International Journal of Research in Computer Science
eISSN 2249-8265 Volume 4 Issue 5 (2014) pp. 11-15
www.ijorcs.org, A Unit of White Globe Publications

TYPICAL GENOMIC FRAMEWORK ON DISEASE ANALYSIS

J. Stanly Thomas[1], Dr. N. Rajkumar[2]
[1]*Research Scholar, Dravidian University, A.P. INDIA*
Email: meetstanly@gmail.com
[2]*Professor and HOD, Sri Ramakrishna Engg. College, INDIA*
Email:nrk29@rediffmail.com

Abstract: *The challenging and major role of the doctor in human life is to predict as well as diagnose the disease which has got infected in the human body. This typical genomic framework on disease analysis algorithm is designed to store and drive each and every gene characteristics like shape, weight, location and normal growth culture. Whenever the disease report is feed into this data mining algorithm triggers the similarity test built upon the data mining classification rules. A gene is usually comprised of hundreds of individual nucleotides arranged in a particular order. There are almost an unlimited number of ways that the nucleotides can be ordered and sequenced to form distinct genes. The algorithm delivers the difference between diseased and healthy status shall guide us to conclude the disease severity, stage and its nature. This powerful Typical Genomic Framework on Disease Analysis (TGFDA) algorithm is built to deliver instant result over Very Large Database using density and weight based Clustering Algorithm.*

Keywords: *Association Rules, Classification Rules, Very Large Database with similarity and density and weight based clustering analysis*

I. INTRODUCTION

In the data mining world, lot of new techniques and tracking algorithm has come for our consideration. However, integrating both fast growing and energetic medical science and computer science in a single entity really benefits the human life cycle. The major emphasis given in this algorithm is to access and analyze the gene sets. The similarity search and density based Clustering Analysis has laid the corner stone for the building called TGFDA algorithm. Being an introduction to the TGFDA algorithm, this algorithm is developed in the format of two capsules as shown below. In the first phase of TGFDA, healthy tissue gene attributes like weight, size, location, shape and strength are fed into a database (Original Data) for the aforesaid clustering and similarity analysis.

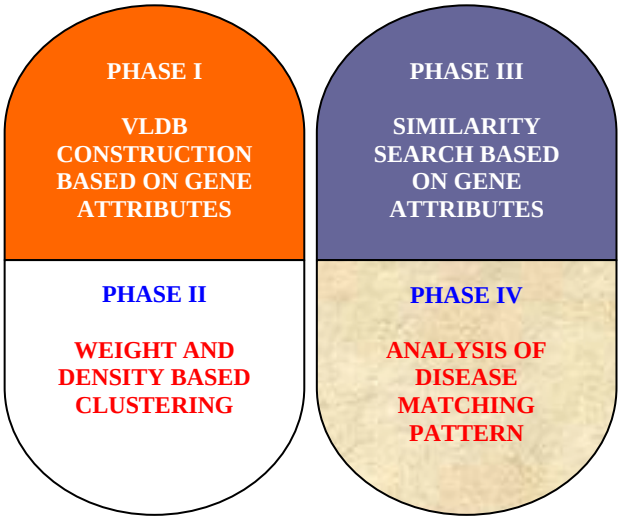

Figure 1: TGFDA Four Phase Capsules

Further indexing of VLDB based on size, creates an easy access to gene attributes. However owing to large volume of data in a data warehouse, it is quite difficult and time consuming task to access a data simply with indexing. Therefore in a second phase weight and density based clustering approach adopted to simplify the access (Boost up the accessing speed). Third phase entirely contributes its service with data analysis by similarity analysis technique based on its weight, location, size, shape and strength compared with healthy tissue status of the gene. Final phase provides the result after the present gene deviation status checked with set of diseases loaded in VLDB.

II.SEMANTIC INTEGERATION DISTRIBUTED GENOME DATABASES

Typical Genomic Framework on Disease Analysis (TGFDA) intakes the gene sets due to the highly distributed, uncontrolled generation and use of a wide variety of DNA data, the semantic integration of such heterogeneous and widely distributed genome databases becomes and important task of systematic and coordinated analysis of DNA databases. TGFDA designed to intake the following attributes of gene sets.

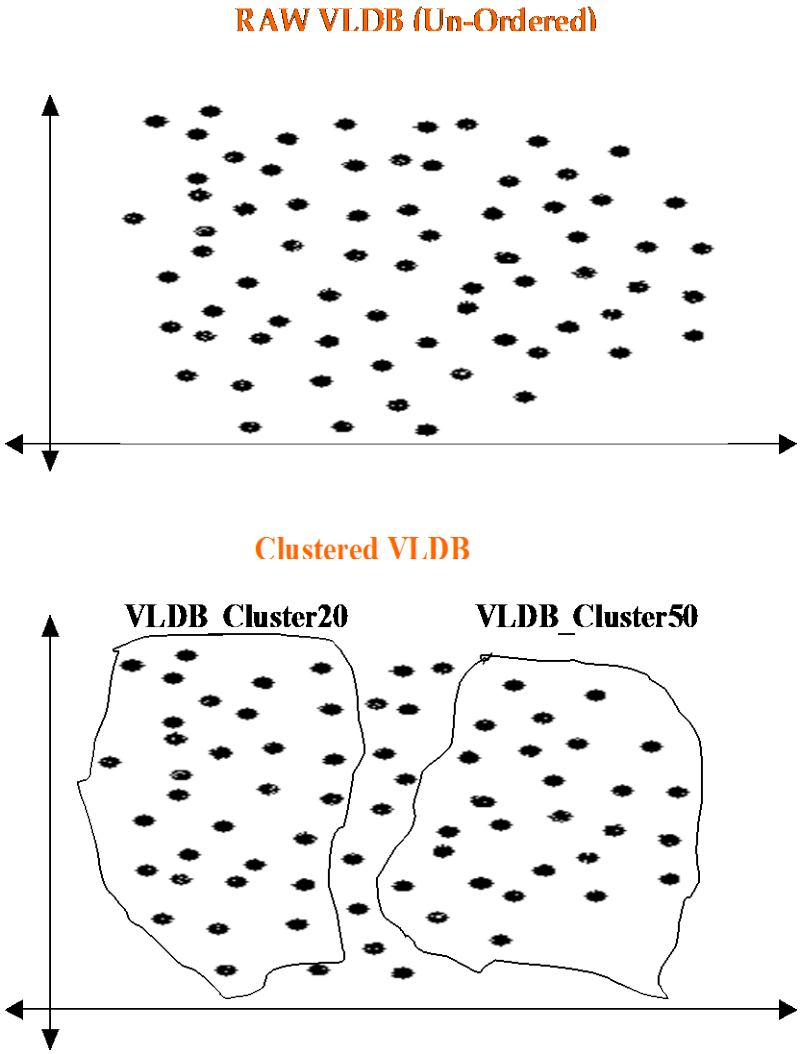

Figure 2:Ordered and Unordered Clustering Derivations

The below shown details (Refer Table I) classified and indexed by means of weight and location of the gene. Disease denoted by upper side for more deviation in increasing fashion (Positive Values) and disease denoted by lower side for more decreasing fashion. Raw VLDB deploys a data without rigid format. Therefore hit count of VLDB is more to obtain a similar data. On the other hand it is also consume more time to execute a result. However if we have above (Refer Figure II) clustering blocks for gene sets shall reduce the same.

III. OPTIMAL LINK STATE

Linking and accessing of VLDB is a major cumbersome area in a real time transaction processing. To reduce the time consumption to found the required data from VLDB, the entire database is clustering into several pieces based on its weight and location so that similarity analysis can be done easily on these clustered data.

For example consider a shopping complex with different variety of items. Kindly think over what will happen if single room dumping with all the items and there is no specific identifier. Absolutely confusion may arise and it will take more time to identify our likings. The same situation will arise in this data mining concept where users are the customers and raw VLDB are the items spread over the data warehouse. We remove this bottleneck by building separate rooms to have different variety of items separately (Cluster). This shall enable us to directly access our fondness instead of searching an entire shopping complex.

Algorithm for clustering

// healthy tissues

Function VLDB_cluster

```
begin
    intake (element)
    if (weight > 20 and weight < =50)     then
    begin

            VLDB_cluster20(attributes)
    else
            VLDB_cluster(attributes)

    end;
end function
```

Function VLDB_cluster20(attributes)
begin
// stores gene, location, size, shape, weight and location
index on weight, location
end function

Function VLDB_cluster50(attributes)
begin
// stores gene, location, size, shape, weight and location
index on weight, location
end function

Table 1: Genome Classification Based on Weight and Location

GENE	SEX	LOCATION	SIZE (X,Y)	WEIGHT (IN MICRONS)	SHAPE	STRENGTH (COUNT)	DISEASE (UPPER SIDE)	DISEASE (LOWER SIDE)
TGF-1	F	29	0.4,0.6	0.08	1	30	BREAST CANCER	SKIN
AML1	M/F	16	0.8,0.8	0.56	5	20	LEUKEMIA	----------
PTTG1	M/F	18	0.2,0,5	0.004	7	30	COLON CANCER	----------
CO	M/F	14	0.5,0.3	0.078	6	15	----------	ALZHEIMER
APOE	M/F	16	0.6,0.1	0.18	8	10	MULTIPLE SCLEROSIS	----------
PON-Q	M/F	18	0.6,0.1	0.07	4	07	GULF WAR SYNDROME	----------
Tlr3	M/F	09	0.7,0.9	0.89	5	29	SENSING SWEET TASTE	----------

Table 2: Genome Classification Based on Weight and Location (Continuation)

GENE	SEX	LOCATION	SIZE (X,Y)	WEIGHT (IN MICRONS)	SHAPE	STRENGTH (COUNT)	DISEASE (UPPER SIDE)	DISEASE (LOWER SIDE)
p56lck	M/F	11	0.1,0.1	0.45	2	25	CARDIAC ATTACK	CARDIAC BLOCK
BRCA1 & BRCA2	F	15	0.6,0.8	0.12	5	09	BLOOD CANCER	INCREASE OF RED CELLS
STEAP (Short for Six-Trans membrane Epithelial Antigen of the	M/F	17	0.4,0.9	0.19	9	18	CANCER	

IV. DNA DATA COMPARISON – SIMILARITY ANALYSIS

Similarity Analysis plays a vital role for this algorithm to enter into its vision. In order to improve the processing rapidness, all the non-numbering attributes such as locations and shapes are mapping into corresponding number values. After an identification of cluster using the aforesaid algorithm, the similarity analysis takes over the control and compares healthy tissue attributes with user input. This process is able to identify too closer disease/s like comparing of photograph with individual. If the key attributes such as gene identification and location is satisfied then it scans an entire user input and checks it values with corresponding healthy tissue attributes for deviation. If any deviation found in increasing or decreasing fashion then it read the disease rules which it belongs to. It may single or multiple based on the above aspects.

Algorithm for Similarity Analysis

Function similarity_test ()

```
begin
  intake(elements)
  if (weight>=1 and weight<=20) then
  begin
    //bypass search operation to VLDB_cluster20
    instead of searching entire database
    foreach weight :=1 to n
    begin
      if (gene_id=VLDB_cluster20.gene_id and
      location=VLDB_cluster20.location) then
       begin
          read(disease rule);
          if (weight or size or strength >disease rule)
          then
             begin
                write (upper_disease)
             else if (weight or size or strength <
          disease rule) then
                write (lower_disease)     else
                write (Combination)
             end;end;
          end;
    else if (weight >20 and weight<=50) then
   begin
       //bypass search operation to VLDB_cluster50
       instead of searching entire database
    foreach weight :=1 to n
    begin
       if (gene_id=VLDB_cluster50.gene_id and
       location=VLDB_cluster50.location) then
       begin
          read(disease rule);
          if (weight or size or strength >disease rule) then
          begin
 //bypass search operation to VLDB_cluster50 instead of
searching entire database
foreach weight :=1 to n
begin if (gene_id=VLDB_cluster50.gene_id and
location=VLDB_cluster50.location) then
begin
    read (disease rule);
    if (weight or size or strength > disease rule) then
    begin
          write (upper_disease)
    else if (weight or size or strength < disease rule) then
          write (lower_disease)
    else
          write (Combination)
    end; end;
else
    // Similar process of VLDB_cluster
end all the case
```

end function

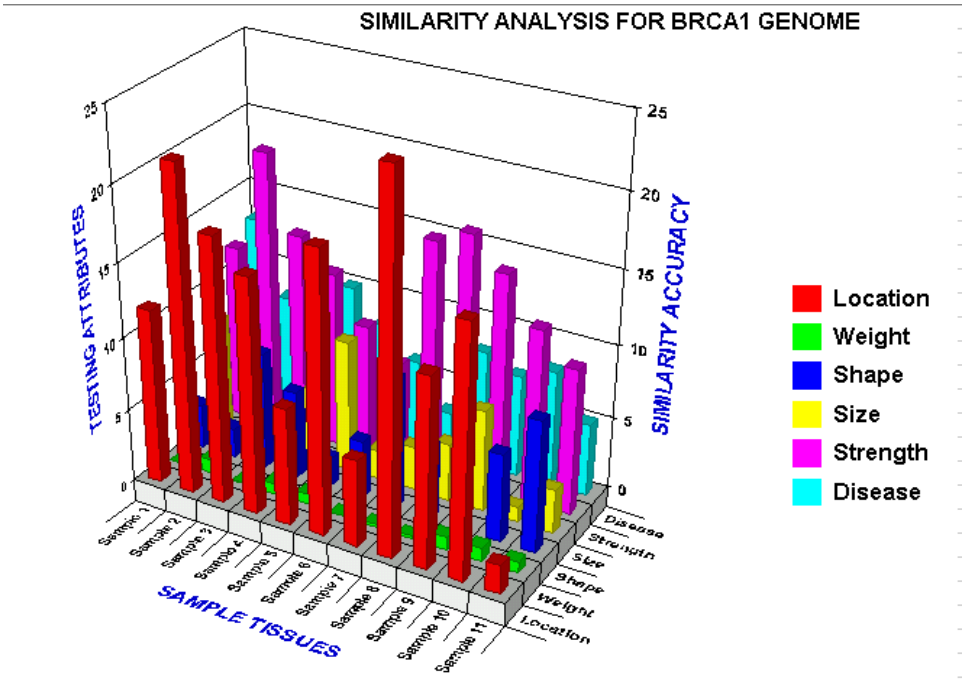

Figure 3: Similarity Search (Normal Attributes)

V. CONCLUSION

The algorithm named "Typical Genomic Framework on Disease Analysis (TGFDA)" satisfies the research area of gene and its behavioral studies. In view of further enhancement of the same, Similarity analysis with path analysis shall be implemented to analyze the group of genes which are reason for the disease/s. While the group of genes may contribute to a disease process, different genes may become active at different stages of the disease. If the sequence of genetic changes across the different stages of disease development is able to identify shall be possibly lead to the development of new medicines that targets to the different stages of diseases in a futuristic manner.

VI. REFERENCES

[1] R. Agrawal and R.Srikant, "Fast Algorithms for Mining Association Rules," Proc. 1994 Int'lconf. Very Large Data Bases, pp. 487-499, Santiago,Chile,Sept. 2010.

[2] Haruka Fuse, Haruka Fukamachi, Mitsuko Inoue and Takeshi Igarshi, "Identification and Functional Analysis of the Gene Cluster",Gene Volume 515,Issue 2, Pages 291-297,25th February 2013, Elsevier Publications

[3] D.W. Cheung, J. Han, V. Ng,A. Fu, and Y. Fu, "A Fast Distributed Algorithm for Mining Association Rules," Proc. 1996 Int'l Conf. Parallel and Distributed Information Systems, PP. 1996 Int'l Conf. Data Enf., PP. 106-114, New Orleans, Feb. 2010. doi: 10.1109/PDIS.1996.568665

[4] L. Li, C.R. Weinberg, T.A. Darden, L.G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method", Bioinformatics 17 (12) (2001) 1131–1142. doi: 10.1007/978-3-642-13089-2_49

[5] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nat. Med. 7 (6) (2001) 673–679.

[6] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Bostein, "Cluster analysis and display of genome-wide expression patterns", Proceedings of the National Academy of Science USA 95 (1998) 14,863–14,868.

[7] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. GaasenBeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri,C.D. Blomfield, E.S. Lander, "Molecular classification of cancer: class discovery and class prediction by gene-expression monitoring", Science 286 (1999) 531–537. doi: 10.1126/science.286.5439.531

[8] E. Frank, I.H. Witten, "Generating accurate rule sets without global optimization", in: , Machine Learning: Proceedings of the 15th International Conference, Morgan Kaufmann Publishers, Los Altos,CA, 1998

[9] Y. Fu and J. Han, V. Ng, A. Fu, and Y. Fu, "A Fast Distributed Algorithm for Mining Association Rules," Proc. 1996 Int'l Conf. Parallel and Distributed Information Systems, PP. 31-44, Miami Beach, Fla.,Dec. 2001.

[10] D.W. Cheung, J. Han,V. Ng, and C.Y. Wong, "Maintenance of Discovered Association Rules in Large Databases: An Incremental Updating Technique," Proc. 1996 Int'l Conf, Data Engg., PP. 106-114, New Orleans, Feb. 2009. doi: 10.1109/ICDE.1996.492094

[11] D.W. Cheung, J. Han, V. Ng,A. Fu, and Y. Fu, "A Fast Distributed Algorithm for Mining Association Rules," Proc. 1996 Int'l Conf. Parallel and Distributed Information Systems, PP. 1996 Int'l Conf. Data Engg., PP. 106-114, New Orleans, Feb. 2010. doi: 10.1109/ PDIS.1996.568665

[12] M.S. Chen, J. Han, and P.S. Yu, "Data Mining: An overview from a Database Perspective," IEEETrans. Knowledge and Data Engg., Vol.8, PP.866-883,1996

[13] R. Agrawal, T. Imielinski, and A. Swami, "Mining Association Rules Between Sets of Items in Large Databases," Proc. 1993 ACM SIGMOD Int'l Conf. Management of Data, pp. 207-216, Wahington, D.C., May 1993. doi: 10.1145/170036.170072

[14] Jiwai Han, Micheline Kamber, "Mining Concepts and Techniques",Elsevier,2005

[15] Jae K Lee, paul D Williams and Sooyoung Cheon, "Data Mining in Genomics" ,Clin Med Lab,2008

[16] I. Witten, E. Frank, "Data Mining", Morgan Kaufmann, San Francisco,CA, 2000.

[17] Anand Rajaraman, Jure Leskovec,Jeffrey D. Ullman, "Mining of Massive Datasets", Individual copright, 2012

How to cite

J. Stanly Thomas, Dr. N. Rajkumar, "*Typical Genomic Framework on Disease Analysis*". *International Journal of Research in Computer Science*, 4 (5): pp. 11-15, January 2015.