceEmail: AdrienCoulet* - adrien.coulet@loria.fr; MalikaSmail-Tabbone - malika.smail@loria.fr; Pascale Benlian - pascale.benlian@sat.ap-hop-paris.fr; Amedeo Napoli - amedeo.napoli@loria.fr; Marie-Dominique Devignes – marie-dominique.devignes@loria.frCorresponding author
- [19] Ming-Huei Chen and Qiong Yang ,l GWAF: an R package for genome-wide association analyses withfamily da-tal, Department of Neurology, Boston University School of Medicine, Boston, MA 118,The National Heart, Lung, lood Institute’s Framingham Heart Study, Framingham, MA 01702 andDepartment of Biostatistics, BostonUni-versity School of Public Health, Boston, MA 02118, USAReceived on September 28, 2009; revised on December 18, 2009; accepted on December 21, 2009Advance Access publication December 29, 2009, Vol. 26 no. 4 2010, pages 580– 581 doi:10.1093/bioinformatics/btp710.
- [20] Timothy D. O’ onnor and Nicholas I. Mundy,l Genotype–phenotype associations: substitution models to detect evolutionary associations between phenotypic variables andgenotypic evolutionary ratel, Department of Zoology, Uni-versity of Cambridge, Cambridge B2 3EJ, UK, Vol. 25 ISMB 2009, pages i94– i100doi:10.1093/bioinformatics/btp231.
- [21] Makio Tamura and PatrikD’haeseleer,l Microbial genotype–phenotype mapping by class association rule miningl, Lawrence Livermore National Laboratory, Computing Applications and Research Department/Chemistry, Materi-als,Earth and Life Sciences Department, Microbial Systems Biology Group, Livermore, CA 94550, USA Received on October 8, 2007; revised on February 28, 2008; accepted on April 26, 2008 Advance Access publication May 8, 2008, Vol. 24 no. 13 2008, pages 1523–1529 doi:10.1093/bioinformatics/btn210

Acknowledgments

We would like to thank the anonymous reviewers for their very helpful comments and suggestions and also we extend our gratitude to Dr.Vikas Modi and family Amrita krupa Hospital Mysore, for helpful discussion.