Soft fuzzy model for mining amino acid associations in peptide sequences of *Mycobacterium tuberculosis* complex

Amita Jain[1,*] and Kamal Raj Pardasani[2]

[1]Department of Computer Application, and
[2]Department of Mathematics, Bioinformatics and Computer Applications, Maulana Azad National Institute of Technology, Bhopal 462 003, India

Analysis of biological data plays an important role in medical and bioinformatics industry. However, uncertainty in this biological information is the most unavoidable challenge of this era. The existing algorithms for association rule mining are inadequate to address the issues of uncertainty in the molecular data. Variation in the length of the sequences leads to variation in the degree of relationships among amino acids. Ignorance of the parameters leads to uncertainty due to the dependencies of the objects and their patterns on the parameters. The degree of relationships among various amino acids present in the molecular sequences also depends on the parameters like length ranges and species, etc. In this article, a soft fuzzy set approach has been proposed for mining fuzzy amino acid associations in peptide sequences of *Mycobacterium tuberculosis* complex (MTBC). The approach is employed to incorporate the degree of relationships among amino acids present in the peptide sequences. The soft sets are employed to model relationships of amino acids with the parameters like length range, species etc. The amino acid associations and their relationships with various parameters in the peptide sequences of MTBC obtained in the present study will be of great use in developing signatures that will provide better insights into the structures, functions and interactions of proteins.

Keywords: Association rule, complex, data mining, fuzzy and soft sets, *Mycobacterium tuberculosis*.

THE biological databases have grown at an enormous rate during the last two decades. This huge volume of biological data provides new opportunities and challenges for analysis to generate new information and knowledge. These have led to the development of methods for collection, storage, management and data mining. The different types of data mining techniques like classification, clustering and association rule mining are available for analysis of data. The concept of association rule mining (ARM) is used in large-scale transactional databases for discovering regularities between products[1]. *A priori* algorithm has been reported as the first algorithm for finding frequent itemsets[2]. Since then, a number of algorithms for ARM have been reported in the literature[3–6]. The available algorithms for ARM have their own advantages and limitations. Kocatas *et al.*[7] employed ARM to study patterns responsible for protein–protein interactions. Rodríguez *et al.*[8] found a relationship between protein sequences and protein features using a modified version of the *a priori* algorithm. Oyama *et al.*[9] implemented ARM to annotate the proteins within the protein–protein interaction network. Kuo *et al.*[10] employed the ARM technique to discover amino acid association patterns on the binding site of a protein complexes and their work is mainly focused on protein complexes which have protein–protein recognition. Capability of handling inherent uncertainties present in the biological data is the major challenge as the existing methods are not fully capable of addressing the issues of uncertainty. The fuzzy set approach is capable of addressing the issues of uncertainty arising due to degree of relationship among amino acids of peptide sequences. Some algorithms are reported in the literature for fuzzy ARM[11–14]. Various researchers have employed fuzzy set approach for finding patterns in molecular sequences[15,16]. Some models have been proposed for mining fuzzy association rules in peptide sequences of class A GPCRs (G protein-coupled receptors), B GPCRs and Alphaproteobacteria[17–19]. Molodtsov[20] proposed the soft set approach which is capable of handling the uncertainties arising due to non-consideration of parameters. Herawan and Mustafa[21] proposed soft set-based algorithm for ARM[21]. From the literature survey it is observed that there are no attempts for mining soft fuzzy amino acid associations in the peptide sequences of *Mycobacterium tuberculosis* complex (MTBC).

According to the World Health Organization, 8.6 million people fell ill with tuberculosis (TB) in 2012 and 1.3 million individuals worldwide have lost their lives from due to the disease[22]. Mycobacteria that cause human and/or animal TB are grouped together within the MTBC which comprises of six members: *Mycobacterium tuberculosis*, *Mycobacterium africanum*, *Mycobacterium microti*, *Mycobacterium bovis*, *Mycobacterium bovis* BCG (bacille Calmette–Guérin), an attenuated variant of

*For correspondence. (e-mail: amita.jain01@gmail.com)

M. bovis and *Mycobacterium canettii*[23]. Increasingly, MTBC has developed resistance towards the drugs that cure TB. Diagnostics, chemotherapy and vaccination are available. However, the disease is far from being eradicated. Globally in 2012, an estimated 450,000 people developed multidrug-resistant TB (MDR-TB) and there were 170,000 deaths estimated from the disease[22].

Saravanan and Selvaraj[24] proposed a model to provide insights into protein folding, design and function. They revisited the mechanism of structural plasticity in unrelated proteins with increased number of structures in the Protein Data Bank by comparing identical octapeptides in unrelated proteins with the dictionary of protein secondary structure extracted from existing experimental data.

Uthayakumar *et al.*[25] studied the ambiguity between sequence–structure relationships in *M. tuberculosis* and examined if sequentially identical peptide fragments adopt similar three-dimensional structures. Brosch *et al.*[26] provided an overview of the diversity and conservation of variable regions in a broad range of tubercle bacilli. Shabbeer *et al.*[27] developed TB-Lineage, an on-line tool for classification and analysis of strains of *M. tuberculosis* complex.

Several molecular sequences of MTBC are available in various biological databases. The associations of amino acids present in peptide sequences of MTBC can be explored and analysed to generate information which will be crucial in understanding the structure, function, binding sites, biochemical properties, protein folding and interactions of these peptide sequences.

In this article we propose the soft fuzzy model for mining amino acid associations in peptide sequences of species of MTBC.

Materials and methods

The data of peptide sequences of all species of MTBC have been taken from NCBI[28]. The sequences with 100% similarity were considered as redundant. Hence these identical sequences were removed to construct a non-redundant dataset. This dataset was then used for mining amino acid association patterns. Totally 83,086 non-redundant sequences were found which comprised of 6176 sequences of *M. africanum*, 8011 of *M. bovis*, 5008 of *M. bovis* BCG, 39,649 of *M. canettii*, 28 of *M. microti* and 24,214 sequences of *M. tuberculosis*.

Figure 1 shows the proposed methodology employed.

A description of the methodology is given below.

Let each sequence be denoted by transaction T, where $T \in D$ and D denotes a database.

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

where A_i is the ith amino acid.

T is transformed into fuzzy transaction $\bar{T}$ as[19]

$$\bar{T} = \{(A_i, \mu_i(A)) | \forall i = 1(1)20\}\ \ \forall \bar{T} \in D', \quad (2)$$

where $\mu_i(A)$ is fuzzy membership of each amino acid in each transaction and D' represents the database of fuzzy transactions.

Soft transaction is denoted by $\check{S}$. Let (F, E) be a soft set over the universe U and $X \sqsubseteq E$.

A soft transaction is defined as[21]

$$\check{S} = \{(A_i, e) \mid \forall i = 1(1)20,\ e \in X\} \quad \forall\ \check{S} \in D^*, \quad (3)$$

where X is a set of parameters, and D^* represents the database of soft transactions and e is a parameter.

A soft fuzzy transaction is defined as

$$\bar{\check{S}} = \{(A_i, \mu_i(A), e) \mid \forall i = 1(1)20,\ e \in X\}\ \forall\ \bar{\check{S}} \in \bar{D}. \quad (4)$$

Here $\bar{D}$ represents the database of soft fuzzy transaction.

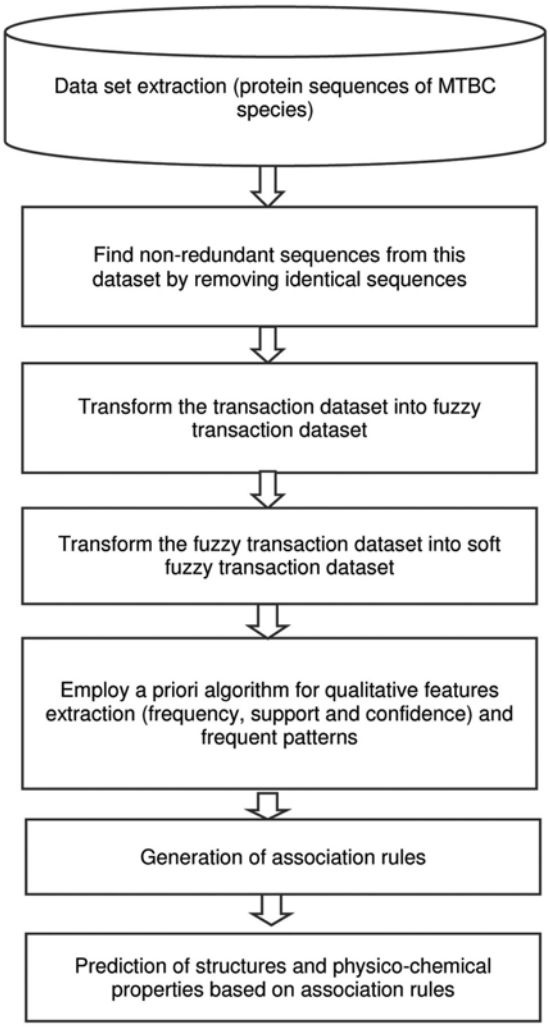

Figure 1. Methodology.

Association patterns denoted by P_i represent the associations of amino acids. They can be written as

$$(P_i) = (A_{r_2} \cup A_{r_1} \cup \cdots A_{r_i}),\ r_i \neq r_j, r_i = 1(1)20,\ (i,\ j) = 1(1)20.$$
$$\text{For } i = 1,\ P_1 = A_{r_1},\ r_1 = 1(1)20.$$
$$\text{For } i = 2,\ P_2 = A_{r_1} \cup A_{r_2},\ r_1 \neq r_2,\ r_1 = 1(1)20,\ r_2 = 1(1)20. \quad (5)$$

Fuzzy association pattern represents the association of amino acids and the fuzzy membership of each association pattern that is calculated for each amino acid and can be expressed as given below:

$$(P_i,\ (P_j)) = ((A_{r_1} \cup A_{r_2} \cup \cdots A_{r_i}),\ (P_i)),$$
$$r_i \neq r_j, r_i = 1(1)20,\ i = 1(1)20,$$
$$(P_i) = \text{Min}\{(A_{r_1}),\ (A_{r_2}) \cdots \mu(A_{r_i})\},$$
$$r_i \neq r_j, r_i = 1(1)20,\ i = 1(1)20, \quad (6)$$

where P_i are the association patterns of amino acids, and $\mu(P_i)$ is the membership of association of amino acids in pattern P_i.

Soft association pattern represents the association of amino acids with their parameter e, which is expressed as

$$(P_i, e) = ((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_i}), e),$$
$$r_i \neq r_j, r_i = 1(1)20,\ i = 1(1)20. \quad (7)$$

Soft fuzzy association pattern represents the fuzzy association of amino acids with their parameter e, which is expressed as

$$(P_i, \mu(P_i), e) = ((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_i}), e(P_i), e),$$
$$r_i \neq r_j, r_i = 1(1)20,\ i = 1(1)20. \quad (8)$$

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 $X \subset E$ as given below

$$X = \{\text{LR, SP}\} \quad (9)$$

where LR denotes the length range of sequences and SP denotes the species of MTBC sequences. Here LR is assumed to be divided into three categories, low, medium and high. The length ranges for these three categories are determined by the following expressions:

$$x = \max\{L_j, j = 1, 2, 3, \ldots, N\}, \quad (10)$$

$$y = \min\{L_j, j = 1, 2, 3, \ldots, N\}, \quad (11)$$

$$\Delta\text{LR} = (x - y)/3. \quad (12)$$

With the help of eqs (11)–(13) length ranges can be given by

$$\text{LR}_1 = [y, y + \Delta\text{LR}], \quad (13)$$

$$\text{LR}_2 = [y + \Delta\text{LR} + 1, y + 2\Delta\text{LR}], \quad (14)$$

$$\text{LR}_3 = [y + 2\Delta\text{LR} + 1, x], \quad (15)$$

where LR_1, LR_2 and LR_3 represent length ranges for low, medium and high respectively. The MTBC has six species. Therefore, the parameter SP is assigned following values

$$\text{SP} = \{\text{SP1, SP2, SP3, SP4, SP5, SP6}\}. \quad (16)$$

Crisp approach

Each sequence is viewed as a transaction containing 20 amino acids as given in eq. (3). For studying the frequent patterns in MTBC sequences the frequency can be calculated as

$$F(A_j) = \sum_i^n f_i(A_j),\ \ j = 1 \text{ to } 20. \quad (17)$$

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

Cumulative length of all the sequences can be calculated as

$$L = \sum_i^n l_i, \quad (18)$$

where l_i is the length of the ith sequence.

Threshold is assumed to be 0.05 as there are 20 amino acids and each will have equal chance of appearing in the sequence. The *a priori* 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 a single amino acid is calculated by

$$\text{Sup}(A_j) = F(A_j)/L, \qquad j = 1 \text{ to } 20 \quad (19)$$

Similarly, support for all the amino acids present in each species is calculated using eq. (19). The amino acids whose support value is greater than the threshold value will be the frequent 1-amino acid sets. These are used to generate frequent 2-amino acids sets and similarly,

frequent *k*-amino acids sets are generated. The *a priori* property is used for efficient generation of level-wise *k*-frequent amino acids sets (*k* = 1 to 20).

Soft approach

The soft support for *k*-amino acids set is calculated by

$$\text{Sup}((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_i}), e) = F((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_{k-1}} \cup A_{r_k}), e)/L, \quad (20)$$

where $A_{r_1} \cap A_{r_2} \cap \ldots A_{r_k} = \phi$, k = 1 to 20 and e is a parameter.

Soft confidence for *k*-amino acids set is calculated by

$$\text{Conf}((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_{k-1}} \cup A_{r_k}), e) = \frac{\sum_i^n F_i((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_{k-1}} \cup A_{r_k}), e)}{\sum_i^n F_i((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_{k-1}}), e)}, \quad (21)$$

where $A_{r_1} \cap A_{r_2} \cap \ldots A_{r_k} = \phi$, k = 1 to 20.

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

Fuzzy set approach

The fuzzy membership of each amino acid is calculated as

$$\mu_i(A) = \sum_i^n f_i(A)/L_i, \quad (22)$$

where L_i is length of the ith sequence, $\sum_i^n f_i(A)$ is frequency of amino acid A in sequence i, and $\mu_i(A)$ is the membership of amino acid A in the ith sequence.

The amino acids that have fuzzy membership value above minimum support value are said to be frequent. The threshold is estimated based on the assumption that there are 20 amino acids and each will have equal chance of appearing in a sequence. Thus the value of membership threshold is assumed to be 0.05.

The *a priori* algorithm is employed to find frequent patterns in all the sequences. The first step is to calculate fuzzy support for all the 20 amino acids in a sequence. The fuzzy support for a single amino acid is calculated by:

$$\text{Sup}(A_j) = \sum_{i=1}^n \mu_i(A_j)/n, \quad j = 1 \text{ to } 20. \quad (23)$$

Similarly, fuzzy support for all the amino acids present in each species is calculated using eq. (23). The amino acids whose support value is greater than the threshold value will be the frequent 1-amino acid sets. These are used to generate frequent 2-amino acids sets and similarly, frequent *k*-amino acids sets are generated. The *a priori* property is used for efficient generation of level-wise *k*-frequent amino acids sets (*k* = 1 to 20).

Soft fuzzy set approach

The soft fuzzy support for *k*-amino acids set is calculated by:

$$\text{Sup}((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_{k-1}} \cup A_{r_k}), e) = \sum_{i=1}^n \mu_i((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_{k-1}} \cup A_{r_k}), e)/n, \quad (24)$$

where $A_{r_1} \cap A_{r_2} \cap \ldots A_{r_k} = \phi$, k = 1(1)20.

Soft fuzzy confidence for *k*-amino acids set is calculated by:

$$\text{Conf}((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_{k-1}} \cup A_{r_k}), e) = \frac{\sum_i^n \mu_i((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_{k-1}} \cup A_{r_k}), e)}{\sum_i^n \mu_i((A_{r_1} \cup A_{r_2} \cup \ldots A_{r_{n-1}}), e)}, \quad (25)$$

where $A_{r_1} \cap A_{r_2} \cap \ldots A_{r_k} = \phi$, k = 1 to 20.

A program has been developed in php language for the above method to perform ARM in peptide sequences of MTBC.

Results and discussion

Table 1 shows the results of fuzzy, soft and soft fuzzy associations. The differences in the results are highlighted in bold in Table 1 in columns 2–7. By comparing columns 2, 5 and 7 in Table 1, the change in frequent-1 patterns for (SP1, R1), (SP1, R2), (SP1, R3), (SP2, R3), (SP3, R1), (SP3, R2), (SP3, R3), (SP4, R1), (SP4, R2), (SP4, R3), (SP5, R1), (SP5, R3), (SP6, R1), (SP6, R2) and (SP6, R3) are observed by the three approaches. Comparing columns 3, 6 and 9 in Table 1, we can observe the change in maximal patterns of all the species in different ranges obtained by the three approaches. There is significant difference in the results obtained using the three approaches. The amino acids A, G, L,V, S, T, R, D, P are predicted as frequent 1-amino acid patterns by fuzzy set approach, soft set approach for (SP1, R1) and soft fuzzy approach for (SP1, R1), as shown in columns 2, 5 and 8 respectively in Table 1. But R and D are not predicted as frequent-1 by soft set approach for (SP1, R2) in column 5, Table 1. Again, for (SP1, R3) amino acids R, D and P are not predicted as frequent 1-amino acids by

Table 1. Amino acid association patterns among species of *Mycobacterium tuberculosis* complex (MTBC) by fuzzy, soft and soft fuzzy approaches

	Fuzzy association		Soft association			Soft fuzzy association		
Species	Frequent 1-amino acid with their support	Maximal association patterns with their fuzzy support	Species and range	Frequent 1-amino acid with their support	Maximal association patterns with their support	Species and range	Frequent 1-amino acid with their support	Maximal association patterns with their support
Mycobacterium tuberculosis (24,214)	A = 0.14 G = 0.13 L = 0.09 V = 0.08 S = 0.06 T = 0.06 R = 0.07 D = 0.05 P = 0.06	AGLR = 0.053 AGLS = 0.050 AGLT = 0.052 AGLV = 0.062 AGRV = 0.051 ALRV = 0.052	SP1 R1 (23,669)	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 R1 (23,669)	A = 0.14 G = 0.13 L = 0.09 V = 0.08 S = 0.06 T = 0.06 R = 0.07 P = 0.06 D = 0.05	AGLR = 0.053 AGLT = 0.051 AGLV = 0.062 AGRV = 0.051 ALRV = 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 R2 (442)	A = 0.13 G = 0.15 L = 0.09 V = 0.08 S = 0.06 T = 0.06 P = 0.06 D = 0.05	AGLPSTV = 0.051
			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	SP1 R3 (103)	A = 0.10 **F = 0.05** G = 0.17 I = 0.06 L = 0.08 V = 0.07 N = 0.09 S = 0.07 T = 0.07 **P = 0.05**	**AGILNST = 0.050**
Mycobacterium bovis (8011)	A = 0.14 G = 0.11 L = 0.09 V = 0.08 S = 0.06 T = 0.06 R = 0.08 D = 0.06 P = 0.06	AGLRV = 0.055	SP2 R1 (7910)	A = 0.13 G = 0.11 L = 0.09 V = 0.08 S = 0.06 T = 0.06 R = 0.07 D = 0.06 P = 0.06	ADGLV = 0.051 AGLPV = 0.052 AGLRV = 0.058 AGLTV = 0.053	SP2 R1 (7910)	A = 0.14 G = 0.10 L = 0.09 V = 0.08 S = 0.06 T = 0.06 R = 0.08 D = 0.06 P = 0.06	AGLRV = 0.055

(*Contd*)

Table 1. (*Contd*)

Species	Fuzzy association		Soft association			Soft fuzzy association		
	Frequent 1-amino acid with their support	Maximal association patterns with their fuzzy support	Species and range	Frequent 1-amino acid with their support	Maximal association patterns with their support	Species and range	Frequent 1-amino acid with their support	Maximal association patterns with their support
			SP2	A = 0.14	AGLST = 0.051	SP2	A = 0.14	AGLST = 0.050
				G = 0.18	ADGLV = 0.051		G = 0.18	ADGLV = 0.051
			R2	L = 0.09	AGLPV = 0.052	R2	L = 0.08	AGLPV = 0.052
				V = 0.07	AGLRV = 0.051		V = 0.07	AGLRV = 0.051
			(89)	S = 0.06	AGLSV = 0.052	(89)	S = 0.06	AGLSV = 0.051
				T = 0.06	AGLTV = 0.052		T = 0.06	AGLTV = 0.051
				D = 0.05			R = 0.05	
				P = 0.05			D = 0.05	
				R = 0.05			P = 0.05	
			SP2	A = 0.13	ADGLRV = 0.051	SP2	A = 0.12	ADGLRV = 0.052
				G = 0.13	ADGLSV = 0.051		G = 0.13	ADGLSV = 0.050
			R3	L = 0.09		R3	L = 0.08	**AGLSTV = 0.051**
				V = 0.08			V = 0.08	
			(12)	S = 0.07		(12)	S = 0.07	
				T = 0.06			T = 0.06	
				R = 0.06			R = 0.05	
				D = 0.06			D = 0.06	
				P = 0.05			P = 0.05	
				I = 0.05				
				N = 0.05				
M. Bovis BCG (5008)	A = 0.13 G = 0.09 L = 0.10 V = 0.09 S = 0.06 T = 0.06 R = 0.08 D = 0.06 P = 0.06 E = 0.05 I = 0.05	ADGLV = 0.051 AGLPV = 0.050 AGLRV = 0.058 AGLTV = 0.052	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	ADGLRV = 0.051	SP3 R1 (4946)	A = 0.13 G = 0.09 L = 0.10 V = 0.08 S = 0.06 T = 0.06 R = 0.08 P = 0.06 D = 0.05	**AGLPV = 0.050** AGLRV = 0.058 **AGLTV = 0.052**
			SP3	A = 0.13	ADGLV = 0.051	SP3	A = 0.13	ADGLV = 0.051
				G = 0.20	AGLTV = 0.051		G = 0.19	AGLTV = 0.051
			R2	L = 0.08		R2	L = 0.08	
				V = 0.07			V = 0.07	
			(88)	S = 0.05		(88)	T = 0.06	
				T = 0.06			D = 0.05	

(*Contd*)

Table 1. (*Contd*)

Species	Fuzzy association		Soft association			Soft fuzzy association		
	Frequent 1-amino acid with their support	Maximal association patterns with their fuzzy support	Species and range	Frequent 1-amino acid with their support	Maximal association patterns with their support	Species and range	Frequent 1-amino acid with their support	Maximal association patterns with their support
				D = 0.05			S = 0.05	
				R = 0.05			R = 0.05	
				P = 0.05			P = 0.05	
				E = 0.05				
			SP3	A = 0.13	ADGLRV = 0.054	SP3	A = 0.13	ADGLRV = 0.052
				G = 0.13	ADGLSV = 0.052		G = 0.13	ADGLSV = 0.050
			R3	L = 0.09		R3	L = 0.09	**AGLSTV = 0.051**
				V = 0.09			V = 0.09	
			(4)	S = 0.07		(4)	S = 0.07	
				T = 0.06			T = 0.06	
				R = 0.05			D = 0.06	
				D = 0.06			R = 0.05	
				P = 0.05			**N = 0.05**	
				E = 0.05				
				N = 0.05				
M. Canettii	A = 0.13	ADGLR = 0.050	SP4	A = 0.13	ADGLRV = 0.051	SP4	A = 0.13	**ADGLV = 0.051**
(13,374)	G = 0.09	ADGLV = 0.051		G = 0.10	AGLRTV = 0.050		G = 0.10	**AGLPV = 0.052**
	L = 0.10	AGLPV = 0.052	R1	L = 0.10		R1	L = 0.10	**ADLRV = 0.051**
	V = 0.09	ADLRV = 0.051		V = 0.08			V = 0.08	**AGLRV = 0.059**
	S = 0.06	AGLRV = 0.059	(13,210)	T = 0.06		(13,210)	T = 0.06	**AGLTV = 0.053**
	T = 0.06	AGLTV = 0.053		R = 0.072			R = 0.07	
	R = 0.08			D = 0.06			D = 0.06	
	D = 0.06			P = 0.06			P = 0.06	
	P = 0.06			E = 0.05				
				S = 0.05				
			SP4	A = 0.14	ADGLPV = 0.051	SP4	A = 0.14	ADGLPV = 0.051
				G = 0.14	ADGLRV-0.054		G = 0.14	ADGLRV = 0.054
				L = 0.09	ADGLSV = 0.050		L = 0.09	AGLPSV = 0.050
			R2	V = 0.08	AGLPSV = 0.051	R2	V = 0.08	ADGLTV = 0.050
				S = 0.06	ADGLTV = 0.050		S = 0.06	AGLPTV = 0.052
			(141)	T = 0.06	AGLPTV = 0.052	(141)	T = 0.06	
				R = 0.072	AGLSTV = 0.051		R = 0.07	
				D = 0.06			D = 0.06	
				P = 0.06			P = 0.06	
				E = 0.05				

(*Contd*)

Table 1. (*Contd*)

Species	Fuzzy association		Soft association			Soft fuzzy association		
	Frequent 1-amino acid with their support	Maximal association patterns with their fuzzy support	Species and range	Frequent 1-amino acid with their support	Maximal association patterns with their support	Species and range	Frequent 1-amino acid with their support	Maximal association patterns with their support
			SP4	A = 0.10	AGILNT = 0.050	SP4	A = 0.10	**AFGILST = 0.050**
				G = 0.17	AFGIST = 0.051		**F = 0.05**	**AGILNST = 0.050**
			R3	I = 0.06	AFGLST = 0.051	R3	G = 0.17	
				L = 0.09	AGILST = 0.057		**I = 0.06**	
			(23)	V = 0.07	AGINST = 0.052	(23)	L = 0.08	
				N = 0.09	AGLSTV = 0.054		V = 0.07	
				S = 0.07	AGLPSV = 0.050		**N = 0.09**	
				T = 0.07			S = 0.07	
				F = 0.05			T = 0.07	
				P = 0.05			P = 0.05	
M. Africanum (6176)	A = 0.13	ADGLR = 0.051	SP5	A = 0.13	ADGLRV = 0.051	SP5	A = 0.13	ADGLR = 0.051
	G = 0.92	ADGLV = 0.052		G = 0.10			G = 0.9	ADGLV = 0.052
	L = 0.10	AGLPV = 0.051	R1	L = 0.10		R1	L = 0. 10	AGLPV = 0.051
	V = 0.09	ADLRV = 0.051		V = 0.08			V = 0.09	ADLRV = 0.051
	S = 0.06	AGLRV = 0.060	(6109)	T = 0.06	AGLRTV = 0.050	(6109)	S = 0.05	AGLRV = 0.060
	T = 0.06	AGLTV = 0.053		R = 0.07			T = 0.06	AGLTV = 0.053
	R = 0.08			D = 0.06			R = 0.08	
	D = 0.06			P = 0.05			D = 0.06	
	P = 0.06			**E = 0.05**			P = 0.08	
			SP5	A = 0.14	AGLPTV = 0.050	SP5	A = 0.14	ADGLV = 0.054
				G = 0.16			G = 0.16	AGLPV = 0.053
			R2	L = 0.09		R2	L = 0.09	AGLRV = 0.055
				V = 0.08			V = 0.08	**AGLSV = 0.053**
			(59)	S = 0.06		(59)	S = 0.06	AGLTV = 0.053
				T = 0.07			T = 0.06	**AGPTV = 0.050**
				R = 0.06			R = 0.06	
				D = 0.06			D = 0.06	
				P = 0.05			P = 0.06	
			SP5	A = 0.11	AGILST = 0.052	SP5	A = 0.11	AGILST = 0.053
				F = 0.05	AGLPSV = 0.053		F = 0.05	AGLPSV = 0.051
			R3	G = 0.15	AGLSTV = 0.052	R3	G = 0.15	AGLSTV = 0.053
				I = 0.05			I = 0.06	
			(8)	L = 0.09		(8)	L = 0.09	
				V = 0.08			V = 0.08	
				N = 0.07			D = 0.05	
				S = 0.07			N = 0.07	

(*Contd*)

Table 1. (*Contd*)

Species	Fuzzy association		Soft association			Soft fuzzy association		
	Frequent 1-amino acid with their support	Maximal association patterns with their fuzzy support	Species and range	Frequent 1-amino acid with their support	Maximal association patterns with their support	Species and range	Frequent 1-amino acid with their support	Maximal association patterns with their support
				T = 0.07 P = 0.05			S = 0.07 T = 0.07 P = 0.05	
M. Microti (28)	A = 0.11 G = 0.09 L = 0.09 V = 0.08 S = 0.06 T = 0.07 R = 0.07 D = 0.06 E = 0.08 P = 0.05	AGLTV = 0.051	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 R1 (13)	A = 0.09 G = 0.08 L = 0.10 V = 0.08 S = 0.06 T = 0.06 R = 0.08 E = 0.10	AGLV = 0.055 EGLV = 0.054
			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 R2 (6)	A = 0.10 G = 0.08 L = 0.09 V = 0.10 S = 0.05 T = 0.08 R = 0.06 D = 0.07 E = 0.07 P = 0.05	ADEGLTV = 0.053
			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	SP6 R3 (9)	A = 0.15 **F = 0.05** G = 0.11 **I = 0.06** L = 0.08 V = 0.07 D = 0.05 **N = 0.09** S = 0.07 T = 0.06 R = 0.05 P = 0.06	AGLPST = 0.052 AGLSTV = 0.052

Table 2. Probable secondary structures and their corresponding 1, 2 and 3 amino acid association patterns using soft fuzzy approach

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

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

Species	Range	Maximal frequent patterns	No. of Maximal frequent patterns	Maximum association patterns	Physico-chemical properties	Probable structure
M. Tuberculosis	R1	4F	5	AGLR, AGLT, **AGLV**, AGRV, ALRV	Hydrophobic, polar uncharged, positively charged	Helix
	R2	7F	1	**AGLPSTV**	Hydrophobic, polar uncharged, positively charged	Coil
	R3	7F	1	**AGILNST**	Hydrophobic, positively charged	Coil
M. Bovis	R1	5F	1	AGLRV	Hydrophobic, polar uncharged positively charged	Helix
	R2	5F	6	AGLST, ADGLV, AGLPV, AGLRV, AGLSV, AGLTV	Hydrophobic, positively charged, polar uncharged	Helix
	R3	6F	3	**ADGLRV**, **ADGLSV**, **AGLSTV**	Hydrophobic	Coil
M. Bovis BCG	R1	5F	3	AGLPV, **AGLRV**, AGLTV	Hydrophobic, polar uncharged	Helix
	R2	5F	2	ADGLV, AGLTV	Hydrophobic, polar uncharged	Coil
	R3	6F	3	**ADGLRV**, **ADGLSV**, **AGLSTV**	Hydrophobic	Coil
M. Canettii	R1	5F	5	ADGLV, AGLPV, ADLRV, AGLRV, AGLTV	Hydrophobic, polar uncharged	Helix
	R2	6F	5	ADGLPV, **ADGLRV**, AGLPSV, ADGLTV, AGLPTV	Hydrophobic, positively charged, polar uncharged	Helix
	R3	7F	2	**AFGILST**, **AGILNST**	Hydrophobic	Coil
M. Africanum	R1	5F	6	ADGLR, ADGLV, AGLPV, ADLRV, **AGLRV**, AGLTV	Hydrophobic, polar uncharged	Helix
	R2	5F	6	ADGLV, AGLPV, **AGLRV**, AGLSV, AGLTV, AGPTV	Hydrophobic, polar uncharged	Helix
	R3	6F	3	AGILST, AGLPSV, AGLSTV	Hydrophobic	Coil
M. Microti	R1	4F	2	**AGLV**, **EGLV**	Hydrophobic, polar uncharged	Helix
	R2	7F	1	**ADEGLTV**	Hydrophobic, polar uncharged	Helix
	R3	6F	2	**AGLPST, AGLSTV**	Hydrophobic	Helix

soft set approach, and amino acids F, I and N are predicted as frequent 1-amino acids for (SP1, R3), as shown in column 5, Table 1. This indicates that the amino acid associations depend upon the parameters like range and species, and ignorance of these parameters is a cause of under and over prediction of patterns. Similarly by soft fuzzy approach for (SP1, R2), R is not predicted but D is predicted as frequent 1-amino acids. Also for (SP1, R2) D and R are not predicted as frequent 1-amino acids, but F, I, N and P are using soft fuzzy approach, as shown in column 7, Table 1. P is not predicted as frequent 1-amino acids for (SP1, R3) by soft set approach as shown in column 5, Table 1. This difference is due to the degree of relationships among amino acids, which is not taken into account in soft set approach leading to under prediction of amino acid P for (SP1, R3), whereas it is predicted by soft fuzzy approach as shown in column 7, Table 1 as the fuzzy set takes care of degree of relationship among amino acids. Similar interpretations can be made for the remaining species SP2–SP6.

Using the set approach for SP1 the maximal amino acid association pattern is 4-amino acids and their number is 6. Using the soft set approach maximal amino acid association patterns predicted for (SP1, R1) are 5-amino acids and their number is 3. For (SP1, R2), the maximal amino acid association pattern is 7 and the number of such patterns is 1. For (SP1, R3), the maximal amino acid association pattern is 6 and the number of such patterns is 4, as shown in column 6, Table 1. Again by soft fuzzy approach the maximal amino acid pattern for (SP1, R1) is 4 and the number of such patterns is 4; for (SP1, R2), the maximal amino acid patterns is 7 and their number is 1, and for (SP1, R3), the maximal amino acid patterns is 7 and number of such patterns is 1, as shown in column 9, Table 1. We observed that the size and number of maximal amino acid patterns change with parameters SP and R and with the approaches as well. We also observed that these maximal amino acid patterns differ in amino acids with respect to parameters SP and R and the approaches used. This implies that the amino acid association patterns depend on the parameters SP and R, and also the degree of relationship among the amino acids. Similar interpretation can be inferred for remaining species SP2 to SP6 in Table 1.

Since it has been observed that amino acid patterns depend on the parameters like species and length range, and

	S	SF	S	SF	S	SF	S	SF	S	SF	S	SF
	MTB		MB		MBCG		MC		MA		MM	
1F R1	9	9	9	9	10	9	10	8	9	9	9	8
1F R2	7	8	9	9	10	9	10	9	9	9	9	10
1F R3	9	10	11	9	11	10	10	10	10	11	7	12
2F R1	26	21	27	26	29	29	29	29	29	29	23	19
2F R2	28	29	27	27	22	24	33	33	29	29	34	32
2F R3	36	37	30	30	31	31	37	35	26	26	36	36
3F R1	34	20	36	27	44	40	46	43	46	43	18	10
3F R2	44	47	37	37	23	27	62	62	44	41	66	59
3F R3	63	68	48	49	36	36	69	63	40	40	33	34
4F R1	19	5	21	9	33	22	37	25	37	28	5	2
4F R2	39	43	24	24	11	13	63	62	33	26	72	62
4F R3	58	68	39	42	29	29	69	66	35	35	28	30
5F R1	3	0	4	1	11	4	14	6	14	6	0	0
5F R2	21	23	6	6	2	2	33	31	11	6	45	38
5F R3	27	37	15	18	12	12	35	41	16	16	12	13
6F R1	0	0	0	0	1	0	2	0	2	0	0	0
6F R2	7	7	0	0	0	0	7	6	1	0	15	13
6F R3	5	10	2	3	2	2	7	14	3	3	2	2
7F R1	0	0	0	0	0	0	0	0	0	0	0	0
7F R2	1	1	0	0	0	0	0	0	0	0	2	2
7F R3	0	1	0	0	0	0	0	2	0	0	0	0

Figure 2. Comparison between soft and soft fuzzy results.

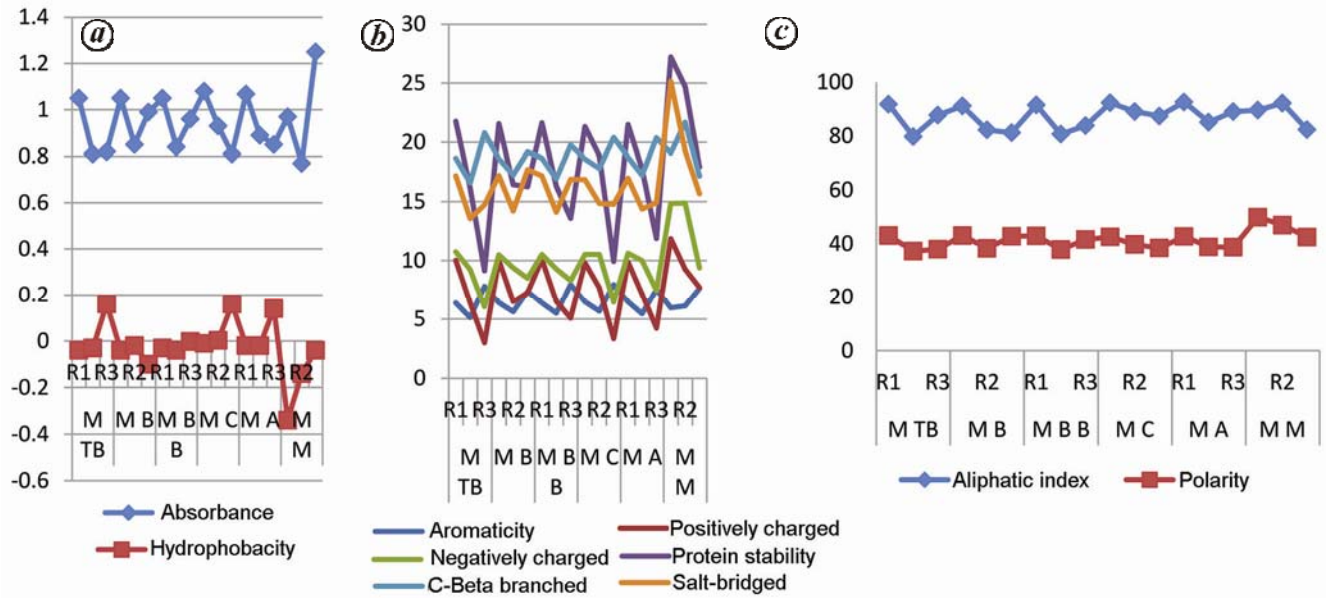

Figure 3 *a*–*c*. Some physico-chemical properties of MTBC.

Table 4. Physico-chemical properties of MTBC species for ranges R1, R2 and R3

Species and range		Absorbance	Hydrophobicity	Aliphatic index	Aromaticity	Protein stability	c-Beta branched	Polarity	Salt-bridged	Helix formation	Beta sheet	Coil	Positively charged	Negatively charged
M. Tuberculosis	R1	1.05	–0.04	91.81	6.36	21.76	18.66	43.12	17.19	45.23	26.04	28.67	10.02	10.73
	R2	0.81	–0.03	79.77	5.11	16.16	16.60	37.28	13.57	38.29	22.26	39.43	6.38	9.24
	R3	0.82	0.16	87.77	7.80	9.12	20.81	38.01	14.72	27.64	28.73	43.62	2.99	6.03
M. Bovis	R1	1.05	–0.04	91.17	6.39	21.57	18.61	43.08	17.21	45.08	26.05	28.84	10.02	10.50
	R2	0.85	–0.02	82.25	5.61	16.44	17.25	38.29	14.20	38.35	23.43	38.21	6.49	9.38
	R3	0.99	–0.10	81.21	7.36	16.26	19.20	42.81	17.70	34.61	27.09	38.28	7.25	8.48
M. Bovis BCG	R1	1.05	–0.03	91.54	6.36	21.65	18.66	42.97	17.17	45.21	26.08	28.68	10.09	10.51
	R2	0.84	–0.04	80.68	5.48	16.32	16.93	37.90	14.11	37.91	22.95	39.13	6.51	9.27
	R3	0.96	–0.002	83.85	7.94	13.59	19.77	41.61	16.89	32.00	27.94	40.05	5.07	8.29
M. Canettii	R1	1.08	–0.01	92.36	6.45	21.32	18.60	42.59	16.85	45.15	26.05	28.78	9.81	10.52
	R2	0.93	0.003	89.09	5.66	18.87	17.79	39.79	14.85	42.24	24.13	33.61	7.67	10.52
	R3	0.81	0.16	87.39	7.94	9.93	20.41	38.46	14.79	28.66	28.50	42.83	3.34	6.44
M. Africanum	R1	1.068	–0.02	92.64	6.45	21.49	18.74	42.75	16.99	45.19	26.18	28.62	9.88	10.63
	R2	0.89	–0.02	85.13	5.45	17.75	17.22	38.84	14.38	40.68	23.33	35.98	7.13	9.99
	R3	0.85	0.14	89.08	7.51	11.88	20.38	38.77	14.91	31.35	28.09	40.55	4.23	7.44
M. Microti	R1	0.97	–0.34	89.62	5.93	27.21	19.09	49.92	25.17	47.06	25.43	27.39	11.86	14.85
	R2	0.77	–0.14	92.22	6.11	24.71	21.68	46.99	19.30	42.87	28.41	28.70	9.25	14.87
	R3	1.25	–0.04	82.35	7.62	17.93	17.16	42.51	15.66	42.24	25.64	32.10	7.70	9.37

degree of relationships among amino acids, the structure and physico-chemical properties will also depend on the parameters like species and length range, and degree of relationships among amino acids. To analyse this fact, the structure and physico-chemical properties based on maximal frequent amino acid patterns were computed (Tables 2 and 3).

It is observed that the fuzzy set approach suffers from the problem of under and over prediction due to ignorance of the parameters SP and R. At the same time the soft set approach suffers from the same problem due to ignorance of degree of relationships among the amino acids. The soft fuzzy approach incorporates the dependence of amino acid patterns on parameters SP and R and the degree of relationships among the amino acids to overcome the uncertainty which arises due to non-consideration of parameters and ignorance of degree of relationships among amino acids. Thus the soft fuzzy approach is superior to the individual soft and fuzzy approaches, as it reaps the advantages of both.

Figure 2 clearly shows the differences in the frequent *i*-amino acid patterns (i = 1(1)7) for different ranges and species by soft and soft fuzzy approaches. The graph clearly depicts the difference in the size and number of amino acid patterns due to different ranges, species and approaches.

Table 2 presents the probable secondary structure of proteins present in all the MTBC species for soft fuzzy approach. In R1, R2 and R3 ranges, the frequent-1, frequent-2 and frequent-3 patterns of amino acid associations supporting formation of helix, coil and sheet are predicted. The results displayed in Table 2 indicate that all the species have a tendency to form the helix structure frequently for R1, R2 and R3. For 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 3 describes the physico-chemical properties of maximal frequent amino acid patterns in R1, R2 and R3 for all species of MTBC by soft fuzzy approach. In all species the amino acids present in maximal frequent pattern are hydrophobic, non-polar, acidic polar and polar. The polar group amino acids are frequent in all species, while basic polar group amino acids are not so frequent in maximal frequent pattern. Table 3 also shows the total frequent patterns for amino acid associations for MTBC species for all three ranges (R1–R3) using soft fuzzy approach. The probable secondary structures in different ranges are also shown.

The amino acid associations have been used to predict the physico-chemical properties (Figure 3 and Table 4) for all three ranges. Figure 4 shows the secondary structures predicted on the basis of amino acid association patterns in MTBC.

In Tables 2 and 3, we observed the differences in secondary structures and physico-chemical properties with respect to length range and species. This is due to the differences in amino acid associations (Table 1) due to length range and species and relationships among amino acids.

The association rules generated on the basis of the above study, specially for R2 and R3 are given below:

I. {A(frequent)/\L(frequent) ⇒ F(infrequent)}→ Tendency for helix formation.
II. {L(frequent)/\R(frequent) ⇒ C(infrequent)}→ tendency for helix formation.
III. {E(frequent)/\L(frequent) ⇒ W(infrequent)}→ tendency for helix formation.
IV. {S(frequent)/\T(frequent) ⇒ M(infrequent)}→ energetically beneficial for protein.
V. {P(frequent)/\S(frequent) ⇒ Q(infrequent)}→ tendency for coil formation.
VI. {A(frequent)/\R(frequent) ⇒ W(infrequent)}→ tendency for helix formation.
VII. {D(frequent)/\G(frequent) ⇒ M(infrequent)}→ tendency for coil formation.
VIII. {D(frequent)/\S(frequent) ⇒ K(infrequent)}→ tendency for coil formation.
IX. {W(infrequent)/\F(infrequent) ⇒ A(frequent)}→ tendency for helix formation.
X. {A(frequent)/\L(frequent) ⇒ E(infrequent)}→ helps in protein folding.
XI. {S(frequent)/\P(frequent) ⇒ H(infrequent)}→ tendency for coil formation.
XII. {A(frequent)/\V(frequent)/\L(frequent) ⇒ K(infrequent)}→aliphatic in nature.
XIII. {G(frequent)/\P(frequent) ⇒ H(infrequent)}→ tendency for coil formation.

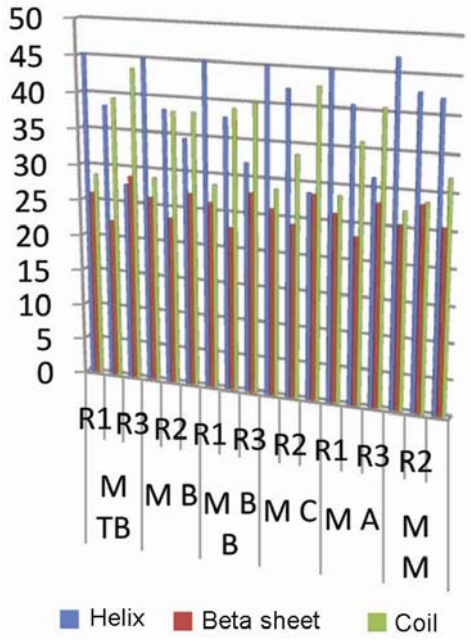

Figure 4. Secondary structure formation.

XIV. {V(frequent)/\L(frequent)/\A(frequent) ⇒ H(infrequent)}→hydrophobic.
XV. {V(frequent)/\T(frequent) ⇒ N(infrequent)}→ tendency for sheet formation.
XVI. {D(frequent)/\E(frequent) ⇒ K(frequent)}→ maintains stability of proteins (This rule applies for *M. Microti* species).
XVII. {A(frequent)/\E(frequent) ⇒ Y(infrequent)}→ tendency for helix formation (This rule applies for *M. Microti* and *M. Bovis* species).
XVIII. {P(frequent)/\G(frequent) ⇒ H(infrequent)}→ tendency for coil formation.
XIX. {W(infrequent)/\G(infrequent) ⇒ L(frequent)}→ tendency for helix formation.
XX. {A(frequent)/\L(frequent) ⇒ Q(infrequent)}→ Helps in protein folding.
XXI. {L(frequent)/\R(frequent) ⇒ Y(infrequent)}→ tendency for helix formation.
XXII. {V(frequent)/\T(frequent) ⇒ M(infrequent)}→ tendency for sheet formation.

According to rule (I) it has been observed that amino acids A and L that favour helix formation are frequent and amino acid F that favours sheet formation is infrequent. Thus this rule implies the tendency to form helix structure. Similarly, from rule (II) it has been observed that amino acids L and R that favour helix formation are frequent and amino acid C that favours sheet formation is infrequent. Thus this rule implies the tendency to form helix structure. Rules (V), (VII) and (VIII) imply the tendency for coil formation as amino acids S, D and P, G are frequent which favours the secondary structure coil, while amino acids Q, M and K which favour helix formation are infrequent. As per rule (IV), as polar amino acids S and T are frequent and non-polar amino acid M is infrequent, it is beneficial for proteins in terms of attaining stability. Similar interpretation can be inferred from the remaining rules.

Conclusion

The soft set and soft fuzzy set approaches are proposed and employed to mine amino acid association patterns in the peptide sequences of MTBC species. The results of soft fuzzy approach have been compared with those obtained by fuzzy set and soft set approaches separately. The difference in the results is due to the inherent uncertainties in the data caused by degree of relationships among amino acids in the peptide sequences of MTBC species and dependence of amino acid association patterns on parameters like length range and species. The soft set approach suffers from the problem of uncertainty due to non-consideration of degree of relationships among amino acids. The fuzzy set approach suffers from the problem of uncertainties due to ignorance of parameters like length range and species. The soft fuzzy approach takes care of both these uncertainties due to degree of relationship and dependence of patterns on parameters, and thus is superior to the individual soft set and fuzzy set approaches. Interesting association rules among amino acid of MTBC species have been generated by the soft fuzzy approach. Also, the physico-chemical properties and secondary structures have been predicted based on amino acid association patterns in peptide sequences of MTBC species. The crisp and fuzzy approaches basically give average amino acid associations and therefore average structures and physico-chemical properties. The granularity provided by soft fuzzy approach gives more specific results of amino acid associations, structures and physico-chemical properties, thus giving the better picture of these patterns than the other approaches. The association patterns and rules generated can serve as signatures for gaining better insights regarding the molecular mechanism of disease, protein structures and the protein–protein interactions.

1. Agrawal, R., Imielinski, T. and Swami, A. N., Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 1993, **22**(2), 207–216.
2. Agrawal, R. and Srikant, R., Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Databases, *VLDB*, 12 September 1994, vol. 1215, pp. 407–419.
3. Patel, R., Swami, D. K. and Pardasani, K. R., Lattice based algorithm for incremental mining of association rules. *Int. J. Theor. Appl. Comput. Sci.*, 2006, **1**(1), 119–128.
4. Pandey, A. and Pardasani, K., Rough set model for discovering multidimensional association rules. *IJCSNS Int. J. Comput. Sci. Network Security*, 2009, **9**(6), 159–164.
5. Panday, A. and Pardasani, K. R., PPCI algorithm for mining temporal association rules in large database. *J. Inf. Knowledge Manage.*, 2009, **8**(04), 345–352.
6. Khare, N., Adlakha, N. and Pardasani, K. R., Karnaugh map model for mining association rules in large databases, *IJCNS Int. J. Comput. Network Security*, 2009, **1**(1), 16–21.
7. Kocatas, A., Gursoy, A. and Atalay, R., Application of data mining techniques to protein–protein interaction prediction. In *Computer and Information Sciences-ISCIS*, Springer, Berlin, Heidelberg 2003, pp. 316–323.
8. Rodríguez, A., Carazo, J. M. and Trelles, O., Mining association rules from biological databases. *J. Am. Soc. Inf. Sci. Technol.*, 2005, **56**(5), 493–504.
9. Oyama, T., Kitano, K., Satou, K. and Ito, T., Extraction of knowledge on protein–protein interaction by association rule discovery. *Bioinformatics*, 2002, **18**(5), 705–714.
10. Kuo, H. C., Ong, P. L., Lin, J. C. and Huang, J. P., Discovering amino acid patterns on binding sites in protein complexes. *Bioinformation*, 2011, **6**(1), p. 10.
11. Intan, R., An algorithm for generating single dimensional fuzzy association rule mining. *J. Informat*., 2006, **7**(1), p. 61.
12. Khare, N., Adlakha, N. and Pardasani, K. R., An algorithm for mining multidimensional fuzzy association rules. *Int. J. Comput. Sci. Inform. Security*, 2009, **5**(1), 72–76.
13. Khare, N., Adlakha, N. and Pardasani, K. R., A fuzzy based model for mining conditional hybrid dimensional association rules. *Int. J. Data Min. Knowledge Eng*., 2010, **2**(5), 69–76.

14. Gautam, P. and Pardasani, K. R., A novel approach for discovery of multilevel fuzzy association rules. *J. Comput.*, 2010, **2**(3), 56–64.
15. Gupta, N., Mangal, N., Tiwari, K. and Mitra, P., Mining quantitative association rules in protein sequences. In *Data Mining*, Springer, Berlin, Hidelberg, 2006, vol. 3755, pp. 273–281.
16. Francisco, J. L., Armando, B., Fernando, G., Carlos, C. and Antonio, M., FUZZY association rules for biological data analysis: a case study on yeast. *BMC Bioinforma.*, 2008, **9**(1), 107.
17. Kumari, T. and Pardasani, K. R., Mining fuzzy associations among amino acids of class A GPCRs. *Online J. Bioinformat.*, 2012, **13**(2), 202–213.
18. Kumari, T. and Pardasani, K. R., Mining amino acid association patterns in class B GPCRs. *Int. J. Bioinformat. Res. Appl.*, 2015, **11**(3), 219–232.
19. Shankar, A. and Pardasani, K. R., Mining fuzzy amino acid association patterns in various orders of class Alphaproteobacteria. *J. Med. Imag. Health Informat.*, 2013, **3**(3), 380–387.
20. Molodtsov, D., Soft set theory – first results. *Comput. Math. Appl.*, 1999, **37**(4), 19–31.
21. Herawan, T. and Mustafa, M. D., A soft set approach for association rules mining. *Knowledge-Based Syst.*, 2011, **24**(1), 186–195.
22. World Health Organization, report 2013 – Global tuberculosis report.
23. Cole, T. S., Comparative and functional genomics of the *Mycobacterium tuberculosis* complex. *Microbiology*, 2002, **148**(10), 2919–2928.
24. Saravanan, M. K. and Selvaraj, S., Search for identical octapeptides in unrelated proteins: Structural plasticity revisited. *Peptide Sci.*, 2012, **98**(1), 11–26.
25. Uthayakumar, M., Patra, S., Nagarajan, R. and Sekar, K., Sequence–structure similarity: do sequentially identical peptide fragments have similar three-dimensional structures? *Curr. Bioinformat.*, 2012, **7**(2), 111–115.
26. Brosch, R. *et al.*, A new evolutionary scenario for the *Mycobacterium tuberculosis* complex. *Proc. Natl. Acad. Sci. USA*, 2002, **99(**3), 684–3689.
27. Shabbeer, A., Cowan, L. S., Ozcaglar, C., Rastogi, N., Vandenberg, S. L., Yener, B. and Bennett, K. P., TB-lineage: an online tool for classification and analysis of strains of *Mycobacterium tuberculosis* complex. *Infect. Genet. Evol.*, 2012, **12**(4), 789–797.
28. http://www.ncbi.nlm.nih.gov/

ACKNOWLEDGEMENTS. We thank the Department of Biotechnology, New Delhi and MPCST Bhopal for providing bioinformatics infrastructure facility at MANIT, Bhopal to carry out this work.

Received 30 January 2015; revised accepted 22 September 2015

doi: 10.18520/cs/v110/i4/603-618