International Journal of Life science & Pharma Research

WEIGHTED SOFT SET APPROACH FOR MINING FREQUENT AMINO ACID ASSOCIATIONS IN PEPTIDE SEQUENCES OF SWINE INFLUENZA VIRUS

ALEKH GOUR*, DR. K.R. PARDASANI

Department of Mathematics, Bioinformatics & Computer Applications, Maulana Azad National Institute of Technology, Bhopal - 462003,India

ABSTRACT

The amino acid associations present in molecular sequences of viruses have correlation with the molecular mechanisms of the disease and other molecular process of an organism. The knowledge of these amino acid associations is crucial for understanding these molecular mechanisms and process. The major challenge in exploring amino acid association patterns in molecular sequences is the presence of uncertainty. In this paper weighted soft set approach is proposed to mine amino acid associations in molecular sequences of swine influenza virus. Soft set has been employed to incorporate the relationship of amino acid associations with the parameters. The parameter like length has been incorporated by assigning weights to the different length ranges. The dataset of 82611 sequences of swine influenza virus is taken from NCBI and filtered to obtain 36434 non redundant sequences. The proposed approach is employed to explore amino acid associations and results have been compared with ordinary soft set approach. It is observed that the weighted soft set approach is able to prune the over estimation of support by ordinary soft set approach. Further the weighted soft set approach is able to reduce the deviation in association patterns thereby improving the consistency of results. It is also observed that weighted soft approach is able to address the 50% to 85% of uncertainity left out by soft set approach which is leading to improvement in the results. Thus weighted soft set approach is superior to ordinary soft set approach for mining amino acid associations in swine influenza virus.

KEYWORDS: *Soft Set, Weighted Soft Set, Amino acid, Association rule mining, Swine Influenza Virus, Uncertainty*

INTRODUCTION

The molecular databases have been growing at an exponential rate leading to the problem of big data and evolution of biological data science. One of the major challenges for data analytics in biological data is the presence of uncertainty which leads to underestimation or overestimation of the information and knowledge present in the datasets. For the past few decades investigators have been developing algorithms for data mining with vide variety of applications in business, science and technology. Some data mining techniques have been widely used in biological datasets like classification and secondary structure prediction in proteins, clustering in biological data, feature selection in bioinformatics and pattern mining[1-4] techniques for exploring biological dataset. Association rule mining is one of the important areas of data mining which has gained interest among various investigators. Apriori algorithm[5] was defined by Agarwal et. al in 1994 for association rule mining. Other algorithms like F-P Growth[6], Eclat Algorithm[7] are also reported for association rule mining which speeds up the process of pattern generation but apriori is most suitable algorithm for transactional dataset. Previously some efforts have been made to mine frequent amino acid association in peptide sequences[8-10] based on ordinary approaches. Although the results obtained by ordinary approaches are very essential they contain lot of uncertainty due to ignorance of degree of membership among amino acid. The attempts are also reported which use the fuzzy sets defined by Zadeh[14] to mine fuzzy association rules in various diseases[11-13] which involves degree of membership among amino acids with the length of sequences. Few attempts are reported which use soft fuzzy set[15] to mine amino acid association in peptide sequences of dengue virus[16] and mycobacterium tuberculosis[17]. The soft set is not

completely capable to incorporate the relationships of parameters with association patterns. Therefore in the present study an attempt has been made to develop a weighted soft set approach for mining amino acid association pattern in swine influenza virus. Swine influenza virus is declared as pandemic disease by WHO in 2010[18] .It involves influenza C and the subtypes of influenza A known as H1N1, H1N2, H2N1, H3N1, H3N2, and H2N3.The swine influenza virus affects around 100 million people worldwide every year[19].The emergence of new strains of swine influenza virus will continue to pose new challenges to scientific communities to study the molecular mechanism of virus and its different strains. In the present study the relationship of parameter with association pattern is incorporated using soft set and further soft set is modified by assigning weight to different length ranges to propose weighted soft set approach. The proposed algorithm is given in the next section.

MATERIAL AND METHOD

Dataset

A universal dataset $\bar{D} = \{\bar{S}_1, \bar{S}_2, ..., \bar{S}_n\}$ is a collection of non empty transactional database where $\bar{D} \triangleright 0$. Where $|\bar{D}|$ represents cardinal number of sets, which means number of elements in the set, and $\bar{S}_i$ represents a peptide sequence[16].

Item Set

Let $\bar{A} = \{a_1, a_2, ..., a_n\}$ be set of amino acids where, Amino acids = {Alanine(A), Asparagine(N), Arginine(R), Aspartic Acid(D), Cysteine(C), Glutamic Acid(E), Glutamine(Q), Glycine(G), Histidine(H), Isoleucine(I), Leucine(L), Lysine(K), Methionine(M), Phenylalanine(F), Proline(P), Serine(S), Threonine(T), Typtophan(W), Tyrosine(Y), Valine(V)}. Here $\bar{A}$ is considered as itemset where each amino acid is an item.

Transaction

Each individual peptide sequence is considered as a single transaction. Here a peptide sequence is a combination of twenty amino acids. Let $\bar{A}$ is a set of amino acids, Thus a transaction can be stated as $\bar{S}' = \{x \mid \forall x \in \bar{A}\}$, i= 1(1) n [16].

Association Rule

An association rule[20] is of the form $[a_j] \rightarrow [a_k]$, where $[a_j]$ is defined as rule body and $[a_k]$ is defined as rule head, specific condition for the rule is both a_j and a_k should not be identical. Two crucial properties used in generating an association rule are support and confidence. Support is defined as the number of transactions supporting the association rule in the database of transactions expressed as[20]

$$\text{Support}[a_j \rightarrow a_k] = \sum_{i=1}^{n} f_i(a_j \cup a_k) / n \qquad \text{Eq: (1)}$$

where $j \neq k$ and j, k = 1(1) 20, f_i represents the frequency of amino acids in i^{th} sequence. Symbol $\cup$ represent association of amino acids a_j and a_k in i^{th} sequence, $j \neq k$. Whereas confidence is the percentage of transactions supporting the rule out of all transactions supporting the rule body and is expressed as[18]:

$$\text{Confidence}[a_j \rightarrow a_k] = \text{Support}[a_j \rightarrow a_k] / \sum_{i=1}^{n} \text{Support}(a_j) \qquad \text{Eq: (2)}$$

where j,k = 1(1) 20 and $j \neq k$. The item $[a_j]$ is called as frequent if, support $[a_j]$ > threshold.

Threshold

The following approach has been implemented to
find threshold described as contribution of one amino acid in all twenty amino acids with an individual length of the sequence, expressed as:

$$T = \sum_{i=1}^{n} (1/20) * l_i \qquad \text{Eq: (3)}$$

Where T is total threshold l_i is length of sequence and i = 1 to n and n is the total number of sequence.

Association Pattern

Association pattern is denoted by $\bar{P}$ and expressed as $\bar{P} = (a_1 \cup a_2 \cup a_3 \cup \ldots \cup a_n)$ where $a_j \neq a_k$ and $(j,k) = 1 \text{ to } 20$. Here $\cup$ represent association of amino acids a_1, a_2,…, a_n in the sequence P_i represents i^{th} pattern.

Soft Set

Soft set theory is basically introduced by D. Molodtsov[21] to deal with uncertainty in datasets due to parameters. Definition: Let there be a universal set $\bar{D}$, $P(\bar{D})$ is the power set of $\bar{D}$, set of parameters E and $A \subset E$ then the soft set pair $(\bar{S}, A)$ over $\bar{D}$ is defined as $\bar{S} : A \rightarrow P(\bar{D})$

Soft Threshold

Soft threshold is the association of threshold with different parameters denoted as $(\bar{T}, e)$ and expressed as

$$(\bar{T}, e) = ((\sum (1/20) * L_i), e) \qquad \text{Eq: (4)}$$

where T_i is threshold and e is the set of parameters.

Soft Transaction

A soft transaction is denoted as $\bar{S}'$ which is explained as the mapping of an item set a with a different set of parameters E expressed as $\bar{S}' = \{(x, e) \mid \forall x \in a, \forall e \in E\}$

Soft Association Pattern

It simply represents an association of patterns with different parameter denoted by $(\bar{P}', e)$ and expressed as $(\bar{P}', e) = ((a_1 \cup a_2 \cup a_3 \cup \ldots \cup a_n), e)$ where e is the set of parameters.

Weighted Soft Set

A soft set is termed as weighted soft set or W-soft set[22] or Quantitative soft set if each element in a parameter of soft set has a weight assigned to it rather than only grade of membership. The weighted soft set $\bar{\bar{S}}$ can be calculated as

$$\bar{\bar{S}} = w_i * a_{ij} \qquad \text{Eq: (5)}$$

where 'w_i' is the weight of elements of parameter and 'a_{ij}' is the ij-th entry in the table of weighted soft set.

Weighted Soft Threshold

Weighted soft threshold is the calculated by combining weight of each element of parameter with the soft threshold. It is defined as

$$(\bar{\bar{T}}, e) = ((\sum (1/20) * L_i), e) * W_i \qquad \text{Eq: (6)}$$

where 'w_i' is the weight of elements of parameter.

Weighted Soft Transaction

Weighted Soft transaction is denoted as $\bar{\bar{S}}'$ which is explained as the mapping of weighted item set a with a different set of parameters E expressed as $\bar{\bar{S}}' = \{(w_i x, e) \mid \forall x \in a, \forall e \in E\}$

Weighted Soft Association Pattern

It simply represents an association of patterns with different weighted parameter denoted by $(\bar{P}'', e)$ and expressed as $(\bar{P}'', e) = ((w_i a_1 \cup w_i a_2 \cup w_i a_3 \cup \ldots \cup w_i a_n), e)$ where e is the set of parameters.

Algorithm

The weighted soft set approach has been proposed and employed for exploring association patterns in peptide sequences of swine influenza virus.

S1: Compute the frequency of each Amino Acid in each peptide sequence and apply threshold as defined in

Eq: (3) to generate frequent amino acid and frequent patterns.

S2: Take sequence length as parameter and calculate L_{max} & L_{min} where L_{max} is maximum length of sequence and L_{min} is minimum length of sequences. Store the sequences in three different arrays (called ranges) according to their category (Range1, Range2 and Range3) using the given formulae

$$\text{Range1} = L_{min} \text{ to } (L_{max} - L_{min})/3 + L_{min}$$
$$\text{Range2} = ((L_{max} - L_{min})/3 + L_{min}) + 1 \text{ to } 2((L_{max} - L_{min})/3 + L_{min})$$
$$\text{Range3} = 2((L_{max} - L_{min})/3 + L_{min}) + 1 \text{ to } L_{max}$$

S3: Within each range calculate frequency of Amino Acid & store it in an array. Apply threshold as defined in Eq: (4) to generate frequent amino acid and soft frequent association pattern on basis of soft sets.

S4: Apply weighted soft set and generate frequent amino acid and weighted soft association pattern by comparing the support with the threshold defined in Eq: (6)

S5: Generate association rules and calculate the support as defined in Eq: (1) and confidence as defined in Eq: (2)

RESULTS AND DISCUSSION

A total of 82611 peptide sequences of swine influenza virus are downloaded from NCBI. The first step is the data preparation. Various filters have been applied to remove redundancy, partial sequences and useless sequences. The final dataset of 36434 is taken for further study. The soft set is employed to divide dataset in three ranges Range 1, Range 2 and Range 3 on the basis of length of sequences. 12882 sequences lie in Range 1, 18189 sequences lie in Range 2 and 5363 sequences lie in Range 3. The frequency of amino acid based on soft set approach is calculated and results obtained are shown in column 2, 3 and 4 of Table 1. The weights are assigned to each range of parameter length. Weight of 0.35 is assigned to range 1, 0.5 weight is assigned to range 2 and 0.15 weight is assigned to range 3. These weights are multiplied to the frequency of amino acids and association patterns. The results obtained by the application of weighted soft set are shown in column 4, 5 and 6 of Table 1. The threshold is fixed in case of soft set i.e. 0.05 in all ranges of length parameter whereas it varies in case of weighted soft set. The threshold for range 1 sequences is 0.0175, for range 2 sequences the threshold is 0.025 and for range 3 the threshold is 0.0075. It is observed that the amino acids whose support value crosses the defined threshold are same in both soft set and weighted soft set. Although the threshold is different in all ranges of weighted soft set but the frequent amino acids reported in all ranges are same. The frequent amino acids reported by both soft set and weighted soft set approaches are Glycine, Alanine,Valine, Leucine, Isoleucine, Asparagine, Serine, Threonine, Glutamic Acid, Lysine and Arginine. The standard deviation in the support value of amino acid using soft set is shown in column 8, 9 and 10 and the standard deviations in the support value of amino acid using weighted soft approach is shown in column 11, 12 and 13 of Table 1.

Table 1
Support of amino acid and deviation in their values by using soft and weighted soft approaches

	Soft			Weighted Soft			Standard Deviation Soft			Standard Deviation Weighted Soft		
	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
G	0.062	0.074	0.062	0.0217	0.0370	0.0093	14.00	40.51	47.16	4.90	20.26	7.07
A	0.058	0.052	0.057	0.0202	0.0261	0.0085	13.91	30.00	42.92	4.87	15.00	6.44
V	0.056	0.057	0.064	0.0196	0.0284	0.0096	12.67	31.40	49.96	4.44	15.70	7.49
L	0.092	0.072	0.079	0.0321	0.0358	0.0119	19.64	41.54	59.95	6.88	20.77	8.99
I	0.061	0.072	0.065	0.0214	0.0360	0.0097	13.31	39.95	48.92	4.66	19.98	7.34
M	0.030	0.024	0.046	0.0104	0.0118	0.0069	6.94	14.99	35.12	2.43	7.50	5.27
P	0.041	0.038	0.040	0.0144	0.0190	0.0060	9.56	21.33	30.35	3.34	10.67	4.55
F	0.039	0.038	0.038	0.0136	0.0190	0.0056	8.60	21.95	28.82	3.01	10.97	4.32
W	0.019	0.021	0.013	0.0066	0.0103	0.0019	4.29	11.58	9.53	1.50	5.79	1.43
N	0.055	0.066	0.052	0.0192	0.0332	0.0078	13.25	36.91	40.27	4.64	18.46	6.04
Q	0.038	0.031	0.044	0.0132	0.0156	0.0065	8.24	17.60	33.12	2.88	8.80	4.97

S	0.077	0.085	0.069	0.0271	0.0424	0.0104	17.26	46.59	52.44	6.04	23.29	7.87
T	0.067	0.061	0.074	0.0236	0.0304	0.0111	15.31	33.58	56.16	5.36	16.79	8.42
Y	0.028	0.035	0.027	0.0100	0.0175	0.0040	7.52	19.82	20.92	2.63	9.91	3.14
C	0.019	0.029	0.010	0.0066	0.0145	0.0015	4.96	16.56	7.92	1.74	8.28	1.19
D	0.044	0.047	0.045	0.0156	0.0235	0.0068	9.70	26.38	34.13	3.39	13.19	5.12
E	0.072	0.067	0.066	0.0251	0.0337	0.0099	15.45	41.24	49.78	5.41	20.62	7.47
K	0.058	0.059	0.064	0.0204	0.0296	0.0096	13.12	34.37	48.41	4.59	17.19	7.26
R	0.060	0.051	0.074	0.0211	0.0257	0.0111	12.96	30.41	56.09	4.53	15.21	8.41
H	0.023	0.021	0.013	0.0082	0.0103	0.0020	5.60	11.75	10.15	1.96	5.87	1.52

It is observed that there is a significant variation in values of standard deviation in support of amino acid obtained by soft set and weighted soft set. There is large deviation in the support of amino acid when calculated by soft set whereas in weighted soft set this deviation is quite low. This indicate that weighted soft set results more accurate than those obtained by soft set. An estimation of uncertainty[23] is made in Table 2. The uncertainty in individual amino acids is combined and the overall uncertainty in support value of amino acid is shown in Table 2. Column 2 of Table 1 shows the uncertainty calculated by soft set approach. In column 3 of Table 2 the combined uncertainty calculated by weighted soft approach is reported. The weighted soft approach is able to handle 64.99% uncertainty in Range 1 sequences, 49.99% in Range 2 sequences and 85% in Range 3. The weighted soft set is able to handle the veracity present in results obtained by soft set approach.

Table 2
Uncertainty Interpretation

Sequences	Remaining Combined Uncertainty (Soft Set Approach)	Remaining Combined Uncertainty (Weighted Soft Approach)	Measurement of Uncertainty Handled (MU)
SIV Range 1 Sequences	0.474953	0.166245	64.99%
SIV Range 2 Sequences	1.006145	0.50311	49.99%
SIV Range 3 Sequences	2.518399	0.377724	85.00%

The frequent amino acids are used to generate the frequent association patterns. Range 1 sequences reports fifteen maximal frequent pattern of length four, Range 2 sequence reports two maximal frequent pattern of length seven and Range 3 reports one maximal frequent pattern of length ten. Table 3 shows the entire maximal frequent patterns observed in different ranges with their soft support, weighted soft support, standard deviation by using soft and weighted soft approach and estimation of uncertainty.

Table 3
Maximal Frequent Patterns in peptide sequences of swine influenza virus.

Maximal Frequent Pattern SIV Range 1 Weight of Range 1 (W_1) = 0.35 Weighted Soft Threshold = 0.0175	**ALST :** Soft Sup 0.051783577, W_1 Soft Sup 0.007767624, Soft SD 12.345425, W_1 Soft SD 1.8518015, MU 97.75%
	EGLS : Soft Sup 0.05267309, W_1 Soft Sup 0.007901079, Soft SD 11.683252, W_1 Soft SD 1.7524909, MU 97.75%
	EILR : Soft Sup 0.05038277, W_1 Soft Sup 0.007557516, Soft SD 10.961134, W_1 Soft SD 1.6441716, MU 97.75%
	EILS : Soft Sup 0.05320929, W_1 Soft Sup 0.007981321, Soft SD 11.502476, W_1 Soft SD 1.7254169, MU 97.75%
	ELRS: Soft Sup 0.053334583, W_1 Soft Sup 0.008000194, Soft SD 11.487818, W_1 Soft SD 1.7231873, MU 97.75%
	ELST : Soft Sup 0.052646633, W_1 Soft Sup 0.00789706, Soft SD 11.708694, W_1 Soft SD 1.7562816, MU 97.75%

	ELSV : Soft Sup 0.050332963, W_1 Soft Sup 0.0075501003, Soft SD 11.225708, W_1 Soft SD 1.6838884, MU 97.75%
	GILS : Soft Sup 0.050766826, W_1 Soft Sup 0.0076149963, Soft SD 11.278329, W_1 Soft SD 1.691728, MU 97.75%
	GLST : Soft Sup 0.0550245, W_1 Soft Sup 0.008253661, Soft SD 12.688937, W_1 Soft SD 1.9033569, MU 97.75%
	GLSV : Soft Sup 0.051573455, W_1 Soft Sup 0.00773603, Soft SD 11.789086, W_1 Soft SD 1.768365, MU 97.75%
	GSTV : Soft Sup 0.050288215, W_1 Soft Sup 0.007543205, Soft SD 11.617153, W_1 Soft SD 1.7425785, MU 97.75%
	ILST : Soft Sup 0.050478105, W_1 Soft Sup 0.0075717075, Soft SD 11.238596, W_1 Soft SD 1.6857933,MU 97.75%
	KLST : Soft Sup 0.05185245, W_1 Soft Sup 0.0077778893, Soft SD 11.698788, W_1 Soft SD 1.7548159, MU 97.75%
	LNST : Soft Sup 0.05126606, W_1 Soft Sup 0.0076898783, Soft SD 12.416843, W_1 Soft SD 1.8625197, MU 97.75%
	LSTV: Soft Sup 0.052451298, W_1 Soft Sup 0.007867789, Soft SD 12.122312, W_1 Soft SD 1.8183223,MU 97.75%
Maximal Frequent Pattern SIV Range 2 Weight of Range 2 (W_2) = 0.5 Weighted Soft Threshold = 0.025	**GIKLNST:** Soft Sup 0.05035464, W_2 Soft Sup 0.02517732, Soft SD 28.226429, W_2 Soft SD 14.1132145, MU 75%
	GILNSTV: Soft Sup 0.05046905, W_2 Soft Sup 0.025234524, Soft SD 28.04891, W_2 Soft SD 14.024455, MU 75%
Maximal Frequent Pattern SIV Range 3 Weight of Range 3 (W_3) = 0.15 Weighted Soft Threshold = 0.0075	**AEGIKLRSTV:** Soft Sup 0.052835546, W_3 Soft Sup 0.007925285, Soft SD 40.226135, W_3 Soft SD 6.033935, MU 94.5%

The probable secondary structures obtained by frequent patterns are shown in Table 4. The 2-Frequent refers to the frequent pattern of length two, 3-Frequent refers to the frequent pattern of length three and 4-Frequent refers to the frequent pattern of length four. It is observed that the coil structure formed by 2F and 3F patterns is identical in all three ranges. The beta sheet formation by 2F and 3F patterns is similar in range 2 and range 3 but there is a slight variation in beta sheet formation in 2F patterns of range 1. The major differences are recorded in helical structures as we observe that range 1 and range 2 sequences formed the helix structure till 3F patterns only but range 3 reports the helix structure till 4F patterns.

Table 4
Probable Secondary Structures

		2-Frequent (2F)	3-Frequent (3F)	4-Frequent (4F)
Helix	**R1**	AL , LE, LK, LR, EK, ER	AEL, EKL, ELR	-
	R2	AL, AE, LE, LK,	EKL	-
	R3	AL, AE, AK, AR, LE, LK, LR, KR, ER, EK	AEL, AER, AKL, AKR, ALR, EKL, EKR, ELR	AEKL, AEKR, AELR, AKLR, EKLR
Beta Sheet	**R1**	VT, IT	ITV	-
	R2	VI, VT, IT	ITV	-
	R3	VI, VT, IT	ITV	-
Coil	**R1**	GN, GS, NS	GNS	-
	R2	GS, GN NS	GNS	-
	R3	NS, GN, GS	GNS	-

Various physico chemical properties have been taken to analyze the behavior of different range sequence using results obtained by weighted soft set for these properties. The results obtained are shown in Table 5.

An analysis of some of these properties is made in Figure 1 and an analyses of secondary structure is made in Figure 2.

Table 5
Weighted Physico-Chemical Properties

Physico-chemical property	SIV Range 1	SIV Range 2	SIV Range 3
Molecular Weight	7919.13	30750.26	12920.27
Extension Coefficient [Assuming all residues of *tyr,trp,cys*]	10381.05	46150.06	12567.61
Absorbance	0.479118	0.767862	0.145865
Hydrophobicity [GRAVY]	-9.36185	-52.2719	-6.96792
Aliphatic Index	0.154801	0.584192	0.255007
Aromaticity	0.016354	0.07031	0.024526
Protein Stability	0.048147	0.191049	0.082063
C-Beta Branched	0.034992	0.142627	0.064377
Polarity	0.102996	0.415706	0.170824
Salt Bridged	0.046541	0.183723	0.081112
Helix Formation	0.081675	0.283738	0.140588
Beta Sheet	0.05493	0.234781	0.092046
Coil	0.053105	0.233114	0.08536
Positive Charged	0.02251	0.083182	0.043712
Negative Charged	0.022051	0.086022	0.035207

Figure 1 indicates how the positive charges and negative charges are related to protein stability. The positive charge and negative charge increases from range 1 to range 2 and decreases from range 2 to range 3 and thus the protein stability also varies accordingly to the positive and negative charges. In Figure 2 it is observed that by the application of weighted soft set the secondary structures are almost equally distributed in range 2 sequences but vary in range1 and range 3 sequences.

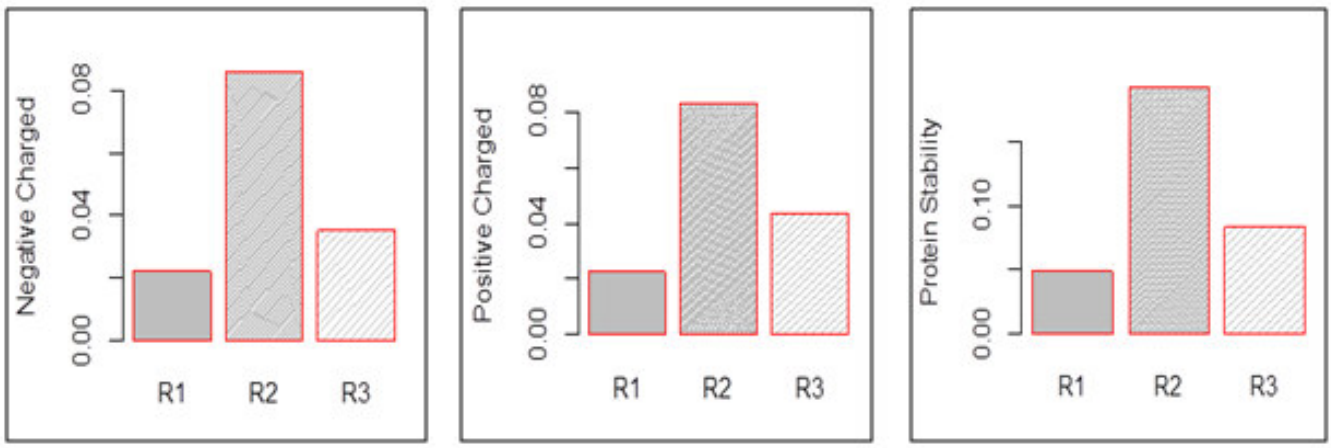

Figure 1
Analyses of Physico chemical properties

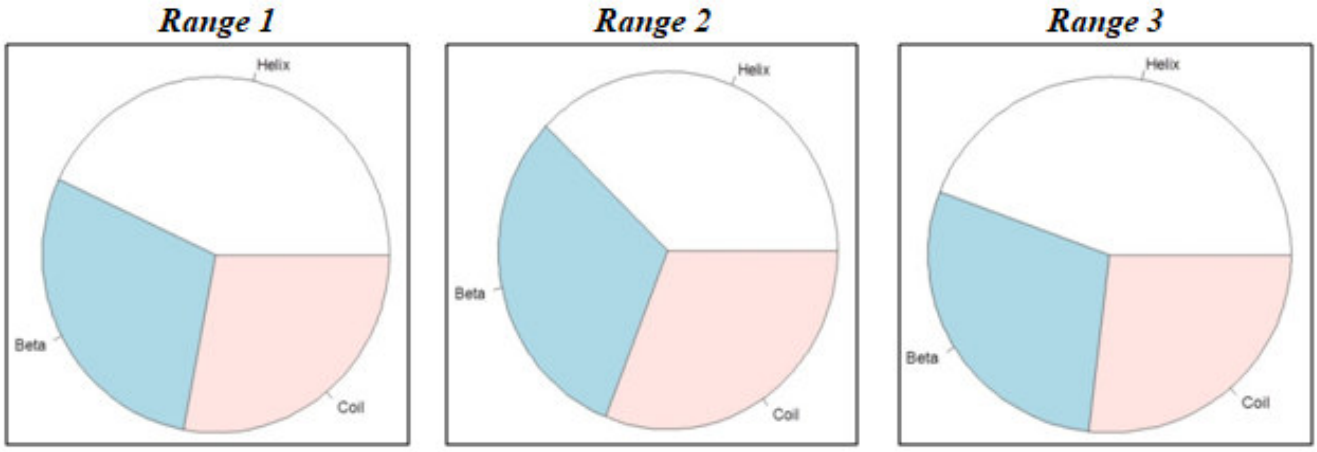

Figure 2
Analyses of secondary structure

CONCLUSION

In this paper a weighted soft set approach is proposed and successfully employed for mining frequent amino acids association present in peptide sequences of swine influenza virus. The simple soft set approach over estimate the support value of amino acids and the association patterns with huge deviations in the actual support and predicted support. The main cause of this over estimation of results is the ignorance of weight of elements of parameters. The weighted soft set is capable to handle this situation of over estimation of results and also minimize the deviation in the actual and predicted values by assigning weights to different elements of the parameter. The association patterns determined by weighted soft set are used to predict secondary structures and physico chemical properties of peptide sequences of swine influenza virus. The weighted soft set is able to handle the uncertainty present due to simple soft set approach. This model can also be implemented in other virus and diseases to generate fruitful results and overcome the bias present in simple soft set approaches. The results generated by this model can further be used for bio medical study and understanding the molecular mechanism of the disease.

ACKNOWLEDGMENT

The authors are highly grateful to the Department of Biotechnology, New Delhi, INDIA and MPCST, Bhopal INDIA for providing Bioinformatics Infrastructure Facility at MANIT Bhopal for carrying out this work.

CONFLICT OF INTEREST

Conflict of Interest declared none.

REFERENCES

1. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics (Oxford, England). 1998 Jan 1;14(4):378-9.
2. Tasoulis DK, Plagianakos VP, Vrahatis MN. Unsupervised clustering of bioinformatics data. InEuropean Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, Eunite 2004 (pp. 47-53).
3. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. Briefings in bioinformatics. 2008 Jun 18; 9(5):392-403.
4. Wang K, Xu Y, Yu JX. Scalable sequential pattern mining for biological sequences. InProceedings of the thirteenth ACM international conference on Information and knowledge management 2004 Nov 13 (pp. 178-187). ACM.

5. Agrawal R, Srikant R. Fast algorithms for mining association rules. InProc. 20th int. conf. very large data bases, VLDB 1994 Sep 12 (Vol. 1215, pp. 487-499).
6. Vijayarani S, Sathya P. Mining frequent item sets over data streams using eclat algorithm. InIJCA Proceedings on International Conference on Research Trends in Computer Technologies. Foundation of Computer Science (FCS) 2013 Feb 27 (Vol. 4, pp. 27-31).
7. Borgelt C. An Implementation of the FP-growth Algorithm. InProceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations 2005 Aug 21 (pp. 1-5). ACM.
8. Gupta N, Mangal N, Tiwari K, Mitra P. Mining quantitative association rules in protein sequences. InData Mining 2006 (pp. 273-281). Springer Berlin/Heidelberg.
9. Kumari T, Pardasani KR. Mining fuzzy associations among amino acids of class A GPCRs. Online J Bioinform. 2012;13(2):202-13.
10. Lavanya Rishishiwar, Bhasker Pant, Kumud Pant, K. R. Pardasani(2011) "Tuber-Gene: Mining Patterns in the Mycobacterium tuberculosis H37Rv Strain". GPB, 9(4-5): 171-178 DOI: 10.1016/S1672-0229(11)60020.
11. Shanker A, Pardasani KR. Mining fuzzy amino acid association patterns in peptide sequences of alphaproteobacteria. Journal of Medical Imaging and Health Informatics. 2013 Sep 1;3(3):380-7.
12. Lopez FJ, Blanco A, Garcia F, Cano C, Marin A. Fuzzy association rules for biological data analysis: a case study on yeast. BMC bioinformatics. 2008 Feb 19;9(1):107.
13. Jain, Amita, and Kamal Raj Pardasani. "Fuzzy-soft-fuzzy set model for mining amino acid associations in peptide sequences of Mycobacterium tuberculosis complex (MTBC)." International Journal of Data Mining and Bioinformatics 17.1 (2017): 1-24.
14. Zadeh, Lotfi A. "Fuzzy sets." Information and control 8.3 (1965): 338-353
15. Yao BX, Liu JL, Yan RX. Fuzzy soft set and soft fuzzy set. InNatural Computation, 2008. ICNC'08. Fourth International Conference on 2008 Oct 18 (Vol. 6, pp. 252-255). IEEE..
16. Gour A, Pardasani KR. Soft Fuzzy Set Approach for Mining Frequent Amino Acid Associations in Peptide Sequences of Dengue Virus. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences.:1-0.
17. Jain A, Pardasani KR. Soft fuzzy model for mining amino acid associations in peptide sequences of Mycobacterium tuberculosis complex. CURRENT SCIENCE. 2016 Feb 25;110(4):603-18.
18. Dandagi GL, Byahatti SM. An insight into the swine-influenza A (H1N1) virus infection in humans. Lung India: Official Organ of Indian Chest Society. 2011 Jan;28(1):34.
19. Goldsmith C. Influenza: The Next Pandemic?. Twenty-First Century Books; 2007.
20. Han J and Micheline K (2006) Data mining: concepts and techniques. Morgan Kaufmann, 2006.
21. Molodtsov D. Soft set theory—first results. Computers & Mathematics with Applications. 1999 Feb 1;37(4-5):19-31.
22. Maji PK. Weighted neutrosophic soft sets approach in a multi-criteria decision making problem. Journal of New Theory. 2015;5:1-2.
23. González AG, Herrador MÁ. A practical guide to analytical method validation, including measurement uncertainty and accuracy profiles. TrAC Trends in Analytical Chemistry. 2007 Mar 31;26(3):227-38.