Int J Pharm Bio Sci 2013 July; 4(3): (B) 880 - 889

Research Article **Bioinformatics**

International Journal of Pharma and Bio Sciences **ISSN 0975-6299**

FINDING THE DOMINATING AMINO ACIDS IN DENGUE VIRUS (TYPE-1) STUDY ON MINING FREQUENT ITEMSETS

D.KERANA HANIREX[*1] AND DR.K.P.KALIYAMURTHIE[2]

[1]Department of CSE, BharathUniversity, Chennai, India
[2]Department of IT, BharathUniversity, Chennai, India

ABSTRACT

Research on biological science is one of the emerging task. One approach is finding the hidden patterns and association rules over the patterns. This proposed system finds the most dominating amino acids in dengue virus type-1 (DEN-1) . There are no specific vaccines and drugs are predicted for dengue virus yet. This system finds the dominating amino acids among the infected protein sequences which causes infections in human. The dominating amino acids have been identified using Apriori Algorithm in data mining. We found that association rule reveals the association between the protein and protein sequence. This is a novel attempt which finds the most informative association rules using Apriori algorithm implemented through Java. The generated information can be useful to the drug designers. Our findings reveals that Leucine(L), Phenylalanine (F),Lysine(K),Serine(S) and Glycine(G) are the dominating amino acids in Dengue Virus Type-1.Hence this system can be a potential candidates for the drug designers to develop the antibiotics for the dengue fever.

KEYWORDS : Bio-data mining, Association Rule Mining, Dengue Virus (DEN-1), Apriori Algorithm, Amino Acids, Protein sequence.

D.KERANA HANIREX
Department of CSE,BharathUniversity,Chennai,India

INTRODUCTION

Data mining is the extraction of finding the hidden information from the large databases. It is also known as Knowledge Data Discovery (KDD) which is a new technology with great potential to help industries to focus on the most important information from their own data bases. Data mining tools predict future trends and behaviours, which allow industries and companies to make knowledge-driven decisions. Data mining techniques can be applied on existing software to enhance the existing resources, to bring the efficient product for the social use. The problem of mining frequent itemset was first proposed by Agarwal [1] in 1993. Data mining tools can analyze the massive databases. Data mining techniques are the result of a long process of research and product development. Data mining can be defined as the process of finding meaningful correlations and patterns by searching large amounts of data which is stored in data warehouses[2].The various data mining tasks are classification[3], clustering[4], Association mining [5,6].

(i) Dengue Viruses

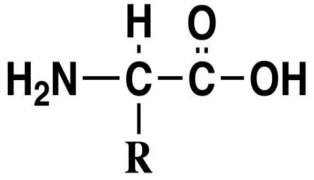
Dengue virus (DENV), a widespread arthropod-borne virus that affects humans, which belongs to the family Flaviviridae, genus Flavivirus. There are distinct serotypes namely (DEN 1–4)[7]. DENV is most occur in tropical and subtropical areas. By current estimates, the impact of DENV infections on human health is enormous; Dengue is caused by any one of four related viruses which is transmitted by mosquitoes. There are no specific vaccines or antibiotics found yet to prevent the infection with dengue virus (DENV). This leads to infection sequential to people at greater risk for dengue hemorrhaghic fever (DHF) and dengue shock syndrome (DSS). DENV genome is a single-stranded positive-sense RNA of approximately 11,000 nucleotides that encodes 3 structural proteins (capsid,membrane and envelope) and 7 nonstructural proteins (NS1,NS2A,NS2NB,NS4A,NS4B and NS5).As per World Health Organization report, nearly 3 billion persons living among all over the countries are now at a greater risk of being infected by at least once annually by 1 of the 4 DENV serotypes [8]. By considering these issues, this proposed system found the most dominating amino acids in the dengue virus.

(ii) Amino Acids

Amino acids are biologically important organic compounds made from amine ($-NH_2$) and carboxylic acid (-COOH) functional groups. Carbon, hydrogen, oxygen, and nitrogen are the major elements of an amino acid .

General Structure of Amino Acid

Figure1
Structure of Amino Acid



C-Carbon,O-Oxygen,H-Hydrogen,N-Nitrogen

Amino acids perform critical roles in processes such as neurotransmitter transport and biosynthesis. 9 of the 20 standard amino acids are called "essential" for humans because they cannot be created from other compounds. It can be taken through food. Amino acids are important in nutrition and are commonly used in nutritional supplements. Industrial uses include

the production of drug. Amino acids are the basic structural units which form proteins. They join together to form short polymer chains called peptides or longer chains called either polypeptides or proteins. The list of amino acids are given in the following table. It may be essential or nonessential amino acids.

Table 1
List of Amino Acids

One letter code	Three letter code	Amino acid	Possible codons
A	Ala	Alanine	GCA, GCC, GCG, GCT
B	Asx	Asparagine or Aspartic acid	AAC, AAT, GAC, GAT
C	Cys	Cysteine	TGC, TGT
D	Asp	Aspartic acid	GAC, GAT
E	Glu	Glutamic acid	GAA, GAG
F	Phe	Phenylalanine	TTC, TTT
G	Gly	Glycine	GGA, GGC, GGG, GGT
H	His	Histidine	CAC, CAT
I	Ile	Isoleucine	ATA, ATC, ATT
K	Lys	Lysine	AAA, AAG
L	Leu	Leucine	CTA, CTC,TTA, TTG, CTG, CTT
M	Met	Methionine	ATG
N	Asn	Asparagine	AAC, AAT
P	Pro	Proline	CCA, CCC, CCG, CCT
Q	Gln	Glutamine	CAA, CAG
R	Arg	Arginine	AGA, CGG, CGT AGG, CGA, CGC,
S	Ser	Serine	AGC, AGT, TCA, TCG, TCT,TCC
T	Thr	Threonine	ACA, ACC, ACG, ACT
V	Val	Valine	GTA, GTC, GTG, GTT
W	Trp	Tryptophan	TGG
X	X	any codon	NNN
Y	Tyr	Tyrosine	TAC, TAT
Z	Glx	Glutamine or Glutamic acid	GAA, GAG, CAA, CAG,
*	*	stop codon	TAA, TAG, TGA

DATA PREPROCESSING

Data need to be processed in order to improve the quality of the data .The various tasks of data mining are data Integration, Data transformation, Data discretization ,Data cleaning and Data reduction. Data Cleaning involves removing incomplete data, noisy data, inconsistent data and intentional data. Data integration combines data from multiple sources. Data Transformation task contains Data Smoothing, Aggregation, Generalization and Normalization. Data reduction strategies include data aggregation, high dimensionality reduction, data compression and discretization.

BIO-DATA MINING

Bio-data mining[9] is the development of mining activities using bio-medical data, or bio-technology information using computer based technology. Now a days bio-data analysis methods and bio-technology leads to the emergent research area called bio-data mining. The biological activity of the protein is determined by the amino acids in proteins[10].It is the combination of 20 amino acids during biosynthesis.

(i) Proteins &Protein Sequence

Proteins are very important molecules in cells. They are involved in all cell functions. Proteins are vary in structure as well as function. Each type of protein has its own unique three-dimensional shape.. If it does not fold up properly it cannot carry out its function in the cell. A protein is held in its correct shape by forming the bonds between amino acids. If the amino acid is not present correctly

the bond cannot be formed and so the protein cannot take up its correct shape and carry out its function[11].Protein database is a collection of sequences from several sources including GenBank as well as records from PDB. Protein sequences are the fundamental determinants of biological structure and function. They are constructed from 20 amino acids. This proposed system deals with Dengue virus type 1 datasets which is taken from NCBI (National Centre for Biotechnology Information) . This system takes polyprotein datasets [Dengue virus 1] from GenBank:AAB27904.1 which consists of 777 amino acids. This system finds the most dominating amino acids among the dengue virus which can be used to generate drugs for the drug designers.

ASSOCIATION MINING

Mining association rule is one of the recent data mining research. Association rules are used to show the relationships among the data items in the database. Association rules are frequently used in various applications such as marketing, advertising and inventory control. This problem is initiated by applications known as market basket analysis to find the relationship between items purchased by customers [3], that is, what kinds of products tend to be purchased together.A database in which an association rule is to be found is viewed as a set of tuples or records, where each tuple contain a set of items further, each item represents an item purchased at a moment. Each tuple is the list of items at a given time. The support(s) of an item is the percentage of transactions in which that item occurs. Given a set of items $I=\{I_1,I_2,\ldots.In\}$ and a database transactions $D=\{t_1,t_2\ldots\ldots t_m\}$ where $t_i=\{I_{i1},I_{i2}\ldots.I_{in}\}$ and $Iij \in I$, an association rule is an implication of the form $X \Rightarrow Y$ where X, $Y \subset I$ are sets of items called itemsets and $X \cap Y = \Phi$. The confidence or strength (α) for an association rule $X \Rightarrow Y$ is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain X.The association rule problem is to identify all association rules with a minimum support and confidence what we assigned earlier. Efficiency of an association rule algorithms can be measured with respect to the number of scans of the database that are required and the maximum number of item sets that must be counted [14]. One of the common approach to find association rules is to break up the problem into 2 parts

1.Find Large Itemsets

2.Generate rule from the frequent Itemsets

A Large (Frequent) Item set is an Itemset whose number of occurrence is above the threshold (s). The Apriori Algorithm is the most well known association rule algorithm and it is

used in most commercial products. This algorithm is based on largest item set property which states that "Any subset of a large item set must be large". The basic idea of Apriori algorithm is to generate item sets of a particular size and then scan the database to count these to see if they are large. Only those candidates that are large are used to generate candidates for the next scan. L_i is used to generate next C_{i+1}. L represent Large Itemset. C represents candidate items. All single item sets are used as candidates for the first pass. The large item sets of the previous pass,L_{i-1} is joined with to determine the candidates. Individual item sets must have all but one item in common in order to be combined.

Steps involved in Apriori algorithm

Input : Input sequence from polyprotein datasets [Dengue virus 1] from GenBank:AAB27904.1

Output: Set of frequent patterns

METHOD

Step 1

1.find all frequent item sets

2.Get frequent items whose occurrence greater than or equal to the minimum support

3.Generate candidate item sets and prune the results

Step 2

Generate association rules which satisfy min.support and min.confidence

The following figure describes the working principle of Apriori algorithm

Working Principle of Apriori Algorithm

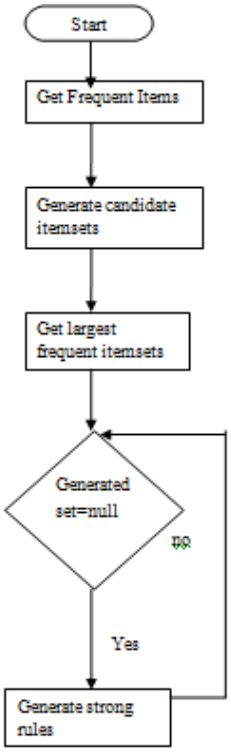


Figure2
Apriori Algorithm

EXPERIMENTAL ANALYSIS

Sample sequential datasets for Dengue virus 1(DEN1): GenBank:AAB2 7904.1 which consists of 777 amino acids

Table 2
Sample datasets

S.NO	List of amino acids											
T12	M	N	N	Q	R	K	K	T	G	R	P	S
T24	F	N	M	L	K	R	A	R	N	R	V	S
T36	T	G	S	Q	L	A	K	R	F	S	K	G
T48	L	L	S	G	Q	G	P	M	K	L	V	M
T60	A	F	I	A	F	L	R	F	L	A	I	P
T72	P	T	A	G	I	L	A	R	W	S	S	F
T84	K	K	N	G	A	I	K	V	L	R	G	F
T96	K	K	E	I	S	S	M	L	N	I	M	N
T108	R	R	K	R	S	V	T	M	L	L	M	L
T120	L	P	T	A	L	A	F	H	L	T	T	R

Here we have taken 12sequences for each transaction which leads to a total of T777 sequences. If we take the minimum support as 10% by varying the confidence from 90% to 50% we are getting the number of association rules which is given in the following table.

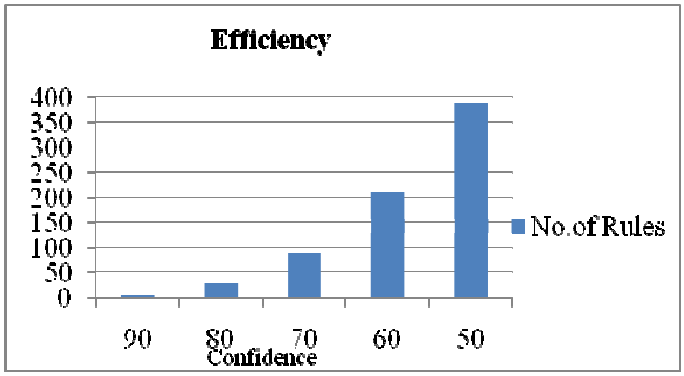
Table 3
No. of rules generated for different confidence value

Confidence	No.of Rules
90	5
80	30
70	87
60	210
50	388
Total number of rules generated	1246

The following graph represents the relationship between the various confidence and the number of association rules generated.

Relationship between various confidence and the number of rules generated

Figure 3
Efficiency of a graph



The above diagram shows that when the confidence value get increased ,the number of rule generated get decreased. This proposed system finds the largest itemsets of size1,2,3and 4. The number of itemsets in largest itemset of size 1 L(1) is 20,L(2) is 121 and L(3) is 123 and L(4) is 19.

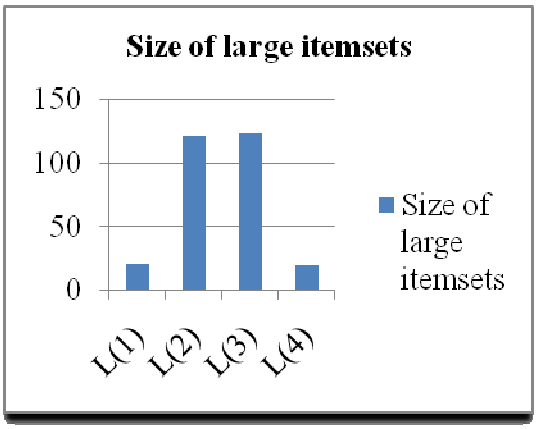
Table 4
Size of large itemsets

large itemsets	Number of itemsets
L(1)	20
L(2)	121
L(3)	123
L(4)	19

The following graph shows the size of 1-itemsetL(1) ,2-itemsetsL(2),3_itemsets L(3),4 itemsets L(4) .

Number of itemsets generated

Figure 4
Number of Large itemsets



The following table describes the number of rules generated by varying the confidence from 90% to 50% as well as varying the support from 10% to 40%

Table 5
Number of rules generated by varying min.confidence and min.support value

No. Of Rules generated				
Confidence	Support			
	40	30	20	10
80	1	2	2	30
70	2	3	13	85
60	6	13	39	29
50	6	22	74	388

The accuracy of the system is measured by the time it takes to find the association rule. The following table shows the time taken in seconds for different confidence measures.

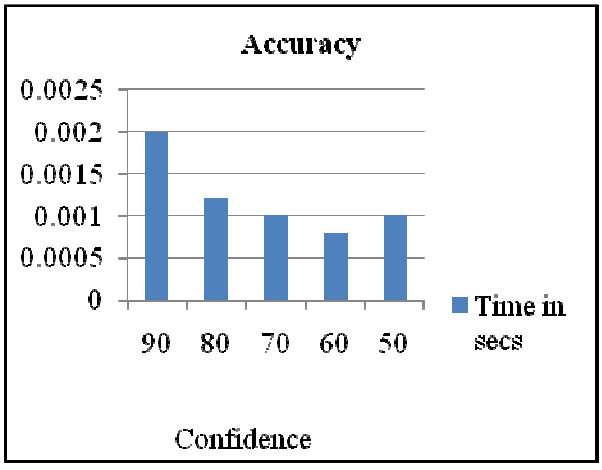
Table 6
Time taken for different confidence value

confidence	Time taken (seconds)
90	.0020
80	.0012
70	.0010
60	.0008
50	.0010

The following diagram shows the relationship between the various confidence and the time in seconds. The Graph shows that the time taken varies from .001 to .002 seconds or 1 to 2 milliseconds for this dengue virus datasets.

Time measure in secs for different confidence value

Figure 5
Accuracy(Time in seconds)



The resulting association rules are described as follows. The following table depicts the existence of strong association among the amino acids which causes dengue virus type 1(DEN1) .In this case we have to take the strong association rules with minimum confidence 90%.

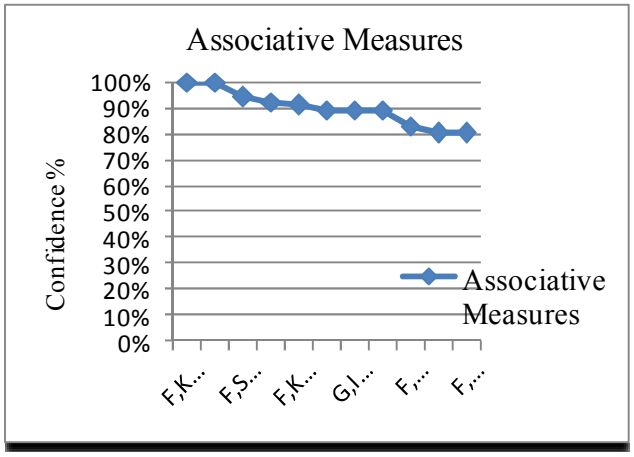
Table 7
Comparison of association rule with minimum confidence 90%

Analysis	Confidence (c)	Compare c with minimum confidence	Result
F,K,L=>G	100%	100%>90%	**Accepted****
F,L,S=>G	100%	100%>90%	**Accepted****
F,S=>G	94%	94%>90%	**Accepted****
F,K=>G	92%	92%>90%	**Accepted****
F,K,S=>G	91%	91%>90%	**Accepted****
A,Q,S=>G	89%	89% <90%	Rejected
G,I,L=>T	89%	89%<90%	Rejected
A,Q,S=>G	89%	89% <90%	Rejected
G,I,L=>T	89%	89%<90%	Rejected
Q,S,T=>G	89%	89%<90%	Rejected
F,G,K=>S	83%	83%<90%	Rejected
A,E=>T	80%	80%<90%	Rejected
F,R=>L	80%	80%<90%	Rejected
D,V=>T	79%	79%<90%	Rejected
A,P=>T	77%	77%<90%	Rejected

Hence from the above table it is very clear that Leucine(L), Phenylalanine(F), Lysine(K), Serine(S) is strongly associated with Glycine(G).The following graph shows the associative measures for different confidence value.

Associative measures for different confidence value

Figure 6
Association measures



CONCLUSION

This proposed system focuses on finding the dominating amino acids which causes dengue fever(DENV-1).The protein sequence named polyprotein for dengue virus-1 is taken from GenBankAAB27904.1 for analysis. From this real dataset, the frequent itemsets were

generated ,in which few amino acids were strongly associated. The dominating amino acids is more beneficial in finding the medicines or drugs to cure the disease caused by the cells. In future this work can be extended to other protein sequences which causes various diseases in human. This system is more beneficial in finding the medicines for the Dengue Fever and Dengue hemorrhaghic fever for which there is no proper medicines at present

.

REFERENCES

1. R. Agrawal and R. Srikant, Fast Algorithms for Mining Association Rules, in Proceedings of the 20th VLDB Conference, 1994, pp. 487−499.
2. Agarwal R., Imielinski T., and Swami A. Mining associations between sets of items in massive databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, Washington D.C. , May 1993, pp. 207-216.
3. JiaweiHan,HongCheng,DongXin and XifengYan,"Frequent pattern mining current status and future directions",Data Mining Knowledge Discovery(2007) 15:55-86
4. Arun K Pujari. Data Mining Techniques 5th ed., Universities Press (India) Private Limited, 2003.
5. Margaret H. Dunham. Data Mining, Introductory and Advanced Topics Pearson Education Inc., 2003.
6. M. Houtsma and A. Swami. Set Oriented Mining for Association Rules in Relational Databases. In Proceedings of 11th International conference on Data Engineering, 1995, pp 25-33,.
7. Stephen Y.W.Shiu,WenR.Jiang,JamesS.Porterfield and Ernest A.Gould,"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
8. World Health Organization. Dengue fever and dengue hemorrhagic fever. Geneva: the Organizationwww.who.int/csr/disease/dengue/; 2009.
9. Anandhavalli,IACSIT,IAENG,Ghose,Gauthaman(2010),"Association Rule Mining in Genomics",International Journal of Computer Theory and Engineering,Vol2,No.2.
10. [Phylogeography of Dengue Virus Serotype 4,Brazil,2010-2011, EID Journl,vol18,November,2012
11. Mohammed J Zaki, George Karypis and JiongYang ,Data Mining in Bioinformatics (BIOKDD),Algorithms for Molecular Biology, April 2007,2:4
12. Tasneemsandozi,vamsi Krishna Emani," Survey Of Prescription Pattern Of Anti-Hypertensive DrugsIn Hypertensives& Hypertension Associated Diabetics"IJPBS,Vol.1,Issue-4,Oct-Dec.2010,p23-26.
13. T. V. Sathe,"Ecology of MosquitoesfromKolhapur District in India",IJPBS,Vol2,Issue 4,Dec 2011,p103-111.
14. D.KeranaHanirex,Dr.M.ADorai Rangaswamy,"Efficient Algorithm for Mining Frequent Itemsets using Clustering Techniques",IJCSE, Vol. 3, No. 3 Mar 2011,p1028-1032