Soft Set Model for Mining Amino Acid Associations in Peptide Sequences of Mycobacterium Tuberculosis Complex (MTBC)

Amita Jain* & Kamal Raj Pardasani**

*Research Scholar, Department of Computer Application, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, INDIA. E-Mail: amita.jain01{at}gmail{dot}com

**Professor, Department of Mathematics, Bioinformatics and Computer Applications, Maulana Azad National Institute of Technology, Bhopal, Madhya Pradesh, INDIA. E-Mail: kamalrajp{at}rediffmail{dot}com

***Abstract*—**The huge amount of molecular data is available in online biological databases for analysis. This data consist of information which can be used in the field of biomedical industry. One of the major issues is the analysis of this data because the uncertainty in relationships among various fields of this data. There are various algorithms existing for association rule mining but they are not fully capable of addressing the issues of uncertainty in molecular data. Some uncertainty arises due to ignorance of the parameters because objects and their patterns are dependent on the parameters. The degree of relationships among various amino acids present in the molecular sequences depends on the parameters like length ranges and species. In this paper a soft set approach has been proposed for mining amino acid associations in peptide sequences of Mycobacterium tuberculosis complex (MTBC). The soft set has been employed to model the degree of relationships of amino acids with the parameters like length ranges and species. The association rules are generated and used to compute the secondary structures and physicochemical properties of peptide sequences of MTBC. The patterns obtained can be used as signatures which will provide better insights of molecular processes of the disease.

***Keywords*—**Association Rule; Confidence; Data Mining; Soft Set; Support.

***Abbreviations*—**Bayesian Network (BN); Multidrug-Resistant Tuberculosis (MDR-TB); Mycobacterial Interspersed Repetitive Units (MIRU); Mycobacterium Tuberculosis Complex (MTBC).

I. INTRODUCTION

ASSOCIATION RULE MINING is a popular method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness [Piatetsky-Shapiro, 1]. Based on the concept of strong rules, Agrawal et al., [2] introduced association rules for discovering regularities between products in large-scale transaction. Various algorithms are reported in the literature [Panday & Pardasani, 3; 4; Khare et al., 5; Patel et al., 6; Khare et al., 7; 8; Gautam & Pardasani, 9] for association rule mining. Some attempts are also reported in literature of mining association rules in molecular data [Gupta et al., 10; Kuo et al., 11; Wei & Chen, 12]. Various research workers have employed fuzzy set approach for finding patterns in molecular sequences [Francisco et al., 13; Kumari & Pardasani; 14; Shanker & Pardasani, 15; Jain & Pardasani, 16]. Available algorithms for ARM have their own advantages and limitations. There are theories, viz. theory of probability, theory of fuzzy sets [Zadeh, 17], theory of intuitionist fuzzy sets [Atanassov, 18], theory of vague set [Gau & Buehrer, 19], theory of interval mathematics [Moore, 20] and theory of rough sets [Pawlak, 21] which can be used as mathematical tools for dealing with uncertainties. But all these theories have their own limitations. Possibly the reason of these limitations is the inadequacy of the parameterization tool of these theories. To overcome these limitations Molodtsov [22] introduced the concept of soft set as a new mathematical tool for dealing with uncertainties. Herawan & Mustafa [23] proposed soft set based algorithm for association rule mining. From the literature survey it is observed that no attempts are reported for mining soft set based amino acid associations in the peptide sequences of MTBC.

Mycobacterium tuberculosis complex is a cause for chronic infection 'tuberculosis (TB)' which is the leading cause of death from various infectious diseases. The M. tuberculosis complex comprises [Cole, 24] six members*:* M. tuberculosis, the causative agent in the vast majority of human tuberculosis cases*;* Mycobacterium africanum, an agent of human tuberculosis in sub-Saharan Africa; Mycobacterium microti, the agent of tuberculosis in voles; Mycobacterium bovis, which infects a very wide variety of mammalian species including humans, and BCG (bacille

Calmette–Guérin), an attenuated variant of M. bovis; and Mycobacterium canettii, a smooth variant that is very rarely encountered but causes human disease. Increasingly MTBC has developed resistance towards the drugs that cure TB. Globally in 2012, an estimated 450 000 people developed multidrug-resistant TB (MDR-TB) and there were an estimated 170 000 deaths from MDR-TB [25].

Aminian et al., [26] presented a novel Bayesian network (BN) to classify strains of Mycobacterium tuberculosis Complex (MTBC) into six major genetic lineages using mycobacterial interspersed repetitive units (MIRUs), a high-throughput biomarker. Brosch et al., [27] provided an overview of the diversity and conservation of variable regions in a broad range of tubercle bacilli. Shabbeer et al., [28] developed TB-Lineage, an online tool for classification and analysis of strains of Mycobacterium tuberculosis complex.

In this paper an attempt has been made to propose the soft set model for mining amino acid associations in peptide sequences of species of MTBC.

II. MATERIAL AND METHODS

The data of peptide sequences of all species of MTBC have been taken from NCBI [29]. The dataset is filtered to obtain non redundant dataset. This dataset of non redundant sequences is used for mining amino acid association patterns. Total 83086 non-redundant sequences are found which comprises 6176 sequences of mycobacterium africanum, 8011 of mycobacterium bovis, and 5008 of mycobacterium bovis BCG, 39649 of mycobacterium canettii, 28 of mycobacterium microti and 24214 of mycobacterium tuberculosis.

Let each sequence be denoted by transaction T where T∈D and D denote a database [Shanker & Pardasani, 15].

$$T = \{A_i / \forall\ i = 1(1)20\} \forall\ T \in D \quad (1)$$

Where A_i is the i^{th} amino acid.

Soft transaction is denoted by $\check{S}$. Let (F, E) be a soft set over the universe U and $X \subseteq E$ [Shanker & Pardasani, 15].

A soft transaction is defined as given below:

$$\check{S} = \{(A_i, e) / \forall\ i = 1(1)20, e \in X\} \forall\ \check{S} \in \quad (2)$$

Where X is a set of parameters and represents database of soft transactions and e is a parameter.

Association patterns denoted by P_i represent the associations of amino acids. It can be written as [Shanker & Pardasani, 15]:

$$(P_i) = (A_{r1} \cup A_{r2} \cup \ldots A_{ri}) \quad r_, r_i = 1(1)20\ (i,j) = 1(1)20$$
$$For\ i=1\ P_1 = A_{r1} \quad = 1(1)\ 20 \quad (3)$$
$$For\ i=2\ P_2 = A_{r1} \cup A_{r2}\ r \quad = 1(1)\ 20,\ r_2 = 1(1)\ 20$$

Soft association pattern represents the associations of amino acids with their parameter e which is represented as

$$(P_i, e) = ((A_{r1} \cup A_{r2} \cup \ldots A_{ri}), e)\ r_i \neq r_j, r_i = 1(1)20, \quad i = 1(1)\ 20 \quad (4)$$

We can define a set of parameters E with respect to which the amino acids patterns can be explored in these peptide sequences. In the present study we consider a set of parameters as given below:

$$X = \{LR, SP\} \quad (5)$$

Where LR denotes the Length Range of sequences and SP denotes the species of MTBC sequences and SUS denotes susceptibility of species. Here the length range LR is assumed to be in three categories LOW, MEDIUM, HIGH. The length ranges for these three categorical values is determined by the following expression.

$$x = Max\{Lj, j = 1, 2, 3, \ldots\ldots N\} \quad (6)$$
$$y = Min\{Lj, j = 1, 2, 3, \ldots\ldots N\} \quad (7)$$
$$\Delta LR = (x - y) / 3 \quad (8)$$

With the help of expression (6), (7) and (8) length ranges can be given by

$$LR1 = [y, y + \Delta LR] \quad (9)$$
$$LR_2 = [y + \Delta LR + 1, y + 2\ \Delta LR] \quad (10)$$
$$LR_3 = [y + 2\ \Delta LR + 1, x] \quad (11)$$

Where LR1, LR2 and LR3 represent length ranges for LOW, MEDIUM and HIGH. The MTBC has 6 species. Therefore the parameter SP is assigned following values

$$SP = \{SP1, SP2, SP3, SP4, SP5, SP6\} \quad (12)$$

2.1. Soft Set Approach

Each sequence is viewed as a transaction containing twenty amino acids as given in expression (3). For studying the frequent patterns in MTBC sequences frequency can be calculated as:

$$F(A_j) = \sum_i^n f_i(A_j) \quad j = 1to20 \quad (13)$$

Where $f_i\ (A_j)$ is frequency of amino acid in sequence i and F (A_j) is the frequency for amino acid in n sequences.

Cumulative length of all the sequences can be calculated as

$$L = \sum_i^n l\, i \quad (14)$$

Where li is the length of the i^{th} sequence.

Threshold is assumed to be 0.05 as there are 20 amino acids and each will have equal chance of appearing in sequence. The apriori algorithm is employed to find frequent patterns in all the sequences. The first step is to calculate support for all the 20 amino acids in a sequence. The support for single amino acid is calculated by [Shanker & Pardasani, 15].

$$Sup\ (A_j) = F(A_j)/L \quad j = 1to\ 20 \quad (15)$$

Similarly support for all the amino acids present in each species is calculated by expression (15). The amino acids whose support value is greater than threshold value will be the frequent 1-amino acid sets. These frequent 1-amino acid sets are used to generate frequent 2-amino acids sets and similarly frequent *k*-amino acids sets are generated. The Apriori property is used for efficient generation of level -wise *k*- frequent amino acids sets (*k* =1to20).

The SOFT support for *k*-amino acids set is calculated by [Herawan & Mustafa, 23]

$$Sup((A_{r1} \cup A_{r2} \cup \ldots\ldots\ldots A_{ri}), e) = F((A_{r1} \cup A_{r2} \cup \ldots\ldots\ldots A_{rk-1} \cup A_{rk}), e) / \quad (16)$$

Where $A_{r1} \cap A_{r2} \cap \ldots \cap A_{rk} = \phi$ k=1to20 and e is the parameter.

SOFT confidence for *k*- amino acids set is calculated by [Herawan & Mustafa, 23]

$$\text{Conf}\ (A_{r1} \cup A_{r2} \cup \ldots\ldots\ldots A_{rk-1} \cup A_{rk}),\ e) = \frac{\ldots\ldots\ldots}{\cup\ldots\ldots\ldots A}$$

$$\text{Where } A_{r1} \cap A_{r2} \cap \ldots \cap A_{rk} = \phi \ k=1to20 \quad (17)$$

These support and confidence are used to generate soft-set based association patterns in MTBC species.

III. Results and Discussion

The results of soft associations are displayed in Table 1. The differences in the results are highlighted in bold in Table 1 in column 3 and 4. The amino acids A, G, L, V, S, T, R, D, P are predicted as frequent-1 amino acid patterns for (SP1,R1), amino acids patterns A, G, L, V, S, T, P for (SP1, R2) and amino acids patterns A, F, G, I, L, V, N, S, T for (SP1, R3) as shown in column 4 in Table 1. Maximal association patterns are AGLPV, AGLRV, and AGLTV for (SP1, R1); AGLPSTV for (SP1, R2) and AFGLST, AGILST, AGINST, AGLSTV for (SP1, R3) as shown in column 5 in Table 1. Same interpretation can be made for remaining species.

Table 1: Amino Acid Association Patterns among Species of MTBC by Soft Set Approach

Species	Affects in	SOFT ASSOCIATION		
		Species and range	Frequent 1-amino Acid with their Support	Maximal Association Patterns with Their Support
M TUBERCULOSIS (24214)	Vast majority of human tuberculosis cases	SP1 R1 (23669)	A=0.14 G=0.12 L=0.09 V=0.08 S= 0.06 T=0.06 R= 0.07 P=0.06 D=0.05	AGLPV=0.051 **AGLRV=0.055** AGLTV=0.052
		SP1 R2 (442)	A=0.13 G=0.15 L=0.09 V=0.08 S= 0.06 T=0.06 P=0.06	**AGLPSTV=0.052**
		SP1 R3 (103)	A=0.11 **F=0.05** G=0.17 **I=0.06** L=0.09 V=0.08 **N=0.08** S= 0.07 T=0.07	**AFGLST=0.051 AGILST=0.055 AGINST=0.051 AGLSTV=0.053**
M BOVIS (8011)	Wide variety of mammalian species including humans	SP2 R1 (7910)	A=0.13 G=0.13 L=0.09 V=0.08 S= 0.07 T=0.06 R= 0.06 D=0.06 P=0.05 **I=0.05 N=0.05**	ADGLV=0.051 AGLPV=0.052 **AGLRV=0.058** AGLTV=0.053
		SP2 R2 (89)	A=0.14 G=0.18 L=0.09 V=0.07 S= 0.06 T=0.06 D=0.05 P=0.05 R=0.05	AGLST=0.051 ADGLV=0.051 AGLPV=0.052 AGLRV=0.051 AGLSV=0.052 AGLTV=0.052
		SP2 R3 (12)	A=0.13 G=0.13 L=0.09 V=0.08 S= 0.07 T=0.06 R= 0.06 D=0.06 P=0.05 **I=0.05 N=0.05**	**ADGLRV=0.051 ADGLSV=0.051**
M BOVIS BCG (5008)	Wide variety of mammalian species including humans	SP3 R1 (4946)	A=0.13 G=0.10 L=0.10 V=0.08 T=0.06 R= 0.07 D=0.06 P=0.06 S=0.05 **E=0.05**	ADGLV=0.051 AGLTV=0.051
		SP3R2 (88)	A=0.13 **G=0.20** L=0.08 V=0.07 S= 0.05 T=0.06 D=0.05 R=0.05 P=0.05 **E=0.05**	ADGLRV=0.052 **AGLRTV=0.051**
		SP3 R3 (4)	A=0.13 G=0.13 L=0.09 V=0.09 S= 0.07 T=0.06 R= 0.05 D=0.06 P=0.05 **E=0.05 N=0.05**	**ADGLRV=0.054 ADGLSV=0.052**
M CANETTII (13374)	Very rarely encountered but causes human disease	SP4 R1 (13210)	A=0.13 G=0.10 L=0.10 V=0.08 T=0.06 R= 0.07 D=0.06 P=0.06 **E=0.05** S=0.05	ADGLRV=0.051 AGLRTV=0.050
		SP4 R2 (141)	A=0.14 G=0.14 L=0.09 V=0.08 S= 0.06 T=0.06 R= 0.072 D=0.06 P=0.06 **E=0.05**	ADGLPV=0.051 **ADGLRV=0.054** ADGLSV=0.050 AGLPSV=0.051 ADGLTV=0.050 AGLPTV=0.052 AGLSTV=0.051
		SP4 R3 (23)	A=0.10 G=0.17 **I=0.06** L=0.09 V=0.07 **N=0.09** S= 0.07 T=0.07 **F=0.05 P=0.05**	AGILNT=0.050 AFGIST=0.051 AFGLST=0.051 **AGILST=0.057** AGINST=0.052 AGLSTV=0.054 AGLPSV=0.050
M AFRICANUM (6176)	Human tuberculosis in sub-Saharan Africa	SP5 R1(6109)	A=0.13 G=0.10 L=0. 10 V=0.08 T=0.06 R= 0.07 D=0.06 P=0.05 **E=0.05**	ADGLRV=0.051 AGLRTV=0.050
		SP5 R2 (59)	A=0.14 G=0.16 L=0.09 V=0.08 S= 0.06 T=0.07 R= 0.06 D=0.06 P=0.05	**AGLPTV=0.050**
		SP5 R3 (8)	A=0.11 **F=0.05** G=0.15 **I=0.05** L=0.09 V=0.08 **N=0.07** S= 0.07 T=0.07P=0.05	AGILST=0.052 AGLPSV=0.053 AGLSTV=0.052
M MICROTI (28)	tuberculosis in voles	SP6 R1 (13)	A=0.10 G=0.08 L=0.10 V=0.09 S= 0.06 T=0.07 R= 0.07 E=0.08 **I=0.05**	AEGV=0.051 AELV=0.051 **AGLV=0.060 EGLV=0.055** GLTV=0.050
		SP6 R2 (6)	A=0.10 G=0.09 L=0.10 V=0.10 T=0.08 R=0 .06 D=0.07 E=0.07 P=0.05	**ADEGLTV=0.052 ADGLRTV=0.052**
		SP6 R3 (**9**)	A=0.15 G=0.11 L=0.08 V=0.07 S= 0.07 T=0.06 P=0.06	**AGLPST=0.052 AGLSTV=0.052**

The Table 2 presents the probable secondary structure of proteins present in all the species of MTBC for soft approach. In R1, R2 and R3 ranges the frequent-1 pattern, frequent-2 patterns and frequent-3 patterns of amino acid associations supporting formation of helix, coil and sheet are predicted in Table 2. The results displayed in Table 2 indicate that all the species have tendency to form helix structure frequently for all three ranges. For ranges R1 and R2 higher frequent amino acid patterns are not found for sheet and coil formation so the tendency to form helix structure frequently is justified.

Table 2: Probable Secondary Structure and their Responsible 1, 2 and 3 Amino Acid Association Patterns by Soft Approach

Species	**Range**	**Helix M,A,L,E,K,R,Q,H**			**Sheet V,I,T,C,W,F,Y**			**Coil N,D,P,S,G**		
		1F	2F	3F	1F	2F	3F	1F	2F	3F
M TB	**R1**	A,L,R	AL,AR,LR	ALR	V,T	TV	NONE	G,D,P,S	DG, GP,GS	NONE
	R2	A,L,R	AL,AR,LR	ALR	V,T	TV	NONE	G,D,P,S	GP,**DG**, GS,**PS**	GPS
	R3	A,L	AL	NONE	V,T,**F,I**	**FT,IT**,TV	**NONE**	G,D,P,S,N	GN,GP, GS, PS,NS	**GNS,GPS**
M BOVIS	**R1**	A,L,R	AL,AR,LR	ALR	V,T, I	TV	NONE	G,D,P,S,N	DG,GP, GS	NONE
	R2	A,L,R	AL,AR,LR	ALR	V,T	TV	NONE	G,D,P,S	DG,GP, GS	NONE
	R3	A,L,**R**	AL,AR,LR	ALR	V,T, I	TV	NONE	G,D,P,S, N	**DG, DS**,GP, GS**, PS**	GPS
M BOVIS BCG	**R1**	A,L,R, E	AL,AR,LR	ALR	V,T	TV	NONE	G,D,P,S,N	DG,GP, GS	NONE
	R2	A,L,R, E	AL,AR,LR	ALR	V,T	TV	NONE	G,D,P,S,N	DG,GP, GS	NONE
	R3	A,L,**R, E**	AL,AR,LR	ALR	V,T	TV	NONE	G,D,P,S,N	**DG, DS**,GS	DGS
M AFRICANUM	**R1**	A,L,R	AL,AR,LR	ALR	V,T	TV	NONE	G,D,P,S	DG, GP,GS	NONE
	R2	A,L,R	AL,AR,LR	ALR	V,T	TV	NONE	G,D,P,S	GP,**DG**, GS,**PS**	GPS
	R3	A,L	AL	NONE	V,T, **I,F**	**FT,IT**,TV	**FIT**	G,D,P,S,N	GN,GP, GS, PS,NS	**GNS,GPS**
M CANETTII	**R1**	A,L,R, E	AL,AR,LR	ALR	V,T	TV	NONE	D,G,P,S	DG,GP, GS	NONE
	R2	A,L,R, E	AL,AR**,LR**	ALR	V,T	TV	NONE	D,P,G,S	DG,**DP, DS**,GP, GS,**PS**	**DGP,DGS,GPS**
	R3	A,L	AL	NONE	V,T,**F,I**	**FI,FT,IT,** TV	**FIT**	G,P,S,N	GN,GP, GS,NS, PS	**GNS, GPS**
M MICROTI	**R1**	A,E,L, R	AE,AL,AR,EL, LR	AEL, ALR	V,T,I	TV	NONE	G,**S**	GS	NONE
	R2	A,E,L, R	AE,AL,AR,EL, LR	AEL, ALR	V,T	TV	NONE	G,**P**,S,**D**	GP,DP, DG	DGP
	R3	A,L	AL	NONE	V,T	TV	NONE	G,P,S	GP,GS, PS	GPS

Table 3 describes the physiochemical properties of maximal frequent amino acid patterns in R1, R2 and R3 for all species of MTBC by soft approach.

Table 3: Maximal frequent amino acids in all species of MTBC

Species	Range	M frequent Patterns	No of M fre patterns	Maximum association patterns	Physiochemical properties
M TUBERCULOSIS	R1	5F	3	AGLPV, AGLRV, AGLTV	Hydrophobic, polar uncharged positively charged
	R2	7F	1	**AGLPSTV**	Hydrophobic, polar uncharged positively charged
	R3	6F	5	**AFGLST,AGILST,AGINST,AGLSTV**	Hydrophobic, positively charged
M BOVIS	R1	5F	4	ADGLV, AGLPV, **AGLRV,** AGLTV	Hydrophobic, polar uncharged positively charged
	R2	5F	6	AGLST,ADGLV,AGLPV,AGLRV,AGLSV, AGLTV	Hydrophobic, positively charged, polar uncharged
	R3	6F	2	ADGLRV,ADGLSV	Hydrophobic
M BOVIS BCG	R1	5F	2	ADGLV, AGLTV	Hydrophobic, polar uncharged
	R2	6F	2	ADGLRV,AGLRTV	Hydrophobic, polar uncharged,
	R3	6F	2	ADGLRV, ADGLSV	Hydrophobic
M CANETTII	R1	6F	2	ADGLRV, AGLRTV	Hydrophobic, polar uncharged
	R2	6F	7	ADGLPV, **ADGLRV,** ADGLSV, AGLPSV, ADGLTV, AGLPTV, AGLSTV	Hydrophobic, positively charged, polar uncharged
	R3	6F	7	AGILNT, AFGIST, AFGLST, **AGILST,** AGINST, AGLSTV, AGLPSV	Hydrophobic
M AFRICANUM	R1	6F	2	ADGLRV, AGLRTV	Hydrophobic, polar uncharged,
	R2	6F	1	AGLPTV	Hydrophobic, polar uncharged,
	R3	6F	3	AGILST, AGLPSV, AGLSTV	Hydrophobic
M MICROTI	R1	4F	5	AEGV, AELV, **AGLV, EGLV,** GLTV	Hydrophobic, polar uncharged,
	R2	7F	2	**ADEGLTV, ADGLRTV**	Hydrophobic, polar uncharged,
	R3	6F	2	**AGLPST, AGLSTV**	Hydrophobic

The association rules are generated on the basis of above study are given below:

Range 1

Rules Applied for Helix Formation:

1. {A(Frequent)/\L(Frequent)}→This rule applies for all species.
2. {L(Frequent)/\R(Frequent)}→ This rule applies for all species.
3. {A(frequent)/\R(frequent)}→ This rule applies for all species.
4. {E(Frequent)/\L(Frequent)}→ This rule applies for M bovis BCG, M Canettii and M Microti species.
5. {A(Frequent)/\L(Frequent)/\R(Frequent)} →This rule applies for all species.
6. {A(Frequent)/\L(Frequent)/\E(Frequent)} →This rule applies for M Microti species.

Rules Applied for Sheet Formation:

1. {V(frequent)/\T(frequent)} → This rule applies for all species.

Rules Applied for Coil Formation:

1. {G(Frequent)/\S(Frequent)}→ This rule applies for all species.
2. {D(Frequent)/\G(Frequent)}→ This rule applies for all species except M Microti species.
3. {G(Frequent)/\P(Frequent)}→ This rule applies for all species except M Microti species.

Range 2:

Rules Applied for Helix Formation:

1. {A(Frequent)/\L(Frequent)}→This rule applies for all species.
2. {L(Frequent)/\R(Frequent)}→ This rule applies for all species.
3. {A(frequent)/\R(frequent)}→ This rule applies for all species.
4. {E(Frequent)/\L(Frequent)}→ This rule applies for M Microti species.
5. {A(Frequent)/\L(Frequent)/\R(Frequent)} →This rule applies for all species.
6. {A(Frequent)/\L(Frequent)/\E(Frequent)} →This rule applies for M Microti species.

Rules Applied for Sheet Formation:

1. {V(frequent)/\T(frequent)} → This rule applies for all species.

Rules Applied for Coil Formation:

1. {G(Frequent)/\S(Frequent)}→ This rule applies for all species.
2. {D(Frequent)/\G(Frequent)}→ This rule applies for all species except M Microti species.
3. {G(Frequent)/\P(Frequent)}→ This rule applies for all species except M Microti species.
4. {P(Frequent)/\S(Frequent)}→ This rule applies for M TB and M Africanum species.

5. {G(Frequent)/\P(Frequent) /\S(Frequent }→ This rule applies for M TB, M Africanum and M Canettii species.
6. {D(Frequent)/\G(Frequent)/\P(Frequent }→ This rule applies for M Canettii and M Microti species.

Range 3:

Rules Applied for Helix Formation:

1. {A(Frequent)/\L(Frequent)}→This rule applies for all species.
2. {L(Frequent)/\R(Frequent)}→ This rule applies for M Bovis and M Bovis BCG species.
3. {A(frequent)/\R(frequent)}→ This rule applies for M Bovis and M Bovis BCG species.
4. {A(Frequent)/\L(Frequent)/\R(Frequent)} → This rule applies for M Bovis and M Bovis BCG species.

Rules Applied for Sheet Formation:

1. {V(frequent)/\T(frequent)} → This rule applies for all species.
2. {F(frequent)/\T(frequent)} → This rule applies for M TB, M Canettii and M Africanum species.
3. {I(frequent)/\T(frequent)} → This rule applies for M TB, M Canettii and M Africanum species.
4. {F(frequent)/\I(frequent)} → This rule applies M Canettii species.
5. {F(Frequent)/\I(Frequent)/\T(Frequent)} → This rule applies for M Canettii and M Africanum species.

Rules Applied for Coil Formation:

1. {G(Frequent)/\S(Frequent)}→ This rule applies for all species.
2. {G(Frequent)/\P(Frequent)}→ This rule applies for all species except M Bovis BCG species.
3. {P(Frequent)/\S(Frequent)}→ This rule applies for all species except M Bovis BCG species.
4. {G(Frequent)/\N(Frequent)}→ This rule applies for M TB, M Canettii and M Africanum species.
5. {N(Frequent)/\S(Frequent)}→ This rule applies for M TB, M Canettii and M Africanum species.
6. {D(Frequent)/\G(Frequent)}→ This rule applies for M Bovis BCG species.
7. {D(Frequent)/\G(Frequent)/\S(Frequent }→ This rule applies for M Bovis BCG species.
8. {G(Frequent)/\P(Frequent) /\S(Frequent }→ This rule applies for all species except M Bovis BCG species.
9. {G(Frequent)/\N(Frequent)/\S(Frequent }→ This rule applies for M TB, M Canettii and M Africanum species.

IV. CONCLUSION

The soft set approach is proposed and employed to mine amino acid association patterns in the peptide sequences of species of MTBC. Soft set approach takes care of uncertainties due to dependence of patterns on parameters and thus is superior to the fuzzy set and ordinary set approaches. The interesting association rules among amino acid of species of MTBC have been generated by the soft approach. Also the physicochemical properties and secondary structures have been predicted based on amino acid association patterns in peptide sequences of species of MTBC. The association patterns and rules generated can serve as signatures for getting better insights of the molecular processes in species of MTBC.

ACKNOWLEDGMENT

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

REFERENCES

[1] G. Piatetsky-Shapiro (1991), "Discovery, Analysis, and Presentation of Strong Rules", Editors: G. Piatetsky-Shapiro & William J. Frawley, *Knowledge Discovery in Databases*, *AAAI/MIT Press*, Cambridge, MA.

[2] R. Agrawal, T. Imielinski & A.N. Swami (1993), "Mining Association Rules between Sets of Items in Large Databases", *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Vol. 22, No. 2, Pp. 207–216.

[3] A. Panday & K.R. Pardasani (2009), "Rough Set Model for Discovering Multidimensional Association Rules", *International Journal of Computer Science and Network Security*, Vol. 9, No. 6, Pp. 159–164.

[4] A. Panday & K.R. Pardasani (2009), "PPCI Algorithm for Mining Temporal Association Rules in Large Database", *Journal of Information & Knowledge Management*, Vol. 8, No. 04, Pp. 345–352.

[5] N. Khare, N. Adlakha & K.R. Pardasani (2009), "Karnaugh Map Model for Mining Association Rules in Large Databases", *International Journal of Computer and Network Security*, Vol. 1, No. 1, Pp. 16–21.

[6] R. Patel, D.K. Swami & K.R. Pardasani (2006), "Lattice based Algorithm for Incremental Mining of Association Rules", *International Journal of Theoretical and Applied Computer Sciences*, Vol. 1, No. 1, Pp. 119–128.

[7] N. Khare, N. Adlakha & K.R. Pardasani (2009), "An Algorithm for Mining Multidimensional Fuzzy Association Rules", *International Journal of Computer Science and Information Security (IJCSIS)*, Vol. 5, P;p. 72–76.

[8] N. Khare, N. Adlakha & K.R. Pardasani (2010), "A Fuzzy based Model for Mining Conditional Hybrid Dimensional Association Rules", *International Journal of Data Mining and Knowledge Engineering*, Vol. 2, No. 5, Pp. 69–76.

[9] P. Gautam & K.R. Pardasani (2010), "A Novel Approach for Discovery of Multilevel Fuzzy Association Rules", *Journal of Computing*, Vol. 2, No. 3, Pp. 56–64

[10] N. Gupta, N. Mangal, K. Tiwari and P. Mitra (2006), "Mining Quantitative Association Rules in Protein Sequences", *Lecture Notes in Computer Science*, Vol. 3755, Pp. 273–281.

[11] H.C. Kuo, P.L. Ong, J.C. Lin & J.P. Huang (2011), "Discovering Amino Acid Patterns on Binding Sites in Protein Complexes", *Bio information*, Vol. 6, No. 1, Pp. 10–14.

[12] Q. Wei & G. Chen (1999), "Mining Generalized Association Rules with Fuzzy Taxonomic Structures", *Proceedings of the 18th International Conference of the North American Fuzzy Information Processing Society(NAFIPS)*, NY, USA, Pp. 477–481.

[13] J.L. Francisco, B. Armando, G. Fernando, C. Carlos & M. Antonio (2008), "Fuzzy Association Rules for Biological Data Analysis: A Case Study on Yeast", *BMC Bioinformatics*, Vol. 9, Pp. 107.

[14] T. Kumari & K.R. Pardasani (2012), "Mining Fuzzy Associations among Amino Acids of Class A GPCRs", *Online Journal of Bioinformatics*, Vol. 13, No. 2, Pp. 202–213.

[15] A. Shanker & K.R. Pardasani (2013), "Mining Fuzzy Amino Acid Association Patterns in Various Orders of Class Alphaproteobacteria", *Journal of Medical Imaging and Health Informatics*, Vol. 3, Pp. 380–3287.

[16] A. Jain & K.R. Pardasani (2015), "Mining Fuzzy Amino Acid Associations in Peptide Sequences of Mycobacterium Tuberculosis Complex (MTBC)", *Network Modeling Analysis in Health Informatics and Bioinformatics*, Vol. 4, No. 1, Pp. 1–14.

[17] L.A. Zadeh (1965), "Fuzzy Sets", *Information and Control***,** Vol. 8, No. 3, Pp. 338–353.

[18] K.T. Atanassov (1986), "Intuitionistic Fuzzy Sets", *Fuzzy Sets and Systems*, Vol. 20, No. 1, Pp. 87–96.

[19] W.L. Gau & D.J. Buehrer (1993), "Vague Sets", *IEEE Transactions on Systems Man and Cybernetics*, Vol. 23, No. 2, Pp. 610–614.

[20] R. Moore (1996), "Interval Arithmatic", Prentice-Hall, Englewood Cliffs, NJ, USA.

[21] Z. Pawlak (1982), "Rough Sets", *International Journal of Computer Science and Information*, Vol. 11, Pp. 341–356.

[22] D. Molodtsov (1999), "Soft Set Theory-First Results", *Computers and Mathematics with Applications*, Vol. 37, Pp. 19–31.

[23] T. Herawan & M.D. Mustafa (2011), "A Soft Set Approach for Association Rules Mining", *Knowledge-based Systems*, Vol. 24, Pp. 186–195.

[24] T.S. Cole (2002), "Comparative and Functional Genomics of the Mycobacterium Tuberculosis Complex", *Microbiology*, Vol. 148, No. 10, Pp. 2919–2928.

[25] WHO Report 2013 – Global tuberculosis report.

[26] M. Aminian, A. Shabbeer & K.P. Bennett (2009), "Determination of Major Lineages of Mycobacterium Tuberculosis Complex using Mycobacterial Interspersed Repetitive Units", *IEEE International Conference on Bioinformatics and Biomedicine*, Pp. 338–343.

[27] R. Brosch, S.V. Gordon, M. Marmiesse, P. Brodin, C. Buchrieser, K. Eiglmeier, T. Garnier, C. Gutierrez, G. Hewinson, K. Kremer, L.M. Parsons, A.S. Pym, S. Samper, D. van Soolingen & S.T. Cole (2002), "A New Evolutionary Scenario for the Mycobacterium Tuberculosis Complex". *Proceedings of the National Academy of Sciences*, USA, Vol. 99, No. 3, Pp. 684–3689.

[28] A. Shabbeer, L.S. Cowan, C. Ozcaglar, N. Rastogi, S.L. Vandenberg, B. Yener & K.P. Bennett (2012), "TB-Lineage: An Online Tool for Classification and Analysis of Strains of Mycobacterium Tuberculosis Complex", *Infection, Genetics and Evolution*, Vol. 12, No. 4, Pp. 789–797.

[29] NCBI (National Center for Biotechnology Information), http://www.ncbi.nlm.nih.gov/.

Amita Jain, is a research scholar in department of computer application, Maulana Azad National Institute of technology; Bhopal (MP). She received her Master's degree in Information Technology at Dr. H. S. Gour University, Sagar, Madhya Pradesh, India. She has published one paper in Network Modeling Analysis in Health Informatics and Bioinformatics, Springer.

Kamal Raj Pardasani is a professor at Department of Mathematics, Maulana Azad National Institute of Technology, MANIT, Bhopal, M.P, India. He Obtained Ph.D. in area of Computational Biology in 1988 from School of Mathematics and Allied Sciences, Jiwaji University, Gwalior. His specializations are Computational Biology, Bioinformatics, Biomathematics, Biocomputing, Data Warehousing & Data Mining, Soft Computing, Finite Element Modeling, Financial Informatics and Law Informatics. He supervised Twenty five Ph.D Dissertations in the field of Mathematics, Bioinformatics, Computer Applications and Mathematical and Computational Bio-Sciences.

