IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS)
e-ISSN: 2278-3008, p-ISSN:2319-7676. Volume 9, Issue 2 Ver. II (Mar-Apr. 2014), PP 122-132
www.iosrjournals.org

Mining quantitative association in peptide Sequences of Flavivirus Subfamilies

Priyanka Rajput and Dr. Usha Chouhan
Department of Bioinformatics.Manit,Bhopal,India

Abstract: *Flavivirus, one of the most important viral pathogen which is also called from the yellow fever virus(flaivirus :yellow in latin),belongs to the family Flaviviridae. The study of relationships among amino acid and other parameters in molecular sequences of these virus are crucial for understanding mechanism of pathogenicity, infection ,regulation and control of the disease. In view of above an attempt has been made to develop a model for mining quantitative association patterns in the amino acid sequence of flavivirus subfamily family. The sequences have been taken form online database "NCBI" .A model is propesed to generate the quantitative association relationships for 12sub-families flavivirus .The results generated have been analyzed for similarities and differences in associations among amino-acid of this family. The association rules have been generated for redundant and non-redundant protein sequences using frequent and in-frequent patterns.*
Key words:-dataset, item set, Threshold, Support, Confidence, Pattern , quantitative association mining.

I. Introduction

The flavivirus family contains many viral agents which produces encephalitis. Flavivirus encephalitis are either mosquito- borne, tick-borne, or have an unknown vector[4]. The viruses of the family Flaviviridae are important arthropod-borne viruses in both human and veterinary medicine. They are transmitted by mosquito and ticks and usually are maintained in a transmission cycle in nature. They are widely distributed throughout the world with the exception of the polar region, although a specific flavivirus may be geographically restricted to a continent or a particular part. They produce a broad spectrum of clinical responses in humans ranging from asymptomatic infection to fulminant encephalitis or hemorrhagic fever. Nearly 60 flaviviruses are known to exist but many are yet to be shown to cause disease in humans[5].Major symptoms include mild acute febrile syndromes, severe neurological, hepatic and/or hemorrhagic disease. The geographical diversity of flavivirus has shown the occurrence of JEV in Asia, causing menigo encephalitis in children. [6], and WNV in west Africa, middle east,and from 1999 in North America[7] .DENV shows worldwide existence ,affecting 2.5 billion[8]. YFV causes serious infections manifested by fulminant hepatitis and severe hemorrhagic disease. YFV still kills a considerable number of people annually, despite the availability of an effective vaccine [9].

In 1993, Agrawal proposed an algorithm for extracting association rules from large databases [10]. Since then, association rule mining has become one of the main techniques for Knowledge Discovery in Databases (KDD).A good number of algorithms are reported in the literature [11-17] for association rule mining

Association analysis has proved to be a powerful approach for analyzing traditional market basket data, and has even been found useful for some problems in bioinformatics in a few instances. However, there are a number of other important problems in bioinformatics, such as finding biomarkers using dense data like SNP data and real-valued data like gene-expression data, where such techniques could prove to be very useful, but cannot currently be easily and effectively applied[18].

The problem of mining association rules in large relational table are introduced by Srikant R and agrawal [19].This technique can generate too many similar rules. They gave an algorithm for mining quantitative association rules.Attempts are also reported in the literature [21-24]for mining associations in molecular sequences.

The quantitative association rules approach read the nature of different amino acid that are present in the protein. This very basic analysis provides understandings into the Co-occurrence of certain amino acids in a protein. Such association rules are desirable for enhancing our understanding of protein composition.[20]

The peptide sequence contains lot of information about the various features and characteristics of the organism which needs to be explored by various techniques for better understanding of function and processes of the organism .In view of above a model for mining association relations in 12 sub families of flavivirus is developed . The similarity and differences in association in amino-acid of these families have been analyzed and the association rules have been generated for redundant and non-redundant protein sequences using frequent and in-frequent patterns.

II. Materials And Methods

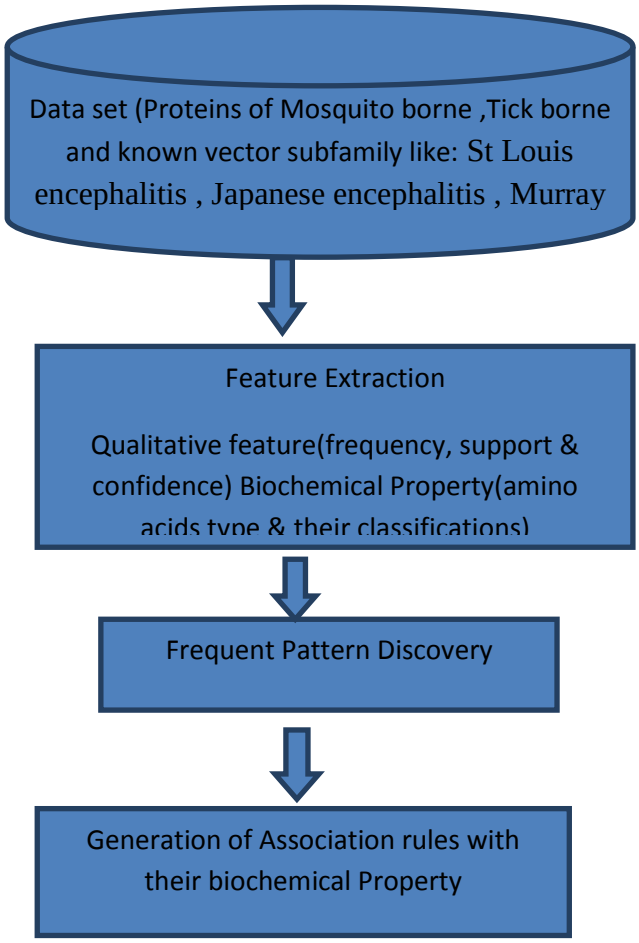
To perform the analysis of Mosquito borne the first step is to collect data of molecular sequences of sub family like :- Mosquito borne ,Tick borne and known vector from NCBI database.

For studying the frequent patterns in redundant and non-redundant sequences frequency can be calculated as

$\mu_i(A) = \sum_i^n f_i(A)$ ---(1)

where $\sum_i^n f_i(A)$ is the sum of frequency of amino acid and $\mu_i(A)$ is the frequency of amino acid .
A is the i^{th} sequence.
In this paper we assume that threshold can be defined by user
The threshold is given by

$$\frac{\sum_{i=1}^{n} \max(ai) + \sum_{i=1}^{n} \min(ai)}{2} \quad \text{---- (2)}$$



Where a is amino acid and i is varies from 1 to 20 Flow chart[2]: Showing the Methodology
n is no. of sequence.
The apriori algorithm is employed to find frequent patterns in all the sequences. These patterns are used to generate ordinary association rule.
The frequency Support for n amino acid can be calculated as:

$\sum \mu_i(A1 \cap A2 \cap A3 \cap \ldots\ldots\ldots\ldots A_{n-1} \cap A_n)$ ---- (3)

Confidence for n amino acids can be calculated by;

$$\frac{\sum_i^L \mu i\,(A1 \cap A2 \cap A3 \cap \ldots\ldots\ldots\ldots An-1 \cap An)}{\sum_i^L \mu i\,(A1 \cap A2 \cap A3 \cap \ldots\ldots\ldots\ldots An-1)} \qquad \text{---- (4)}$$

III. Results And Discussion

After applying the apriori algorithms on flavivirus subfamilies datasets, it has been found that there are variations in frequent amino acid patterns for redundant and non-redundant data set. In mosquito borne subfamilies for redundant dataset, for Japanese encephalitis **G has high frequency**(118562.0) whereas **C has low frequency**.(26913.0), for Marry valley encephalitis **L has high frequency** (4083.0) along with **C has low frequency**(841.0), for St. Louis encephalitis **G has high frequency** (19961.0) along with **H has low frequency**(4264)) ,for illhues **W has high frequency** (3620.0) along with **C has low frequency**(524.0), for West Nile **L has high frequency**(297889.0) along with **C** has low frequency(63989.0) and for non-redundant dataset, for Japanese encephalitis **G** has high frequency.(82991.0)) along with **C has low frequency**(17496.0), for Marry valley encephalitis **L has high frequency** (2905.0) along with **Chas low frequency**(602), for St. Louis encephalitis **Ghas high frequency** (15593.0) along with **H** has low frequency.(3347.0), for illhues **W has high frequency**(2043.0) along with **C has low frequency**(367.0), for West Nile **Lhas high frequency**(196577.0)) along with **C** has low frequency.(40681.0)).

In tick borne subfamilies for redundant and non-redundant dataset, for Central European **L has high frequency**(181311.0) along with **W has low frequency** (27134.0), for Louping ill **G has high frequency**(2555.0)) along with **C has low frequency** (533.0), for Powassan **G has high frequency** (7969.0) along with **C has low frequency**(1780.0), for Russion spring-rodent **V has high frequency**(5529.0) along with **Mhas low frequency**(508.0), for Summar encephalitis **G has high frequency**(89.0) along with **C,Y are low frequency**(L.F.(19.0)).

In Known vectors for redundant dataset, for Ricio **L has high frequency**(13898.0)along with **W has low frequency** (2254.0) , for Rio bravo **L has high frequency**.(2202.0)) along with **C has low frequency** (373.0)) and for non-redundant dataset , for Ricio **L has high frequency**(12700.0)) along with **W has low frequency**(2072.0) , for Rio bravo **L has high frequency**(1482.0) along with **C has low frequency** (278.0)) mosquito borne subfamilies, G(glycine) and L(luecine) are most frequency amino acid and histidine(H) and cysteine(C) are least frequent amino for redundant and non- redundant dataset.

Frequency of amino acid depends on the total number of sequences for example west Nile virus and central european have lager number of sequences than other viruses so amino acid frequency of west Nile virus and centeral european are higher than other subfamilies.

The figures 1,2 and 3 show the frequency of amino acids for various subfamilies of flavi virus.

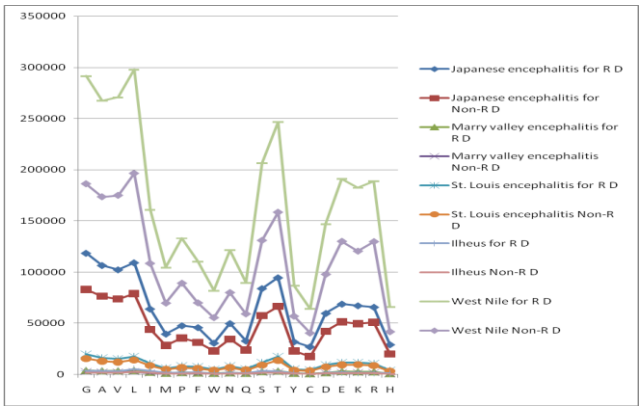


Figure1: Variation in Frequency of Redundant and Non-redundant amino acid Mosquito Subfamily

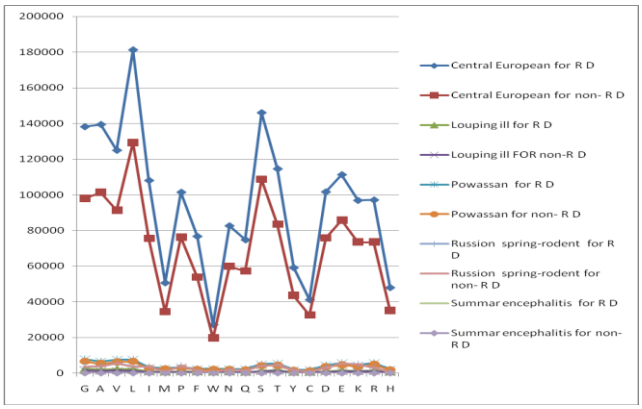


Figure2: Variation in Frequency of Redundant and Non-redundant amino acid of Tick borne subfamily

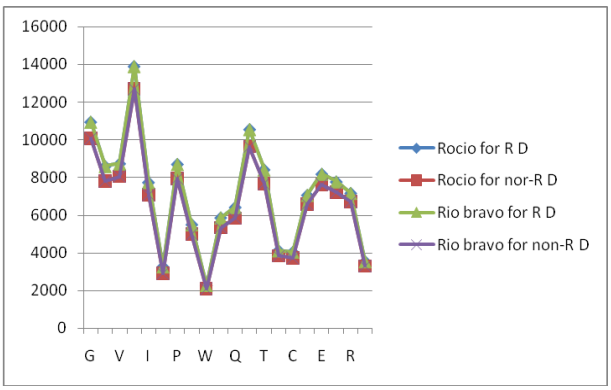


Figure3: Variation in Frequency of Redundant and Non-redundant amino acid Known Vector Subfamily

Table-1 Maximum support in case of Mosquito Borne

Japanese encephalitis		Marry vally encephalitis		St. Louis encephalitis		Ilheus		West Nile	
Redundant	Non-redundant	Redundant	Non-redundant	Redundant	Non-redundant	Redundant	Non-redundant	Redundant	Non-redundant
SUPPORT OF SIX FREQUENT PATTERN [G, A, L, T, V, S]	SUPPORT OF SIX FREQUENT PATTERN [E, G, A, L, T, V, S]	SUPPORT OF EIGHT FREQUENT PATTERN [E, G, A, L, K, T, V, S]	SUPPORT OF SEVEN FREQUENT PATTERN[E, G, A, L, K, T, V, S, R]	SUPPORT OF FIVE FREQUENT PATTERN [G, A, L, T, V]	SUPPORT OF FIVE FREQUENT PATTERN [G, A, L, T, V]	SUPPORT OF THREE FREQUENT PATTERN [G, A, L, I, T, S]	SUPPORT OF THREE FREQUENT PATTERN [G, A, L, I, T, V, S]	SUPPORT OF SIX FREQUENT PATTERN[E, G, A, L, T, V, S, R]	SUPPORT OF SEVEN FREQUENT PATTERN [E, G, A, L, T, V, S, R]
GALTVS 79153.0	EGALTV 50686.0	EGALKTVS SUPPORT = 2490.0	EGALTVS SUPPORT = 1812.0	GALTV SUPPORT = 15208.0	GALTV SUPPORT = 12020.0	GAL SUPPORT = 3084.0	GAL SUPPORT = 1737.0	EGALTV SUPPORT = 185116.0	EGALTVR SUPPORT = 123574.0
TOTAL NO OF SIX FREQUENT PATTERN: 1	GALTVS 55074.0	TOTAL NO OF EIGHT FREQUENT PATTERN: 1	GALKTVS SUPPORT = 1765.0	TOTAL NO OF FIVE FREQUENT PATTERN: 1	TOTAL NO OF FIVE FREQUENT PATTERN: 1	GLS SUPPORT = 2999.0	GLV SUPPORT = 1582.0	GALTVS SUPPORT = 195490.0	TOTAL NO OF SEVEN FREQUENT PATTERN: 1
	TOTAL NO OF SIX FREQU		TOTAL NO OF SEVEN FREQUE	SUPPORT OF FIVE FREQUE		ALT SUPPORT = 2955.0	GLS SUPPORT = 1595.0	GALTVR SUPPORT =	

	ENT PATTERN: 2		NT PATTERN: 2	NT PATTERN [G, A, L, T, V]				184127.0	
				GALTV SUPPORT = 15208.0		LIS SUPPORT = 3061.0	ALT SUPPORT = 1586.0	TOTAL NO OF SIX FREQUENT PATTERN: 3	
						TOTAL NO OF THREE FREQUENT PATTERN: 4	TOTAL NO OF THREE FREQUENT PATTERN: 4		

Table-2 Maximum support in case of Tick Borne

Central European		Louping ill		Powassan		Russion spring-rodent		Summar encephalitis	
Redundant	Non-redundant	Redundant	Non-redundant	Redundant	Non-redundant	Redundant	Non-redundant	Redundant	Non-redundant
SUPPORT OF FOUR FREQUENT PATTERN [G, A, L, T, V, S]	SUPPORT OF FOUR FREQUENT PATTERN [E, G, A, L, T, V, S]	SUPPORT OF FIVE FREQUENT PATTERN [G, A, L, T, V]	SUPPORT OF FIVE FREQUENT PATTERN [G, A, L, T, V]	SUPPORT OF SIX FREQUENT PATTERN [E, G, A, L, T, V, S, R]	SUPPORT OF SIX FREQUENT PATTERN [E, G, A, L, T, V, S, R]	SUPPORT OF TEN FREQUENT PATTERN [E, G, A, L, I, K, T, V, P, S]	SUPPORT OF TEN FREQUENT PATTERN [E, G, A, L, I, K, T, V, P, S]	SUPPORT OF FOUR FREQUENT PATTERN [G, A, L, T, V]	SUPPORT OF FOUR FREQUENT PATTERN [G, A, L, T, V]
GALV SUPPORT = 106432.0	GALV SUPPORT = 77644.0	GALTV SUPPORT = 1783.0	GALTV SUPPORT = 1092.0	EGALTV SUPPORT = 4971.0	EGALTV SUPPORT = 4118.0	EGALIKTVPS SUPPORT = 3593.0	EGALIKTVPS SUPPORT = 3591.0	GALV SUPPORT = 61.0	GALV SUPPORT = 61.0
GALS SUPPORT = 105082.0	GALS SUPPORT = 76939.0	TOTAL NO OF FIVE FREQUENT PATTERN: 1	TOTAL NO OF FIVE FREQUENT PATTERN: 1	EGALVS SUPPORT = 4928.0	EGALVS SUPPORT = 4097.0	TOTAL NO OF TEN FREQUENT PATTERN: 1	TOTAL NO OF TEN FREQUENT PATTERN: 1	GLTV SUPPORT = 59.0	GLTV SUPPORT = 59.0
TOTAL NO OF FOUR FREQUENT PATTERN: 2	GLVS SUPPORT = 75114.0			EGALVR SUPPORT = 4989.0	EGALVR SUPPORT = 4280.0			TOTAL NO OF FOUR FREQUENT PATTERN: 2	TOTAL NO OF FOUR FREQUENT PATTERN: 2
	ALVS SUPPORT = 76605.0			EGATVS SUPPORT = 4876.0	GALTVS SUPPORT = 4040.0				
	TOTAL NO OF FOUR FREQUENT PATTERN: 4			GALTVS SUPPORT = 4886.0	TOTAL NO OF SIX FREQUENT PATTERN: 4				
				TOTAL NO OF SIX					

				FREQUENT PATTERN: 5					

Table-3 Maximum support in case of Known Vector

Rocio		Rio bravo	
Redundant	Non-redundant	Redundant	Non-redundant
SUPPORT OF THREE FREQUENT PATTERN [G, A, L, T, V, P, S]	SUPPORT OF THREE FREQUENT PATTERN [E, G, A, L, T, V, P, S]	SUPPORT OF THREE FREQUENT PATTERN [G, A, L, V, S]	SUPPORT OF THREE FREQUENT PATTERN [G, L, T, V, S]
GLT SUPPORT = 8132.0	GLT SUPPORT = 7407.0	GLV SUPPORT = 1344.0	GLV SUPPORT = 919.0
GLV SUPPORT = 8237.0	GLV SUPPORT = 7576.0	GLS SUPPORT = 1288.0	GLS SUPPORT = 885.0
GLP SUPPORT = 8190.0	GLP SUPPORT = 7511.0	TOTAL NO OF THREE FREQUENT PATTERN: 2	TOTAL NO OF THREE FREQUENT PATTERN: 2
GLS SUPPORT = 8440.0	GLS SUPPORT = 7699.0		
TOTAL NO OF THREE FREQUENT PATTERN: 4	TOTAL NO OF THREE FREQUENT PATTERN: 4		

The Tables 1,2 and 3 show maximum support for k-frequent amino acid in case of Mosquito borne ,Tick borne and known vector sub families respectively. Based on the maximum support for frequent amino acid among the subfamilies of mosquito borne,tick borne and known vector family of Flavivirus it is found that **A,G,L ,T,** and **V** (Alanine, Glycine, Leucine, Threonine & Valine)are frequent in all subfamilies except of Rio bravo(known vector subfamilies). **S(**Serine**)** is also frequent in all subfamilies except St. Louis encephalitis's , Louping ill and Summar encephalitis .E(Glutamate) is frequent in Japanese encephalitis(redundant dataset),st.Louis encephalitis(redundant and non-redundant dataset) and ilhues, west nile,central European(non-redundant),powassan, russion spring rodent and Rocio(non-redundant).R(Arginine) is frequent for marry vally (non-redundant dataset),west nile and powassan.K(Lysine) is frequent for Marry vally , Russion spring-roden. I(Isoleucine) is frequent for illeus , Russion spring-rodent. P(Proline) is frequent for Russion spring rodent ,Rocio subfamilies of Flavi virus.

Table 4 –Probable Structures and physicochemical Prosperities of Protein Sequences of Sub families

Subfamily	Redundant			Non-redundant		
	Frequent Amino Acid	Probable Structure	Physicochemical Properties	Frequent Amino Acid	Probable Structure	Physicochemical Properties
Mosquito Borne Subfamilies						
Japanese encephalitis	G, A, L, T, V, S	Helix, Sheet and Coil	Polar aliphatic (G), polar uncharged(S, T),non-polar aliphatic(A,L,V) and hydrophobic(G,L,V), CBetaBranched(T,V)	E, G, A, L, T, V, S	Helix	Acidic Negative charged protein stable (E), Polar aliphatic (G),polar uncharged(S,T),non-polar aliphatic(A,L,V) and hydrophobic(G,L,V), CBetaBranched(T,V)
Marry vally encephalitis	E, G, A, L, K, T, V, S	Helix	Acidic Negative charged protein stable(E), Polar aliphatic (G), non-polar aliphatic(A,L,V), and polar uncharged(T,S),basic charged(K) and hydrophobic(G,L,V), CBetaBranched(T,V)	E, G, A, L, K, T, V, S, R	Helix	Acidic Negative charged protein stable (E), Polar aliphatic hydrophobic (G), non-polar aliphatic(A,L,V), polar uncharged(T,S),basic charged(K,R) and hydrophobic(G,L,V), CBetaBranched(T,V)
St. Louis encephalitis	G, A, L, T, V	Helix & Sheet	Polar aliphatic (G),polar uncharged(T),non-polar aliphatic(A,L,V) and hydrophobic(G,L,V),	G, A, L, T, V	Helix	Polar aliphatic (G),polar uncharged(T),non-polar aliphatic(A,L,V) and hydrophobic(G,L,V),

			CBetaBranched(T,V)			CBetaBranched(T,V)
Ilheus	G, A, L, I, T, S	Sheet	Polar aliphatic (G),polar uncharged(S, T),non-polar aliphatic(A,L,I) and hydrophobic(G,L) CBetaBranched(T)	G, A, L, I, T, V, S	Sheet	Polar aliphatic (G),polar uncharged(S, T),non-polar aliphatic(A,L,I,V) and hydrophobic(G,L,V), CBetaBranched(T,V)
Tick Borne subfamilies						
Central European	G, A, L, V, S	Helix And coil	Polar aliphatic (G), polar uncharged(S),non-polar aliphatic(A,L,V) and hydrophobic(G,L,V), CBetaBranched(V)	E, G, A, L, T, V, S	Helix, Coil and Sheet	Acidic Negative charged protein stable (E), Polar aliphatic (G),polar uncharged(S,T),non-polar aliphatic(A,L,V) and hydrophobic(G,L,V), CBetaBranched(T,V)
Louping ill	G, A, L, T, V	Helix, Coil	Polar aliphatic (G), non-polar aliphatic(A,L,V), and polar uncharged(T) and hydrophobic(G,L,V), CBetaBranched(T,V)	G, A, L, T, V	Helix And sheet	Polar aliphatic hydrophobic (G), non-polar aliphatic(A,L,V), polar uncharged(T) and hydrophobic(G,L,V), CBetaBranched(T,V)
Powassan	E, G, A, L, T, V, S, R	Helix & Sheet	Acidic Negative charged protein stable (E), Polar aliphatic hydrophobic (G), non-polar aliphatic(A,L,V), polar uncharged(T,S),basic charged(R) and hydrophobic(G,L,V), CBetaBranched(T,V)	E, G, A, L, T, V, S, R	Helix & Sheet	Acidic Negative charged protein stable (E), Polar aliphatic hydrophobic (G), non-polar aliphatic(A,L,V), polar uncharged(T,S),basic charged(R) and hydrophobic(G,L,V), CBetaBranched(T,V)
Russian Spring - rodent	E, G, A, L, I, K, T, V, P, S]	Helix, Coil and Sheet	Acidic Negative charged protein stable (E), Polar aliphatic hydrophobic (G), non-polar aliphatic(A,L,V), polar ,Aliphatic Neutral Non-polar(P),uncharged(T,S),basic charged(R) and hydrophobic(G,L,V), CBetaBranched(T,V)	E, G, A, L, I, K, T, V, P, S]	Helix, Coil and Sheet	Acidic Negative charged protein stable (E), Polar aliphatic hydrophobic (G), non-polar aliphatic(A,L,V), polar ,Aliphatic Neutral Non-polar(P),uncharged(T,S),basic charged(K) and hydrophobic(G,L,V), CBetaBranched(T,V)
Summer encephalitis	G, A, L, T, V	Helix and Sheet	Polar aliphatic (G), non-polar aliphatic(A,L,V), and polar uncharged(T), hydrophobic(G,L,V), CBetaBranched(T,V),	G, A, L, T, V	Helix and sheet	Polar aliphatic (G), non-polar aliphatic(A,L,V), and polar uncharged(T), hydrophobic(G,L,V), CBetaBranched(T,V),
Known Vector Subfamilis						
Rocio	G, A, L,T, V,P, S	Coil and Helix	Polar aliphatic (G), polar uncharged(T,S),non-polar aliphatic(A,L,V) and hydrophobic(G,L,V), Non-polar(P), CBetaBranched(T,V)	E, G, A, L, T, V, P,S	Helix and Coil	Acidic Negative charged protein stable (E), Polar aliphatic (G),polar uncharged(S,T),non-polar aliphatic(A,L,V) and Non-polar(P),hydrophobic(G,L,V), CBetaBranched(T,V)
Rio bravo	G, A, L, V,S	Helix, Coil	Polar aliphatic (G), non-polar aliphatic(A,L,V), and hydrophobic(G,L,V), CBetaBranched(T,V)	G, L, T, V	Coil	Polar aliphatic hydrophobic (G), non-polar aliphatic(L,V), polar uncharged(T) and hydrophobic(G,L,V), CBetaBranched(T,V)

It has been found in Table 4 that maximal frequent amino acid patterns of flavi virus subfamilies are different. Most of the amino acids are common and belong to hydrophobicity, CBetaBranched, polar aliphatic and uncharged, non-polar aliphatic group. The variations among the subfamilies are in Acidic Negative Charged, Protein stability and basic positive charged group.

Talbe-5 Probable Helix Structure of Protein of flavi virus based on amino acid association

Subfamily	**Helix Formation**							
	(A,R,E,Q,L,K,M,H)							
	Redundant				**Non-redundant**			
	1-Frequent Patterns	**2-Frequent Patterns**	**3-Frequent Patterns**	**4-Frequent Patterns**	**1-Frequent Patterns**	**2-Frequent Patterns**	**3-Frequent Patterns**	**4-Frequent Patterns**
Japanese encephalitis	**A,L**	**AL**	**None**	**None**	**A,L,E,R**	**AL,AE,AR, LE,LR**	**EAL,ALR**	**None**
Marry vally encephalitis	**A,L,E,K, R**	**AL,AE,AK, AR,LE,LK,L R,EK,ER**	**EAL, EAK, EAR,ELK, ELR,ALK, ALR**	**EALK**	**A,L,E,K, R**	**AL,AE,AK, AR,LE,LK, LR,EK,ER**	**EAL,EAK, EAR,ELK, ELR,ALK, ALR**	**EALK**
St. Louis encephalitis	**A,L**	**AL**	**None**	**None**	**A,L,E,K**	**AL**	**None**	**None**
Ilheus	**A,L**	**AL**	**None**	**None**	**A,L**	**AL**	**None**	**None**
West Nile	**A,L,E,K, R**	**AL,AE,AR, LE,LR,ER**	**ELR,ALR**	**None**	**A,L,E,K,R**	**AL,AE,AR, LE,LK,LR, ER**	**EAL,ELR, ALR,EAR**	**EALR**
Central European	**A**,**L**	**AL**	**None**	**None**	**A,L,E,**	**AL,AE,LE**	**None**	**None**
Louping ill	**A,L**	**AL**	**None**	**None**	**A,L**	**AL**	**None**	**None**
Powassan	**A,L**,E,R	**AL**,AE,AR,LE ,LR,ER	**EAL,EAR ,ELR,ALR**	**EALR**	**A,L,E,R**	**AL,AE,AR ,LE,ER**	**ALR,ELR**, EAL	**EALR**
Russian Spring – rodent	**A,L,E,K**,	**AL**,LE,EK,LK	**EAL**,EAK ,ELK,ALK,	EALK	**A,L,E,K**	**AL,AE,AK**, **EK,LK**	**EAL,EAK ,ALK**,	**EALK**
Summer encephalitis	**A,L,**	**AL**	**None**	**None**	**A,L**	**AL**	**None**	**None**
Rocio	**A**,**L**	**AL**	**None**	**None**	**A,E**,**L**	**AL**,**EL**	**None**	**None**
Rio bravo	**A**,**L**	**AL**	**None**	**None**	**L**	**None**	**None**	**None**

Talbe-6 Probable Sheet Structure of Protein of flavivirus based on amino acid association

Subfamily	Sheet Formation							
	(V,I,T,C,W,F,Y)							
	Redundant				Non-redundant			
	1-Frequent Patterns	**2-Frequent Patterns**	**3-Frequent Patterns**	**4-Frequent Patterns**	**1-Frequent Patterns**	**2-Frequent Patterns**	**3-Frequent Patterns**	**4-Frequent Patterns**
Japanese encephalitis Threshold 12112.5	**V,T**	**VT**	**None**	**None**	**V,T**	**VT**	**None**	**None**
Marry vally encephalitis	**V,T**	**VT**	**None**	**None**	**V,T**	**VT**	**None**	**None**
St. Louis encephalitis	**V,T**	**VT**	**None**	**None**	**V,T**	**VT**	**None**	**None**
Ilheus	**V,I,T**	**None**	**None**	**None**	**V,I,T**	**None**	**None**	**None**
West Nile	**V,T**	**VT**	**None**	**None**	**V,T**	**VT**	**None**	**None**
Central European	**V**,**T**	**None**	**None**	**None**	**V,T**	**VT**	**None**	**None**
Louping ill	**V,T**	**VT**	**None**	**None**	**V,T**	**VT**	**None**	**None**
Powassan	**V**,**T**	**VT**	**None**	**None**	**V,T**	**VT**	**None**	**None**

Russian Spring – rodent	**V,I,T**	**VL**,VI	**LIV**	**None**	**V,I,T**	**VI,VL,LI**	**ITV**	**None**
Summer encephalitis	**V,T**	**VT**	**None**	**None**	**V,T**	**VT**	**None**	**None**
Rocio	**V**,**T**	**None**	**None**	**None**	**V,T**	**None**	**None**	**None**
Rio bravo	**V**	None	**None**	**None**	**V,T**	None	**None**	**None**

Talbe-7 Probable Coil Structure of Protein of flavivirus based on amino acid association

Subfamily	Coil							
	(N,D,P,S,G)							
	Redundant				Non-redundant			
	1-Frequent Patterns	2-Frequent Patterns	3-Frequent Patterns	4-Frequent Patterns	1-Frequent Patterns	2-Frequent Patterns	3-Frequent Patterns	4-Frequent Patterns
Japanese encephalitis Threshold 12112.5	**G,S**	**GS**	**none**	**None**	**G,S**	**GS**	**None**	**None**
Marry vally encephalitis	**G,S**	**GS**	**None**	**None**	**G,S**	**GS**	**None**	**None**
St. Louis encephalitis	**G**	**None**	**None**	**None**	**G**	**None**	**None**	**None**
Ilheus	**G,S**	**GS**	**None**	**None**	**G,S**	**GS**	**None**	**None**
West Nile	**G,S**	**GS**	**None**	**None**	**G,S**	**SG**	**None**	**None**
Central European	**G,S**	**GS**	**none**	**None**	**G**	**None**	**None**	**None**
Louping ill	**G**	**None**	**None**	**None**	**G**	**None**	**None**	**None**
Powassan	**G, S**	**GS**	**None**	**None**	**G**,**S**	**GS**	**None**	**None**
Russian Spring – rodent	**G,S,P**	**GS,GP,PS**	**GPS**	**None**	**G,S,P**	**GS,GP,PS**	**GPS**	**None**
Summer encephalitis	**G**	**None**	**None**	**None**	**G**	**None**	**None**	**None**
Rocio	**G,P,S**	**GP,GS**	**none**	**None**	**G**,P,S	**GP,GS**	**None**	**None**
Rio bravo	**G**,**S**	**GS**	**None**	**none**	**G**, **S**	**GS**	**None**	**None**

The Table 5 presents the amino acid association patterns which are responsible for formation of secondary structure Helix, in various Flavi virus subfamilies. It has been observed that redundant dataset of Japanese encephalitis (Mosqiti borne subfamily),Central European (Tick borne subfamily) and Rocio (Known Vector) have two frequent patterns AL (amino acid Alanine and Lucien) which are responsible for formation of Helix and non-redundant dataset of Japanese encephalitis (Mosqiti borne subfamily),Central European (Tick borne subfamily) and Rocio(Known Vector) have three frequent patterns A, L, E (amino acid Alanine, Lucien and Glutamic acid) which are responsible for formation of secondary structure Helix. Along with Japanese encephalitis has amino acid R(Arginine) which are responsible for formation of Helix. Table 6 and 7 presents the amino acid association patterns which are responsible for formation of secondary structure Sheet and Coil respectively in various flavi virus sub families .

In Table 6 it has been observed that redundant and non- redundant dataset of Japanese encephalitis (Mosqiti borne subfamily),Central European (Tick borne subfamily) and Rocio (Known Vector) have two frequent patterns VT (amino acid Valine and Threonine) which are responsible for formation of Sheet.

In Table 7 it's observed that most of subfamilies have two frequent patterns GS which are responsible for formation of secondary structure Coil.and amino acid P(proline) is also present in the some subfamilies which is

important in coil formation. Parallel analysis can be ready for rest of the subfamilies in table 5,6 and 7 to produce association rule.
It has been shown in table 5,6 and 7 that in all the 12 subfamilies of Flavi virus family ,association patterns of amino acid exposed high tendency to form Helix rather than Sheet and Coil.

The association rules generated on the basis of aboveFlavi Virus Subfamilies are given below:-

Mosquito Borne Subfamilies:-

For 2 frequent Patterns:-

I. {A(Frequent)∩L(Frequent)=>Tendency for Helix Formation}
II. {A(Frequent)∩E(Frequent)=>Tendency for Helix Formation}
III. {A(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
IV. {A(Frequent)∩R(Frequent)=>Tendency for Helix Formation}
V. {L(Frequent)∩E(Frequent)=>Tendency for Helix Formation}
VI. {L(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
VII. {L(Frequent)∩R(Frequent)=>maintain charge of protein and help in protein stability }
VIII. {E(Frequent)∩K(Frequent)=>Tendency for Helix Formation and Protein Solubility}
IX. {E(Frequent)∩R(Frequent)=>Tendency for Helix Formation and Protein Solubility }
X. {V(Frequent)∩T(Frequent)=>Tendency for Sheet Formation}
XI. {G(Frequent)∩S(Frequent)=>Tendency for Coil Formation}

For 3 frequent Patterns:-

I. { E(Frequent)∩A(Frequent)∩L(Frequent)=>Tendency for Helix Formation}
II. {A(Frequent)∩L(Frequent)∩R(Frequent)=>Tendency for Helix Formation}
III. { E(Frequent)∩A(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
IV. { E(Frequent)∩A(Frequent)∩R(Frequent)=>Tendency for Helix Formation}
V. { E(Frequent)∩L(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
VI. { A(Frequent)∩L(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
VII. { E(Frequent)∩L(Frequent)∩R(Frequent)=>Tendency for Helix Formation}

For 4 frequent Patterns:-

I.{E(Frequent)∩A(Frequent)∩L(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
II. {E(Frequent)∩A(Frequent)∩L(Frequent)∩R(Frequent)=>Tendency for Helix Formation}

For Tick Borne Subfamilies:-

For 2 Frequent Patterns

{A(Frequent)∩L(Frequent)=>Tendency for Helix Formation}
I. {A(Frequent)∩E(Frequent)=>Tendency for Helix Formation}
II. {A(Frequent)∩R(Frequent)=>Tendency for Helix Formation}
III. {L(Frequent)∩E(Frequent)=>Tendency for Helix Formation}
IV. {L(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
V. {L(Frequent)∩R(Frequent)=>maintain charge of protein and help in protein stability }
VI. {E(Frequent)∩K(Frequent)=>Tendency for Helix Formation and Protein Solubility}
VII. {E(Frequent)∩R(Frequent)=>Tendency for Helix Formation and Protein Solubility }
VIII. {V(Frequent)∩T(Frequent)=>Tendency for Sheet Formation}
IX. {V(Frequent)∩I(Frequent)=>Tendency for Sheet Formation}
X. {L(Frequent)∩I(Frequent)=>Tendency for Sheet Formation}
XI. {G(Frequent)∩S(Frequent)=>Tendency for Coil Formation}
XII. {G(Frequent)∩P(Frequent)=>Tendency for Coil Formation}
XIII. {P(Frequent)∩S(Frequent)=>Tendency for Coil Formation}

For 3 frequent Patterns:-

VIII. { E(Frequent)∩A(Frequent)∩L(Frequent)=>Tendency for Helix Formation}
IX. {A(Frequent)∩L(Frequent)∩R(Frequent)=>Tendency for Helix Formation}
X. { E(Frequent)∩A(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
XI. { E(Frequent)∩A(Frequent)∩R(Frequent)=>Tendency for Helix Formation}
XII. { E(Frequent)∩L(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
XIII. { A(Frequent)∩L(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
XIV. { E(Frequent)∩L(Frequent)∩R(Frequent)=>Tendency for Helix Formation}
XV. {I(Frequent)∩T(Frequent)∩V(Frequent)=>Tendency for Sheet Formation}
XVI. {G(Frequent)∩P(Frequent)∩S(Frequent)=>Tendency for Coil Formation}

For 4 frequent Patterns:-

I. I.{E(Frequent)∩A(Frequent)∩L(Frequent)∩K(Frequent)=>Tendency for Helix Formation}
II. {E(Frequent)∩A(Frequent)∩L(Frequent)∩R(Frequent)=>Tendency for Helix Formation}

For Known Vector:-

For 2 Frequent Patterns:-

I. {A(Frequent)∩L(Frequent)=>Tendency for Helix Formation}
II. {G(Frequent)∩S(Frequent)=>Tendency for Coil Formation}
III. {G(Frequent)∩P(Frequent)=>Tendency for Coil Formation}

IV. Conclusion

The approach for mining quantitative associations is appropriate in view of analysis and prediction of secondary structure of sequences for each subfamily of flavivirus. Further quantitative associations have been obtained successfully to generate association rules for predicting the physiochemical properties and secondary structure. This approach will lead to develop fuzzy set and soft set based approaches for mining association in flavivirus sequences and their structures. These studies can be performed to generate association rules, signatures and their relationships along with physiochemical properties.

V. Conflict Of Interest

The authors are no conflict of interest.

ACKNOWLEDGMENT

The authors are grateful to Department of Biotechnology, New Delhi for Providing Bioinformatics Infrastructure facility at MANIT Bhopal to carry out this.

References

[1]. M. W. Gaunt, A. A. Sall, X. de Lamballerie, A. K. I. Falconar, T. I. Dzhivanian, and E. A. Goul(2001)d, "Phylogenetic relationships of flaviviruses correlate with their epidemiology, disease association and biogeography," *Journal of General Virology*, vol. 82, part 8, pp. 1867–1876.
[2]. Shi, P-Y (editor) (2012). *Molecular Virology and Control of Flaviviruses*. Caister Academic Press. ISBN 978-1-904455-92-9.
[3]. A. Oya and I. Kurane(2007), "Japanese encephalitis for a reference to international travelers," *Journal of Travel Medicine*, vol. 14, no. 4, pp. 259–268.
[4]. PanisadeeAvirutnan,Anja Fuchs, Richard E. Hauhart(March 2010),"Antagonism of the complement component C4 by flavivirusnonstructural protein NS1"J Exp Med 2010 207:793-806
[5]. B. J. Blitvich(2008), "Transmission dynamics and changing epidemiology of West Nile virus," *Animal Health Research Reviews*, vol. 9, no. 1, pp. 71–86.
[6]. S. B. Halstead(2007), "Dengue," *The Lancet*, vol. 370, no. 9599, pp. 1644–1652.
[7]. E. Gould and T. Solomon(2008), "Pathogenic flaviviruses," *The Lancet*, vol. 371, no. 9611, pp. 500–509.
[8]. Jaiwei Han and Yongjian (1995)"Discovery of multiple level association rule from large database",VLDB Conference, pp. 420-431.
[9]. PanisadeeAvirutnan,Anja Fuchs, Richard E. Hauhart(March 2010),"Antagonism of the complement component C4 by flavivirusnonstructural protein NS1"J Exp Med 2010 207:793-806
[10]. Gowtham Atluri at al "Association Analysis Techniques for Bioinformatics Problems"{gowtham,rohit,gangfang,gaurav,steinbac,kumar}@cs.umn.edu,http://www.cs.umn.edu/~kumar/dmbio
[11]. Pandey, Anjana and KamalRaj Pardasani."Rough set Model for Discovering Multidimensional Association Rules."International Journal of Computer Science and Network Security9.6(2009):159-164.
[12]. Khare,Neelu,Neeru Adlakha, and K.R. Pardasani."An Algorithm for Mining Multidimensional Association Rules using Boolean Matix."Recent Trends in Information,Telecommunication and Computing (ITC),2010 International Conference on.IEEE,2010.
[13]. Khare,N.;Adlakha,N.;Pardasani,K.R."An algorithm for mining conditional hybrid dimensional association rule using Boolean Matix,"Computer and automation engineering (ICCAE),2010 The 2nd International Conference on,vol.2,no.,pp.644,648,26-28 Feb.2010
[14]. Khare ,Neelu Adlakha, and K.R. Pardasani."Karnaugh MAP Model for Mining Association Rules in Large Databases."(IJCNS) International Journal of Computer and Network Security 1.1(2009).
[15]. Pandey,Anjana,Niket Bhargava,and K.R. Pardasani."Counting inference appoarch to discover calendar based temporal association rules." SPIT-IEEE Colloquium. Vol.5.2007.
[16]. Panday,Anjana, and K.R. Pardasani."PPCI algorithm for mining temporal association rules in large database."Journal of Information & Knowledge Management 8.04(2009):345-352.
[17]. Khare,Neelu,Neelu Adlakha, and K.R. Pardasani."An Algorithm for Mining Multidimensional Fuzzy Association Rules."arXiv preprint arXiv:0909.5166(2009).
[18]. Agrawal,R. and Srikant,R.(1994).Fast algorithms for mining association rules .In proc of the 20th Int'l Conference on Very Large Databases,407-419
[19]. Srikant R and Agrawal(1996) "Mining quantitative association rules in large relational tables" SIGMOD '96 Proceedings of the 1996 ACM SIGMOD international conference on Management of data, Volume 25 Issue 2, June 1996 Pages 1-12
[20]. Nitin Gupta, Nitin Mangal, Kamal Tiwari, Pabitra Mitra(2006) "Mining Quantitative Association Rules in Protein Sequences" Lecture Notes in Computer Science Volume 3755, pp 273-281
[21]. Anuja Shankar & KamalRAj Pardasani"Mining Fuzzy amino acid association patterns in various orders of class Apphaproteobacteria"Journal of Medical Imaging and Health Informatics.Accepted for Publication 2013
[22]. Kumari T,Pardasani KR.,"Mining Fuzzy amino acid Association Patterns in Class C GPCRs"Computational Life sciences,Springer.Accepted of Publication 2013
[23]. Kumari T, Pardasani KR.,"Mining Fuzzy associations among amino acids of class A GPCRs."OnlineJ Bioinform.,13(2):202-213,2012
[24]. Anuja Shankar & Kamal Raj Pardasani,"Amino acid composition based model for prediction and identification of Alpa and Epsilon-proteobacteria"Online Journal of Bioinformatics,14(1)(2013)

Publiched date 14 mar 2014