IJPSR (2017), Vol. 8, Issue 6 **(Research Article)**

INTERNATIONAL JOURNAL OF PHARMACEUTICAL SCIENCES AND RESEARCH

Received on 02 December, 2016; received in revised form, 16 January, 2017; accepted, 17 February, 2017; published 01 June, 2017

PERFORMANCE OF ASSOCIATION RULES FOR DENGUE VIRUS TYPE1 AMINO ACIDS USING AN INTEGRATION OF TRANSACTION REDUCTION AND RANDOM SAMPLING (TRRS) ALGORITHM

D. Kerana Hanirex *, K. P. Thooyamani and V. Khanaa

Bharath University,Chennai-73, Tamilnadu, India.

Keywords:

Association Rule Mining, Association Rules, Frequent Item set, Dengue Virus Type1, Apriori, FP Growth, Two Dimensional Transaction Reduction (TDTR), Random Sampling, Transaction Reduction and Random Sampling Algorithm (TRRS).

Correspondence to Author:
D. Kerana Hanirex

Assistant Professor/MCA,
173, Agharam Road, Bharath
University, Chennai-600073
Tamilnadu, India.

E-mail: keranahanirex.cse@bharathuniv.ac.in

ABSTRACT: Association rule mining is the recent research area in data mining. Frequent Itemset Mining techniques is one of the prominent techniques in pattern mining. In this system frequent itemset mining is used to find the frequent amino acids patterns in Dengue Virus Type1 data set. This system uses an integration of Transaction Reduction and Random Sampling (TRRS) approach to identify the frequent patterns in Dengue Virus Type1 amino acids sequence. Our system reveals the association between the amino acids. Our System first identifies the number of amino acids sequences suitable for each transaction and finds the number of association rules using Transaction Reduction and Random Sampling methods by varying the sample size. Experimental results show that the performance of our proposed Transaction Reduction and Random Sampling Algorithm(TRRS) works efficiently when compared to Apriori algorithm, FP Growth algorithm, Two Dimensional Transaction Reduction(TDTR) Algorithm Improved TDTR algorithm, Set Oriented Mining (SETM) algorithm and Improved SETM Algorithm (ISETM) in terms of number of association rules generated and the time taken to generate the association rules.

INTRODUCTION: Mining association rule is the recent research area. Data mining is used to find the hidden pattern and useful information from the data base [1, 9]. Association rule mining is one of the major task of the Data mining [2, 3]. The other data mining tasks include clustering [37, 30], classification [4, 21]. Frequent itemset mining is not only used in market basket analysis but also used in bioinformatics such as gene expression data and protein analysis [10, 23].

QUICK RESPONSE CODE

DOI: 10.13040/IJPSR.0975-8232.8(6).2578-87

Article can be accessed online on: www.ijpsr.com

DOI link: http://dx.doi.org/10.13040/IJPSR.0975-8232.8 (6).2578-87

The algorithm which is developed for market basket problem can also be applied to solve various bioinformatics problem such as analysing frequent patterns in amino acids. Here each transaction is identified by the sequence of amino acids. Various algorithms have been proposed to find the frequent itemsets but it differs in its computational efficiency. Frequent itemsets can be converted into association rules that can be used in further applications.

Association Rule Mining: Association rule mining problem can be done in 2 steps. First find frequent itemsets and then generate rules from the frequent itemsets. A brute-force method for frequent itemset mining is to generate support and confidence measure for all generated association rules. This method is expensive because the search space is exponential to the number of itemsets present in the database [2]. Most algorithms that are used for association rule mining are differentiated by its search space and their computation of support value. At each stage it generates candidate itemsets and calculate its support count and remove its itemsets that are having less support which is infrequent in the database. Search space traversal may be either depth first search or breadth first search. In breadth first traversal, itemsets are generated starting from the singleton sets. In depth first traversal, it uses divide and conquer strategy. Apriori algorithm uses breadth first approach and FP Growth algorithm uses depth first approach. If the data is not fit into the memory it takes each transaction from the database one by one. Various optimization techniques such as sampling and partitioning are used to fit the transactions from the database into memory. FP Tree algorithm uses compressed tree structure that are memory resident [18]. Éclat [43] is an algorithm which uses depth first approach.

Association rule is of the form X=> Y where X and Y are the itemsets and X∩Y =Ø. This implies that a transaction which contains X also contains Y. X is the antecedent of the rule and Y is the consequent of the rule.

Confidence and support are the two important measures of Association rule mining. Support determines how frequently the rules occur in the database.

Support (X=>Y, D) = Support (XUY, D)………(1)

Confidence of an association rule X=>Y is the ratio of total occurrences of X and Y to the total number of occurrence of X.

Confidence (X=>Y,D) = Support(XUY,D) /support (X, D) ………………….(2)

Association Rule Mining is used to generate a set of Association Rules. Various kinds of interesting measures have been proposed [38, 14] for biological datasets. The other interesting measures are lift and coverage. Lift describes the ratio between the observed support for X=>Y to the expected support value when X and Y are independent. The coverage of an association rule states that how often the rule is present in the database.

Related work: Researchers proposed various algorithms and methods to find the frequent itemset. Apriori algorithm is the well-known standard algorithm for association rule mining which requires large number of data scans and candidate generation. FP Growth algorithm is used to find the frequent itemset generation without candidate itemset generation. FP Growth is the fastest algorithm than Apriori which is based on prefix tree representation can save memory [19]. Hashing technique [22] can also be used to improve the way of finding frequent itemsets. FP-Streaming [15] and Regression parameter [11] has also been proposed to find the frequent itemsets. The paper [5, 16] uses Genetic algorithm for association rule mining. Genetic algorithm uses fitness function and genetic operators such as selection, crossover and mutation to find the frequent itemsets.

Soumadip Ghosh *et al.,* proposes [36] genetic algorithm to find frequent itemsets. Apriori algorithm has certain limitations [3] such as producing large number of redundant patterns. This redundant rules can be removed by various techniques such as by reducing search space, considering either maximal frequent itemsets or top-k frequent itemsets [8, 40, 17]. Cai R, Hao Z, Wen W, *et al.,* [13] proposes kernel density estimation measurement to reduce the irrelevant rules. Closed itemsets has also been considered for generation of association rules [34].

Bioinformatics: Frequent itemset mining is used in Bioinformatics to analyze gene. It is used to analyze co-occurring frequent annotation patterns in molecular biology. It can also be used in cross ontology mining as well as gene ontology [24, 32, 6]. Frequent items mining can also be used in finding structural patterns or motif discovery in biomolecules [31, 39]. Association rule mining can be used in analysis of gene expression data [12, 7]. It has also been used to identify the strong factors associated with the particular diseases [33]. Frequent sub graph can be identified from the molecular graph [41]. Association Rules can be used to build a classifier which is more accurate to solve biological problems [20, 25].

Data Collection: Dengue virus (DENV) belongs to the family Flaviviridae and Genus Flavivirus. There are four serotypes namely (DEN 1-4) [35]. At present there are no proper vaccines for diseases like Dengue, Ebola and Anticancer drugs. This system will find the hidden patterns in polyprotein Dengue virus type1 DNA sequence and find the dominating amino acids using data mining techniques and to improve the quality of finding drugs for the pharmacists.

This system will increase the demands in all pharmaceutical companies through innovation and to treat the patients carefully. The association between dominating amino acids will be useful for the drug designers to develop the antibiotics for the virulent diseases caused by viruses such as ebola, dengue and anticancer drugs. This proposed system uses protein dengue virus type1 datasets from NCBI (National Centre for Biotechnology Information). It uses GenBank: AB189120.1 which consists of 3392 amino acids.

Sample Amino Acids sequence for GenBank: AB189120.1

```
MNNQRKKTGR PSFNMLKRAR NRVSTVSQLA
KRFSKGLLSG QGPMKLVMAF IAFLRFLAIP
PTAGILARWG SFKKNGAIKV LRGFKKEISN
MLNIMNRRKR SVTMLFMLLP TALAFHLTTR
GGEPHMIVSK QERGKSLLFK TSAGVNMCTL
IAMDLGELCE DTMTYKCPRI TETEPDDVDC
WCNATETWVT YGTCSQTGEH RRDKRSVALA
PHVGLGLETR TETWMSSEGA WRQIQKVETW
ALRHPGFTVM ALFLAHAIGT SITQKGIIFI
LLMLVTPSMA MRCVGIGNRD FVEGLSGATW
```

FIG. 1: AMINO ACIDS SEQUENCE FOR DENGUE VIRUS TYPE1 DATASET

Data Preprocessing: To improve the quality of data, it requires pre-processing techniques. Data mining task includes data cleaning, data transformation, normalizing, data aggregation and discretization. This system uses data transformation as a pre-processing techniques which transforms the data suitable for analysis.

Selecting suitable number of amino acids sequence for a transaction: Our first part of the research work is to identify the suitable number of sequences. Polyprotein Dengue Virus Type 1 Dataset consists of sequence of 3392 amino acids. Our research work first identifies how many amino acids sequence we can take together for a transaction. For the analysis, amino acids are taken as sequence of 10,11,12,13,14,15,16,17,18,19 and 20 from the data set. The number of association rules generated for each sequence T10, T11, T12, T13, T14, T15, T16, T17, T18, T19 and T20 are measured by varying confidence and support measure. Association rules are measured using Apriori algorithm in R tool. The following table shows the number of rules generated for some of the amino acids sequence.

The following table **Table 1, 2, 3**. shows the number of Association Rules generated for different aminoacids sequence by taking support=0.1 and varying confidence.

TABLE 1: NUMBER OF ASSOCIATION RULES GENERATED FOR DIFFERENT AMINOACIDS SEQUENCE BY TAKING SUPPORT = 0.1 AND BY VARYING CONFIDENCE FROM 0.9 TO 0.1

Confidence	T10	T12	T14	T16	T18	T20
0.9	-	-	-	3	220	802
0.8	-	-	21	227	1687	4645
0.7	-	-	302	1500	5687	10807
0.6	16	75	1133	3909	11113	18033
0.5	87	321	2134	5702	14071	22610
0.4	268	572	2609	6631	15916	24985
0.3	331	693	2895	7076	16580	26075
0.2	381	733	2976	7196	16777	26279
0.1	385	740	2980	7197	16778	26279

TABLE 2: NUMBER OF ASSOCIATION RULES GENERATED FOR DIFFERENT AMINOACIDS SEQUENCE BY TAKING SUPPORT = 0.2 AND BY VARYING CONFIDENCE FROM 0.9 TO 0.1

Confidence	T10	T12	T14	T16	T18	T20
0.9	-	-	-	-	7	38
0.8	-	-	-	20	229	684
0.7	-	-	58	262	842	1783
0.6	1	17	213	721	1788	2991
0.5	24	59	398	925	2075	3505
0.4	61	110	447	1071	2282	3730
0.3	68	126	480	1093	2315	3785
0.2	71	128	484	1099	2326	3793
0.1	71	128	484	1099	2326	3793

TABLE 3: NUMBER OF ASSOCIATION RULES GENERATED FOR DIFFERENT AMINOACIDS SEQUENCE BY TAKING SUPPORT=0.3 AND BY VARYING CONFIDENCE FROM 0.9 TO 0.1

Confidence	T10	T12	T14	T16	T18	T20
0.9	-	-	-	-	-	3
0.8	-	-	-	-	58	182
0.7	-	-	18	84	211	483
0.6	-	12	59	226	440	751
0.5	5	26	105	261	483	840
0.4	11	34	113	292	518	869
0.3	16	40	116	294	521	872
0.2	16	40	116	294	521	872
0.1	16	40	116	294	521	872

The above tables show that T20 sequence reveal association rules for higher confidence and for increasing support values.

The following **Fig. 2** represents the number Association Rules generated for different amino acids sequence by taking support=0.3 and by varying confidence.

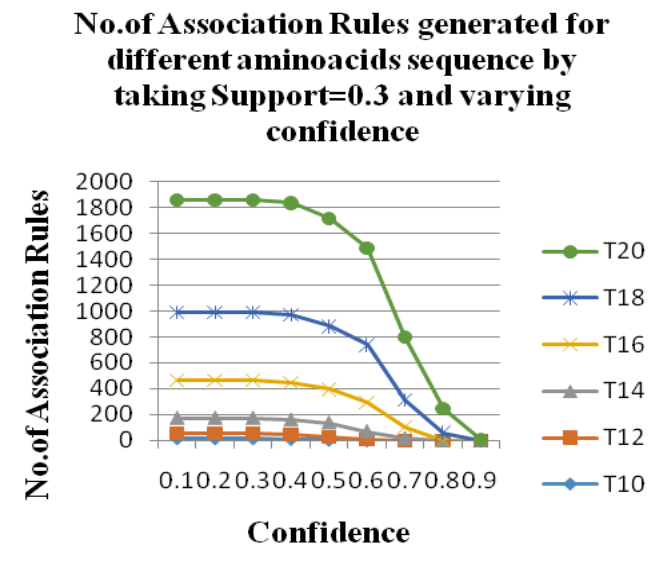


FIG. 2: NUMBER OF ASSOCIATION RULES GENERATED FOR DIFFERENT AMINO ACIDS SEQUENCE FOR SUPPORT =0.3

The above graph reveals that T20 sequence exhibits association rules for all confidence measure. Hence 20 amino acids sequence is taken for each transaction which will be considered for our further analysis.

In further analysis, the number of association rules generated based on T20 sequence by varying different confidence and support measure. The following graph **Fig. 3** shows the number of association rules generated for T20 sequence by varying different confidence and support measure.

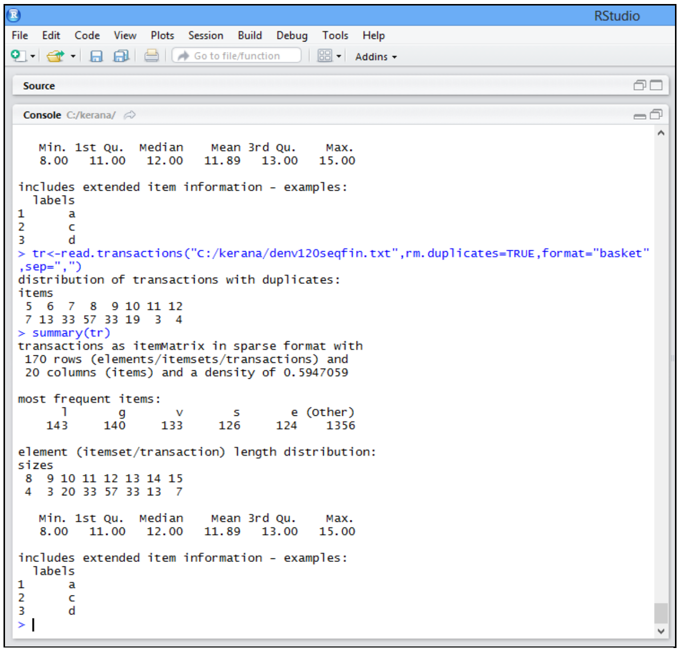
This data set consists of minimum of 8 different items and a maximum of 15 different items in a transaction. The following table **Table 4** shows the number of transactions that contain different items of different size or length.

TABLE 4: MOST FREQUENT ITEMS IN THE SPARSE MATRIX AND ITS NUMBER OF OCCURRENCES

Items	L	G	V	S	E	other
Number of occurences	143	140	133	126	124	1356

TABLE 5: LENGTH DISTRIBUTION OF DIFFERENT ITEMS IN A TRANSACTION

Item Length	8	9	10	11	12	13	14	15
Number of Transactions	4	3	20	33	57	33	13	7



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.00   11.00   12.00   11.89   13.00   15.00

includes extended item information - examples:
  labels
1      a
2      c
3      d
> tr<-read.transactions("C:/kerana/denv120seqfin.txt",rm.duplicates=TRUE,format="basket"
,sep=",")
distribution of transactions with duplicates:
items
 5  6  7  8  9 10 11 12
 7 13 33 57 33 19  3  4
> summary(tr)
transactions as itemMatrix in sparse format with
 170 rows (elements/itemsets/transactions) and
 20 columns (items) and a density of 0.5947059

most frequent items:
      l       g       v       s       e (Other)
    143     140     133     126     124    1356

element (itemset/transaction) length distribution:
sizes
 8  9 10 11 12 13 14 15
 4  3 20 33 57 33 13  7

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   8.00   11.00   12.00   11.89   13.00   15.00

includes extended item information - examples:
  labels
1      a
2      c
3      d
> |
```

FIG. 3: ANALYSIS OF DENGUE VIRUS AMINO ACIDS DATA SET

From the above table **Table 5**, it reveals that there are 57 transactions of length 12 different items in the dataset. Also there are 7 transactions having itemset of length 15. The following figure **Fig. 4** shows the frequencies of amino acids sequence in the dengue virus type 1 amino acids data set.

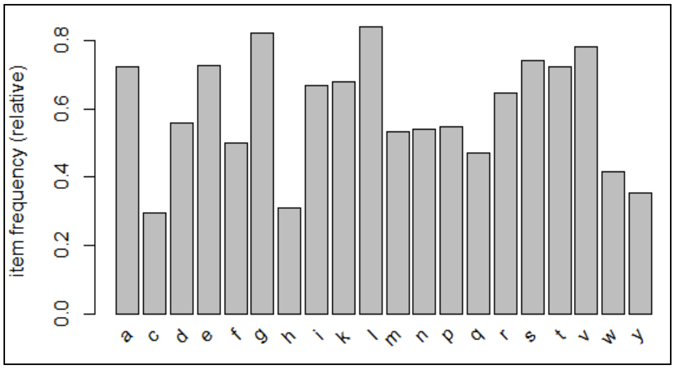


FIG. 4: ITEM FREQUENCIES OF AMINO ACID SEQUENCE

Apriori Algorithm: The Apriori Algorithm is the most well-known association rule algorithm. It uses downward closure property. This algorithm is based on largest item set property which states that “Any subset of a large item set must be large” [1, 2].

In our earlier research work, the Apriori algorithm is implemented for Dengue Virus Type1 Dataset with 777 aminoacids [26]. This paper implements Apriori algorithm for GenBank: AB189120.1 which consists of 3392 amino acids. This Apriori algorithm is implemented using R tool.

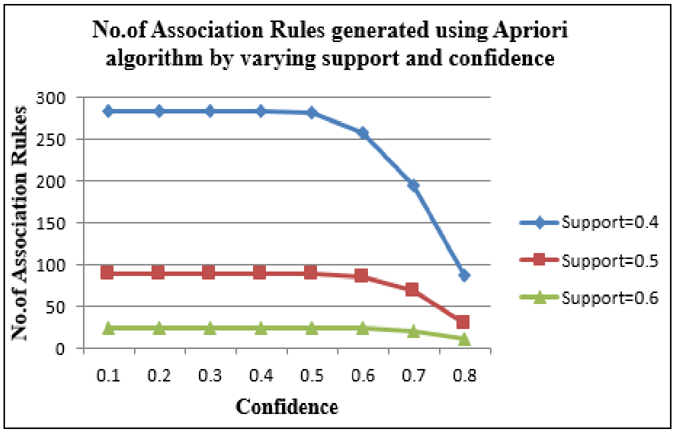


FIG. 5 NUMBER OF ASSOCIATION RULES GENERATED USING APRIORI ALGORITHM BY VARYING SUPPORT AND CONFIDENCE MEASURE

From the above **Fig. 5** we can understand that from the confidence value 0.1 to 0.5 the system reveals similar number of association rules and above the confidence value 0.5 the number of rules generated gets decreased. The above graph shows that Apriori algorithm generates more association rules for decreasing confidence values and less association rules for increasing support values.

FP Growth Algorithm: FP Growth algorithm is used to find the frequent itemset generation without candidate itemset generation. FP Growth algorithm is a two-step process. In the first step, it builds compact data structure called the FP-tree. It builds this FP Tree using 2 scans over the database. From the FPTree it generates frequent itemsets. FP Growth algorithm is implemented for this data set. The number of association rules generated by the FP Growth algorithm is measured by varying support value from 0.4 to 0.6 and confidence value 0.1 to 0.8.This algorithm is implemented using Rapid Miner tool.

The following **Fig. 6** shows that FP Growth algorithm generates more association rules for decreasing confidence values and less association rules for increasing support values. For the support value 0.4 and 0.5 it reveals similar number of association rules.

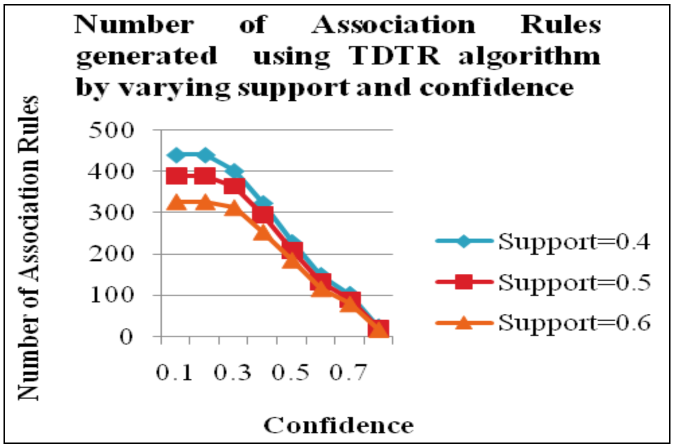


FIG. 6: NUMBER OF ASSOCIATION RULES GENERATED USING TDTR ALGORITHM BY VARYING SUPPORT AND CONFIDENCE

The above graph shows that TDTR algorithm generates more association rules for decreasing confidence values and less association rules for increasing support values. This TDTR algorithm clearly exhibits the association rules for higher confidence value.

Transaction Reduction and Random Sampling (TRRS) Algorithm: Our second part of the research work implements Transaction Reduction and Random Sampling Algorithm (TRRS) which integrates transaction reduction and random sampling approach to find the frequent dominating amino acids and to generate association rules in Dengue Virus Type1 dataset This algorithm integrates our Two Dimensional Transaction Reduction (TDTR) Algorithm with the Random Sampling method. Our earlier approach integrates TDTR algorithm with systematic sampling with 777 amino acids sequence. By integrating TDTR algorithm with sampling the efficiency of the algorithm gets increased [27]. This TRRS algorithm integrates Two Dimensional Transaction Reduction (TDTR) Algorithm with the Random Sampling for GenBank: AB189120.1 which consists of 3392 amino acids.

TRRS Algorithm:
//Algorithm to find frequent itemset and to generate association rules:

```
for each $t_i \in D$ do
begin
count the number of items in count1[i]
```

```
if the count1[i] $\geq$ min_sup then put the transactions in to $D_1$
end
for each $I_i \in D_1$ do
begin
count the number of transactions in count2[i]
If the count2 [i] <min_sup then remove that $I_i$ from $D_1$
end
begin
select optimal sample size s (Random sampling) from $D_1$.
find the frequent item sets from $D_1$ using FP-GROWTH algorithm with min_sup
generate association rules with min_conf with the optimal sample size
end
```

Algorithm: Transaction Reduction and Random Sampling: This TRRS algorithm first implements TDTR algorithm to reduce the size of the Database for further analysis. Then it selects the optimal sample size using Random sampling method. Optimal sample is selected based on the number of association rules revealed. Based on the optimal size, the frequent item sets are generated from the reduced database using FP-GROWTH algorithm with min_sup and association rules are found based on min_conf.

Generation of Association Rules: TRRS Algorithm (With Replacement)

TABLE 6: NUMBER OF ASSOCIATION RULES GENERATED USING TRRS ALGORITHM BY VARYING CONFIDENCE AND SAMPLE SIZE FOR SUPPORT = 0.6(WITH REPLACEMENT)

	Sample size						
Confidence	30%	40%	50%	60%	70%	80%	90%
0.1	332	516	506	404	400	350	332
0.2	332	516	506	404	400	350	332
0.3	321	485	477	392	386	342	322
0.4	264	407	397	333	324	289	265
0.5	212	317	297	248	247	216	201
0.6	149	212	205	168	163	131	123
0.7	76	134	130	110	117	86	80
0.8	41	66	48	48	44	26	19

The above graph shows that TRRS algorithm generates more association rules for decreasing confidence values and less association rules for increasing support values. The association rules are measured by varying sample size (With Replacement). From the above graph we can understand that sample size 40% reveals large number of association rules when compared with other sampling size. Our research work identifies that instead of considering the entire database it is enough to consider 40% of the transactions from the original database.

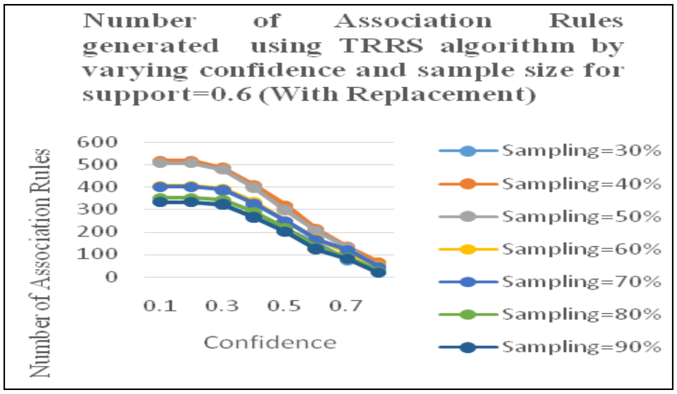


FIG. 7: NUMBER OF ASSOCIATION RULES GENERATED USING TRRS ALGORITHM BY VARYING CONFIDENCE AND SAMPLE SIZE FOR SUPPORT = 0.6 (WITH REPLACEMENT)

Generation of Association Rules: TRRS Algorithm (Without Replacement)

TABLE 7: NUMBER OF ASSOCIATION RULES GENERATED USING TRRS ALGORITHM BY VARYING CONFIDENCE AND SAMPLE SIZE FOR SUPPORT = 0.6 (WITHOUT REPLACEMENT)

	Sample size						
Confidence	30%	40%	50%	60%	70%	80%	90%
0.1	402	628	402	472	402	332	336
0.2	402	628	402	472	402	332	336
0.3	390	583	394	447	385	323	320
0.4	325	504	324	374	317	268	264
0.5	245	367	247	280	227	199	200
0.6	177	223	162	181	150	122	129
0.7	103	150	104	112	102	81	84
0.8	56	88	51	48	40	12	20

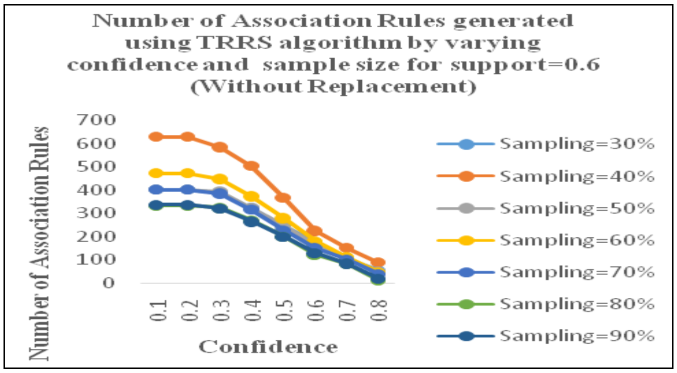


FIG. 8: NUMBER OF ASSOCIATION RULES GENERATED USING TRRS ALGORITHM BY VARYING CONFIDENCE AND SAMPLE SIZE FOR SUPPORT = 0.6 (WITHOUT REPLACEMENT)

The above graph shows that sampling with replacement and without replacement doesn't produce much variations in finding association rules in association rule mining.

Performance Comparison:

Comparison of Algorithms: TRRS with Apriori

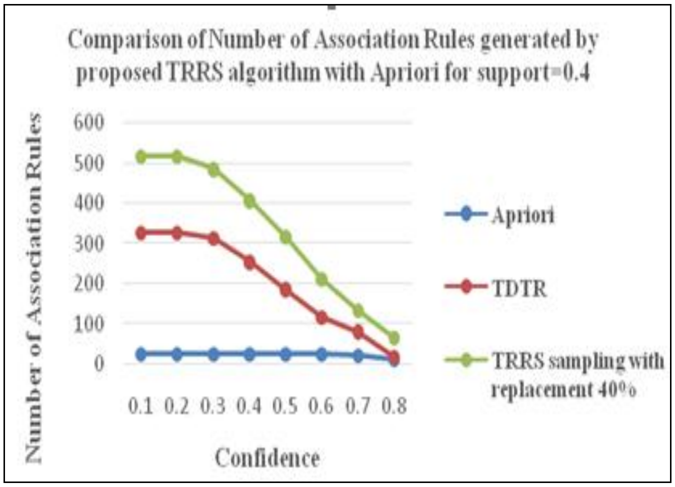


FIG. 9: COMPARISON OF NUMBER OF ASSOCIATION RULES GENERATED BY PROPOSED TRRS ALGORITHM WITH APRIORI FOR SUPPORT = 0.4

Comparison of Algorithms: TRRS with FP Growth

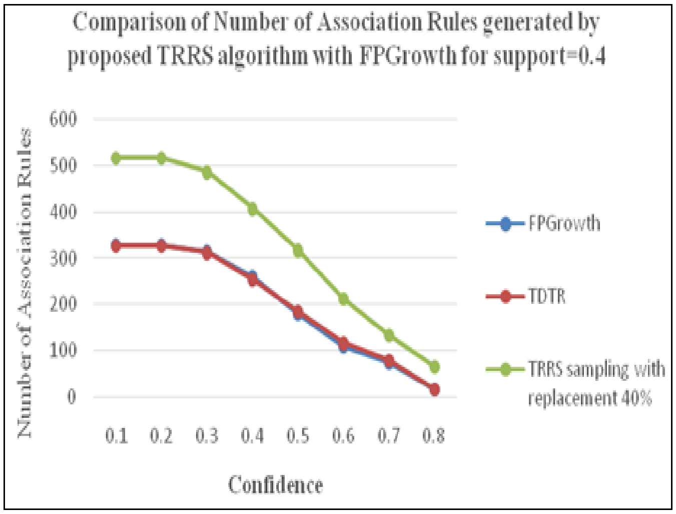


FIG. 10: COMPARISON OF NUMBER OF ASSOCIATION RULES GENERATED BY PROPOSED TRRS ALGORITHM WITH FP GROWTH FOR SUPPORT=0.4

TRRS Algorithm (Random Sampling Vs Stratified Sampling):

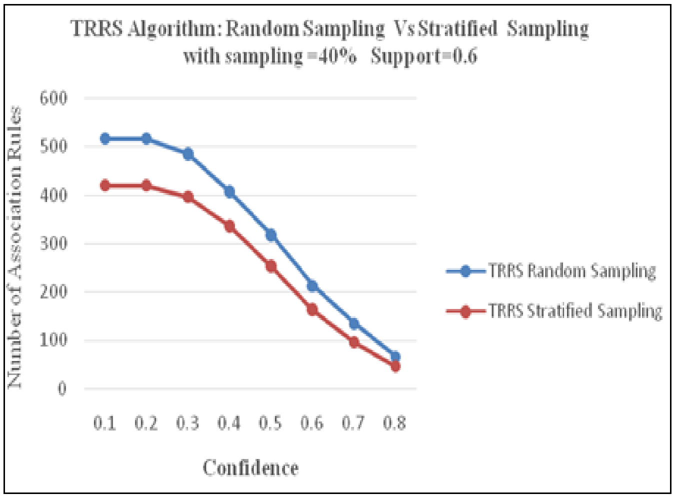


FIG. 11: COMPARISON OF NUMBER OF ASSOCIATION RULES GENERATED BY PROPOSED TRRS ALGORITHM (RANDOM SAMPLING VS STRATIFIED SAMPLING)

Comparison of TRRS Algorithm with SETM, ISETM and ITDTR:

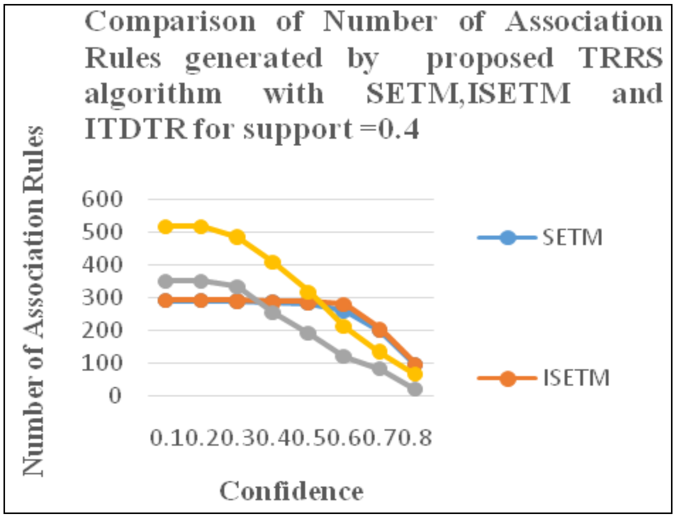


FIG. 12: COMPARISON OF NUMBER OF ASSOCIATION RULES GENERATED BY PROPOSED TRRS ALGORITHM WITH SETM, ISETM AND ITDTR FOR SUPPORT = 0.4

Research Findings:

1. Transaction Reduction approach can also be combined with other sampling methods.

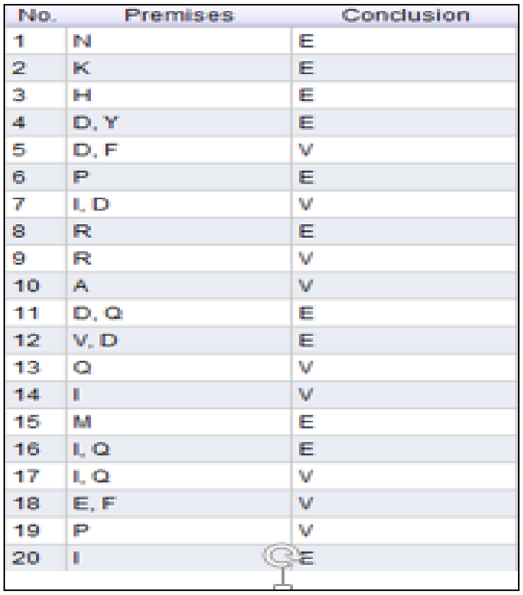
2. Random sampling method is an efficient sampling method to be combined with transaction reduction.

3. TRRS algorithm works efficiently than Apriori algorithm.

4. It reveals stable and similar behaviour with FP Growth algorithm. It produces more number of association rules than FP Growth algorithm.

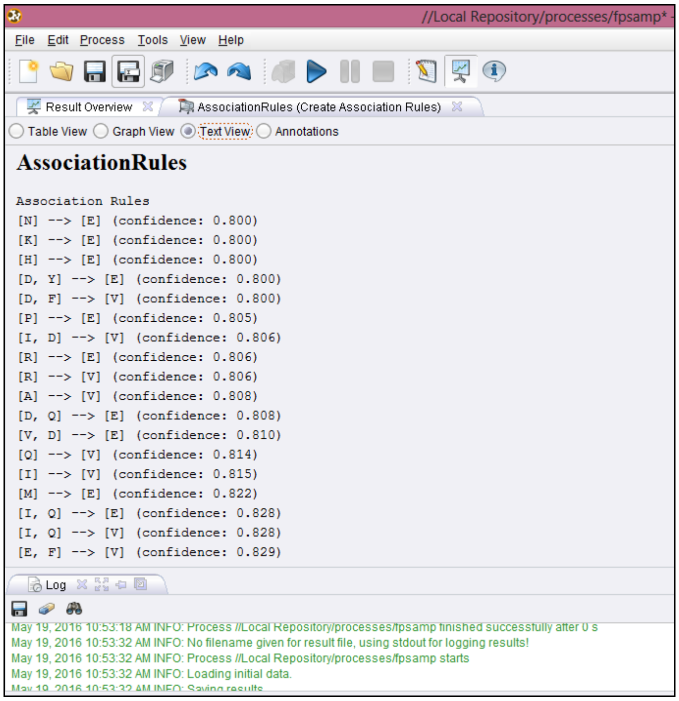
5. TRRS algorithm produce optimal association rules. It produces more association rules for decreasing confidence values and less association rules for increasing support values.

Sample Screen shots:



No.	Premises	Conclusion
1	N	E
2	K	E
3	H	E
4	D, Y	E
5	D, F	V
6	P	E
7	I, D	V
8	R	E
9	R	V
10	A	V
11	D, Q	E
12	V, D	E
13	Q	V
14	I	V
15	M	E
16	I, Q	E
17	I, Q	V
18	E, F	V
19	P	V
20	I	E

FIG. 13: ASSOCIATION RULES IN TABLE VIEW FOR SUPPORT=0.6, SAMPLING=40%, CONFIDENCE=0.8


//Local Repository/processes/fpsamp* -
File
Edit
Process
Tools
View
Help
Result Overview
AssociationRules (Create Association Rules)
Log

○ Table View ○ Graph View ● Text View ○ Annotations

AssociationRules

```
Association Rules
[N] --> [E] (confidence: 0.800)
[K] --> [E] (confidence: 0.800)
[H] --> [E] (confidence: 0.800)
[D, Y] --> [E] (confidence: 0.800)
[D, F] --> [V] (confidence: 0.800)
[P] --> [E] (confidence: 0.805)
[I, D] --> [V] (confidence: 0.806)
[R] --> [E] (confidence: 0.806)
[R] --> [V] (confidence: 0.806)
[A] --> [V] (confidence: 0.808)
[D, Q] --> [E] (confidence: 0.808)
[V, D] --> [E] (confidence: 0.810)
[Q] --> [V] (confidence: 0.814)
[I] --> [V] (confidence: 0.815)
[M] --> [E] (confidence: 0.822)
[I, Q] --> [E] (confidence: 0.828)
[I, Q] --> [V] (confidence: 0.828)
[E, F] --> [V] (confidence: 0.829)
```

```
May 19, 2016 10:53:18 AM INFO: Process //Local Repository/processes/fpsamp finished successfully after 0 s
May 19, 2016 10:53:32 AM INFO: No filename given for result file, using stdout for logging results!
May 19, 2016 10:53:32 AM INFO: Process //Local Repository/processes/fpsamp starts
May 19, 2016 10:53:32 AM INFO: Loading initial data.
May 19, 2016 10:53:32 AM INFO: Saving results
```

FIG. 14: ASSOCIATION RULES IN TEXT VIEW FOR SUPPORT=0.6, SAMPLING=40%, CONFIDENCE=0.8

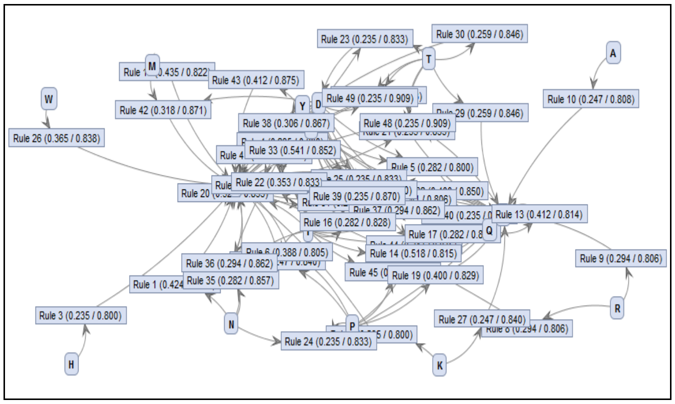


FIG. 15: ASSOCIATION RULES IN GRAPH VIEW FOR SUPPORT=0.6, SAMPLING= 40%, CONFIDENCE = 0.8

Research Findings: From the above figure we can identify the Largest Frequent Itemsets are {I,D,F,V},{I,D,F,E},{V,D,N,E},{V,D,F,I},{V,T,D} for Support = 0.6, Sampling = 40%, Confidence = 0.8. The Dominating Amino acids are Isolecuine, Aspartic Acid, Phenylalanine, Valine, Glutamic acid, Asparagine, Threonine which give knowledge to the pharmacists in drug discovery.

CONCLUSIONS: In this research work, we have implemented TDTR (Two Dimensional Transaction Reduction) Algorithm. It is integrated with Random Sampling methods which is our proposed Transaction Reduction and Random Sampling method to find frequent itemsets and hence to generate association rules. This algorithm proves its efficiency and accuracy.

CONTRIBUTIONS:

1. T20 amino acid sequence is suitable for each transaction.

2. TDTR algorithm generates more association rules for decreasing confidence values and less association rules for increasing support values.

3. TRRS algorithm generates more association rules for decreasing confidence values and less association rules for increasing support values.

4. Sampling size 40% is suitable for generating frequent itemsets in association rule mining.

5. Sampling with replacement and without replacement doesn't produce much variations in association rule mining.

6. Transaction Reduction can also be combined with other sampling methods.

7. Random sampling method is efficient to be combined with transaction reduction.

8. Generated frequent itemsets and association rules give knowledge to the pharmacists in drug discovery

9. TRRS method exhibits similar behaviour with FPGrowth algorithm which is superior than Apriori algorithm

FUTURE WORK: In future, the transaction reduction approach can be combined with Partitioning or Distributed methods to deal with large datasets. This research work concentrates on frequent items sets. Finding infrequent itemsets and negative association rules is an open topic for the future research work.

BENEFITS OF THIS RESEARCH WORK: This research work will find the hidden patterns in Polyprotein Dengue Virus Type1 DNA sequence and find the dominating amino acids using data mining techniques and to improve the quality of finding drugs for the pharmacists. This system will increase the demands in all pharmaceutical companies through innovation and to treat the patients carefully. The association between dominating amino acids will be useful to the drug designers to develop the antibiotics for the virulent diseases caused by viruses such as Ebola, Dengue, Zika and Anticancer drugs. Our research work efficiently finds the dominating amino acids in Dengue Virus Type1 Dataset using an integration of transaction reduction and random sampling approach.

REFERENCES:

1. Agrawal, R., T. Imielinksi and A. Swami, "Database mining: A performance perspective", IEEE Transactions on Knowledge and Data Engineering, 1993; 5(6): 914-925.
2. Agrawal, R. and R. Srikant," Fast algorithms for mining association rules", Proc. 20 Int. Conf. Very Large Data Bases, VLDB, edited by J.B. Bocca, M. Jarke and C. Zaniolo, Morgan Kaufmann, 1994; 12: 487-499.
3. Agrawal, R., T. Imielinski and A. Swami, "Mining Association rules between sets of items in large databases", Proc. of the ACM SIGMOD Int. Conf. on Management of Data ACM SIGMOD '93,Washington, USA, 1993; pp: 207-216.
4. Arun K Pujari," Data Mining Techniques", 5th ed., Universities Press (India) Private Limited, 2003.
5. Anandhavalli M., Suraj Kumar Sudhanshu Ayush Kumar and M.K. Ghose," Optimized association rule mining using genetic algorithm", Advances in Information Mining, 2009; 1(2): 01-04.
6. Artamonova II, Frishman G, Gelfand MS, *et al.*, "Mining sequence annotation databanks for association patterns", Bioinformatics, 2005; 21:iii49–57.
7. Becquet C, Blachon S, Jeudy B, *et al.*, "Strong-association- rule mining for large-scale gene-expression data analysis: a case study on human SAGE data". Genome Biology, 2002; 3.
8. Bayardo RJ," Efficiently mining long patterns from databases", Proceedings of the ACM SIGMOD International Conference on Management of Data Seattle WA USA, ACM, 1998; 85–93.
9. Chen, M.S., J. Han and P.S. Yu," Data Mining: An Overview from a Database Perspective", IEEE Trans. Knowledge and Data Engg, 1996; 866-883.
10. Carmona-Saez P, Chagoyen M, Rodriguez A, *et al.*, "Integrated analysis of gene expression by association rules discovery", BMC Bioinformatics, 2006; 7:54.
11. Chang J W. Lee," A Sliding Window Method for Finding Recently Frequent Itemsets over Online Data Streams", Int. journal of Information Science and Engg., 2004; 20: 753-762.
12. Creighton C, Hanash S, "Mining gene expression databases for association rules", Bioinformatics, 2003; 19: 79–86.
13. Cai R, Hao Z, Wen W, *et al*.," Kernel based gene expression pattern discovery and its application on cancer classification", Neurocomputing, 2010; 73: 2562–70.
14. Franceschini A, Szklarczyk D, Frankild S, *et al*., STRING v 9.1: "protein-protein interaction networks with increased coverage and integration" Nucleic Acids Res 2012; 41: D808–15.
15. Giannella. C,J. Han, J. Pei, X. Yan P.S. Yu, "Mining Frequent patterns in data streams at multiple time granularities in Data Mining Next Generation Challenges", 2003.
16. Ghosh, A. and B. Nath, "Multi-objective rule mining using genetic algorithms", Information Sciences, 2004; 163: 123-133.
17. Gouda K, Zaki MJ, "Gen Max: an efficient algorithm for mining maximal frequent itemsets", Data Min Knowl Discov, 2005; 11: 223–42.
18. Han J, Pei J, Yin Y, *et al.*, "Mining frequent patterns without candidate generation: a frequent-pattern tree approach", Data Min Knowl Discovery" 2004; 8:53–87.
19. Han,. Pei J and Y. Yin," Mining Frequent Patterns without candidate generation ", Proceedings of the Conference on the Management of Data SIGMOD'00 2000.
20. He J, Hu H, Chen B, *et al.,"* Rule extraction from SVM for protein structure prediction Rule", 2008; 80: 227–52.
21. Jiawei Han, Hong Cheng, Dong Xin and Xifeng Yan, "Frequent pattern mining current status and future directions", Data Mining Knowledge Discovery, 2007, 15:55-86.
22. Jin, C, W. Qian, C. Sha, J.X. Yu, A. Zhou ," Dynamically Maintaining Frequent Items Over a Data Stream", In CIKM the International Conference on Information and Knowledge Management, 2003.
23. Koyuturk M, Kim Y, Subramaniam S, *et al*.," Detecting conserved interaction patterns in biological networks", Journal of Computational Biology, 2006; 13:1299-322.

24. Karpinets TV, Park BH, Uberbacher EC," Analyzing large biological datasets with association networks", Nucleic Acids Res, 2012; 40:e131.
25. Karabatak M, Ince MC," An expert system for detection of breast cancer based on association rules and neural network", Expert Syst Appl 2009; 36: 3465–9.
26. Kerana Hanirex. D, K.P. Kaliyamurthie," Finding the Dominating Amino Acids in Dengue Virus (Type-1) Study on mining frequent itemsets", Int. Journal of Pharama and Bio Sciences, 2013; 4(3): (B), 880 – 889.
27. Kerana Hanirex. D, K.P. Kaliyamurthie," An Adaptive Transaction Reduction Approach for Mining Frequent Itemsets: A Comparative Study on Dengue Virus Type1", Int. Journal of Pharma and Bio Sciences, 2015; 6(2): (B) 336-340.
28. Kerana Hanirex. D," An Efficient TDTR Algorithm for Mining Frequent Itemsets", International Journal of Electronics and Computer Science Engineering, 2012; V2(N1): 251-256.
29. Kerana Hanirex. D, K.P. Kaliyamurthie, A. Kumaravel," Analysis of Improved TDTR Algorithm for mining frequent itemsets using Dengue virus type1 Dataset: A combined approach", Int. Journal of Pharma and Bio Sciences, 2015; 6(2): (B) 228-295.
30. Kerana Hanirex. D, M.A. Dorai Rengaswamy," Efficient Algorithm for Mining Frequent Itemsets using Clustering techniques"; IJCSE, 2011; 3(3):1028- 1032.
31. Leung K-S, Wong K-C, Chan T-M, *et al.,* "Discovering protein DNA binding sequence patterns using association rule mining", Nucleic Acids Res, 2010, 38: 6324–37.
32. Manda P, Ozkan S, Wang H, *et al.,* "Cross-ontology multi- level association rule mining in the gene ontology"; PLoS One, 2012; 7: e47411.
33. Ma L, Assimes T, Asadi N, *et al.,"* An almost exhaustive search-based sequential permutation method for detecting epistasis in disease association studies", Genet Epidemiol, 2010; 34:434–43.
34. Pan F, Tung A, Cong G, *et al*., "COBBLER: combining column and row enumeration for closed pattern discovery", Proceedings of the 16th International Conference on Scientific and Statistical Database Management SSDBM, Santorini Island, Greece, Washington, DC: IEEE Computer Society 2004; 21–30.
35. Shiu SY, Jiang WR, Porterfield JS, Gould EA," Envelopeprotein sequences of dengue virus isolates TH-36 and TH-Sman and identification of a type specific genetic marker for dengue and tick borne flaviviruses", Journal of General Virology, 73; 207-212.
36. Soumadip Ghosh, Sushanta Biswas, Debasree Sarkar, Partha Pratim Sarkar "Mining Frequent Itemsets Using Genetic Algorithm", International Journal of Artificial Intelligence & Applications (IJAIA), 2010; 1(4): 133-143.
37. Singh Vijendra, Laxman Sahoo Kelkar Ashwini, "An Effective Clustering Algorithm for Data Mining", IEEE conference on Data Storage and Data Engineering (DSDE), 2010; 978-1-4244-5678-9.
38. Tan P-N, Kumar V, Srivastava J, "Selecting the right interestingness measure for association patterns", Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Edmonton Alberta Canada. New York ACM, 2002; 32–41.
39. Tweedie-Cullen RY, Brunner AM, Grossmann J, *et al*. "Identification of combinatorial patterns of post-translational modifications on individual histones in the mouse brain", 2012; PLoSOne; 7: e 36980.
40. Tuzhilin A, "Handling very large numbers of association rules in the analysis of microarray data", Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Edmonton Alberta Canada, New York, ACM, 2002; 23–6.
41. Van Leemput K, Verschoren A," Modeling networks as probabilistic sequences of frequent subgraphs", http://win.ua.ac.be/adrem/bibrem/pubs/MLSB08.pdf .
42. World Health Organization Dengue fever and dengue hemorrhagic fever; Geneva, www.who.int/csr/disease/ dengue /2009
43. Zaki M, Parthasarathy S, Ogihara M, Li W, "New algorithms for fast discovery of association rules", Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD), Newport Beach CA USA Palo Alto CA: AAAI Press, 1997; 283–6.

How to cite this article:

Hanirex DK, Thooyamani KP and Khanaa V: Performance of association rules for Dengue Virus Type1 amino acids using an Integration of Transaction Reduction and Random Sampling (TRRS) algorithm. Int J Pharm Sci Res 2017; 8(6): 2578-87.doi: 10.13040/IJPSR.0975-8232.8(6).2578-87.