Proc. of The Seventh International Conference on Advances in Computing, Electronics and Communication - ACEC 2018
Copyright © Institute of Research Engineers and Doctors, USA. All rights reserved.
ISBN: 978-1-63248-157-3 doi: 10.15224/978-1-63248-157-3-12

Pattern Identification on Protein Sequences of Neurodegenerative Diseases Using Association Rule Mining

M Shahedul Islam[1], Swapnil Saha[2], Mohammad Shamsur Rahman[3], Md. Abul Kashem Mia[4]

Abstract—**Proteins are the integral part of all living beings, which are building blocks of many amino acids. To be functionally active, amino acids chain folds up in a complex way to give each protein an unique 3D shape, where a minor error may cause misfolded structure. Neurodegenerative diseases e.g. *Alzheimer, Parkinson, Sickle cell anemia*, etc. arise due to misfolding in protein sequences. Thus, identifying the patterns of the amino acids is important for inferring the protein associated genetic diseases. Recent studies in predicting patterns of amino acids focused on only the simple chronic neurodegenerative disease *Chromaffin Tumor* by applying association rule mining. However, more complex diseases are yet to be attempted. Moreover, the association rules obtained by these studies were not verified by usefulness measuring tools. In this paper, we have analyzed the protein sequences associated with more complex neurodegenerative protein misfolded diseases by association rule mining technique, where only the useful rules are finally sorted out with the use of interestingness measures. Adopting the quantitative experimental method, our work forms more reliable and strong association rules among the most domination amino acids of corresponding misfolded proteins and identify the dominating patterns of amino acid of complex genetic disease.**

Keywords — ***amino acid, association rule, disease, frequent pattern, protein misfolding, protein sequence.***

I. Introduction

Frequent Patterns (FP) are small patterns that repeatedly occurs in a database, specially high in bio-sequences. The challenging task in pattern finding of bio-sequences is to find FP [1]. Data Mining has recently increased its popularity in classifying the biological sequences and structures based on their critical features and functions [2].

Protein is one among the important factors and acts as constituents of all living organisms [2]. Protein are building blocks of hundreds of Amino acids joined together by peptide bonds. To be functionally active, amino acids chain folds up in complex way to give each protein a unique 3D shape. Protein folding is crucial for living organism as it affects gene skeleton. A small error in the folding process results in a misfolded structure, which can sometimes be lethal [3]. Protein misfolding is one of the primary cause of genetic disorder diseases such as Alzheimer's disease, Parkinson's disease, Huntington's disease, Sickle cell anemia, Cystic fibrosis, Cancer and many other degenerative and neurodegenerative disorders [4]. Protein misfolding may occurs due to an unwanted mutation in their amino acids or because of an error in the folding process. Thus, the relationship between these amino acids is very vital in case of protein misfolded diseases.

Frequent pattern mining is helpful to find the recurring relationships, association and correlation in a given data set [1]. Patterns can be represented as association rules and association rules are said to be strong if it satisfies both a minimum support threshold and a minimum confidence threshold. Therefore, frequent pattern mining can provide the solution for association rules formation among the most dominating amino acids for different protein misfolded diseases. To the best of our knowledge, three studies [2, 5, 6] have been identified on this issue. But all these were focused to predict pattern and association rules of the most dominating amino acids which causes the *Chromaffin Tumor* disease only. However, predicting the pattern and associations between more complex diseases are yet to be attempted in literature. Moreover, association rules obtained by these studies were not verified by usefulness measures.

The aim of this paper was to analyze protein sequences associated with complex protein misfolded diseases (i.e *Sickle Cell Anemia*, *Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa-4*) and identify frequent patterns among their amino acids. Here, association rule mining is used to predict patterns. Association rules were considered to be strong if it satisfied a minimum support and a confidence threshold. Then only useful rules were finally sorted out with the use of interestingness measures (i.e *Lift* and *Improve*). Adopting quantitative experimental method, this work forms more reliable and strong association rules among the most domination amino acids of corresponding proteins and identify the dominating patterns of amino acid of complex protein misfolded diseases. Identification of the most dominating amino acids and their patterns may open up new opportunities in medical science to handle the concerned genetic disorder diseases.

[1] M Shahedul Islam
Bangabandhu Sheikh Mujibur Rahman Maritime University
Bangladesh

[2] Swapnil Saha
East West University
Bangladesh

[3] Mohammad Shamsur Rahman
Monash University
Australia

[4] Md. Abul Kashem Mia
Bangladesh University of Engineering and Technology
Bangladesh

ISBN: 978-1-63248-157-3 doi: 10.15224/978-1-63248-157-3-12

This paper is organized as follows. Section 2 presents theoretical background of related issues. Section 3 highlights an overview of the related works. The experimental design is presented in section 4 and section 5 represents the data analysis and results. The concluding remarks and the future work are presented in the final section.

II. Theoretical Background

Some of the concepts and issues such as protein structure, protein associated diseases, association rule mining and their interestingness measures which have been considered in this paper are discussed below.

A. *Amino Acid and Protein*

Amino acids are made from carbon, hydrogen, nitrogen, and oxygen. Amino acid is characterized as unique due to its R-group. The 20 amino acids that are found within proteins convey a vast array of chemical versatility [7]. To survive, living being need proteins. Proteins are complex molecules, made up of hundreds of amino acids that are attached to one another by peptide bonds (Fig. 1), forming a long chain [8].

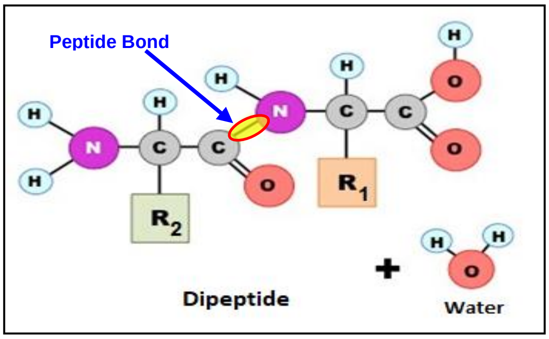

Figure 1. Amino Acids joined together through *Peptide Bonds*

B. *Protein Misfolding*

To be functionally active, amino acids chain folds up in a complex way to give each protein a unique 3D shape (Fig. 2). Protein may lose its usual function or may gain deleterious function if not folded properly. Proteins that are not able to achieve native state, due either to an unwanted mutation in their amino acid sequence or simply because of an error in folding process, are recognized as misfolded.

C. *Protein Misfolding Diseases*

According to the prion researcher Susan Lindquist, ‘protein misfolding could be involved in up to half of all human diseases’ [9]. Protein misfolding is believed to be the primary cause of genetic disorder diseases such as *Alzheimer*, *Parkinson*, *Huntington*, *Sickle cell anemia*, *Cystic fibrosis*, *Cancer* and many other degenerative and neurodegenerative disorders [4]. Over last two decades, protein misfolding and its pathogenic effect have become a significant area of human bio-molecular research. In this work, three protein misfolded diseases (i.e. *Sickle Cell Anemia* [11], *Nephrogenic Diabetes Insipidus* [12] and *Retinitis Pigmentosa 4* [12]) have been experimented.

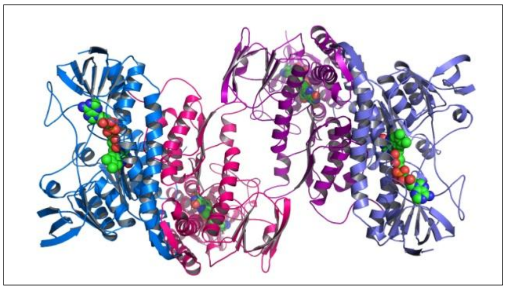

Figure 2. Quaternary/Final protein structure

D. *Frequent Pattern Mining in Bioinformatics*

Frequent patterns are either itemsets or subsequences or substructures which appear in a data set with a frequency that is equal to or higher than a threshold specified by the user. Pattern mining is useful in bioinformatics for predicting rules of certain elements in genes, for protein function prediction, for gene expression analysis, for protein fold recognition and for motif discovery in DNA sequences [13]. Thus frequent pattern mining can be used to find recurring relationships, association and correlation between amino acids for protein misfolded diseases.

E. *Association Rule Mining*

Association rule mining is one sorts of pattern mining which is built from frequent itemset mining. In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases [14]. Patterns can be represented as association rules and the association rules are said to be strong if it satisfies both a minimum support threshold and a minimum confidence threshold. Therefore, frequent pattern mining can provide solution for association rules formation among the most dominating amino acids for different protein misfolded diseases. To analyse, predict and manage bulk biological data, numerous computer algorithms and methods are developed which help to compare and align biological sequences and predict bio-sequence patterns [1]. In this work, as a tools of association rule mining, Apriori algorithm was used to analyse, predict and identify desired pattern of dominating amino acids in the protein sequences.

In this work, strong and interesting association rules were generated and selected by using constraints on various measures of interest and significance (i.e. *support*, *confidence*, *lift* and *improve* [15]).

*1) **Support***: The *support* of an itemset *X*, *supp (X)* is defined as proportion of transaction in data set in which the item *X* appears. It indicates popularity of an itemset.

$$supp\ (X) = \frac{No.of\ transactions\ in\ which\ itemset\ X\ appeared}{Total\ no.of\ transactions} \tag{1}$$

*2) **Confidence***: The *confidence* of a rule is defined as:

$$conf\ (X \rightarrow Y) = \frac{supp\ (X \cup Y)}{supp(X)} \tag{2}$$

*3) **Lift**:* The *lift* of a rule is defined as:

$$lift(X \rightarrow Y) = \frac{supp(X \cup Y)}{supp(Y) * supp(X)} \qquad (3)$$

The rule (*X*→*Y*) will be considered as positively correlated rule if its *Lift* value is greater than 1. Thus, those rules are useful only whose *Lift* value is greater than 1.

*4) **Improve**: Improve* is a relatively new interestingness measure method of association rules defined as:

$$Improve\ (X \rightarrow Y) = [\ P(Y\mid X) - P(Y)\ ] \qquad (4)$$

III. Literature Review

Frequent Contiguous Patterns (FCP) are small patterns that repeatedly occurs in a database, specially high in bio-sequences. Biological sequences such as DNA and protein sequences consist of long linear chain of chemical components and typically contain a large number of items [16]. Frequent pattern mining is helpful to find the recurring relationships, association and correlation in a given data set [1]. In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases [14]. Data Mining has increased popularity in classifying biological sequences and structures based on their critical features and functions [2].

Protein is one among the important factors and acts as the constituents of all living organisms [2]. Proteins are made up of smaller building blocks called amino acids, joined together in chains [17]. The chains of amino acids fold up in complex ways, giving each protein a unique 3D shape. Thus, relationship between these amino acids is very vital in case of protein misfolded diseases. To the best of our knowledge, three studies [2, 5, 6] have been identified on this issue.

G. Lakshmi and S. Hariharan [5] aimed to predict patterns applying strong association rules over the frequent itemsets of the protein sequence named *Succinate dehydrogenase* which is involved in *chromaffin tumor* disease. The system generated frequent itemsets from the protein sequence and construct a frequent pattern tree. Thereafter strong association rules were generated based on 90% confidence threshold to identified the dominating amino acids.

G. Lakshmi and S. Hariharan [2] conducted another similar research in finding the most dominating amino acids (in *Succinate dehydrogenase* protein) which causes the disease *chromaffin tumor*. Here, Apriori algorithm was used in finding frequent items using candidate generation and then generating association rules from those frequent itemsets. In predicting the pattern, this work considered 5 as minimum *Support* count and 90% *Confidence* threshold.

S. Dhumale carried out similar work [6] to find dominating amino acids responsible to cause five diseases, i.e.*Epilepsy*, *Hartnup*, *Cystinuria*, *Alzheimer* and *Chromaffin Tumor*. As deduction, the author claimed five amino acid patterns (association rules), each to be responsible for an individual diseases. This work suffers serious limitations. Firstly, the experimented protein sequence is anonymous. Secondly, the author arbitrarily increased the minimum *Support* count from 2 to 5 and declared set of amino acid pattern as responsible for each of the diseases. But basis of such deduction was not authenticated.

The above three works were focused to predict the pattern and association rules of amino acids which causes the *Chromaffin Tumor* disease only. However, finding patterns of other protein associated diseases ate yet to be attempted.

IV. Methodology

In this study, three protein misfolded diseases were taken in consideration. The protein sequences associated with each of the diseases were collected from a well-recognised protein data bank. To generate the strong association rules from the amino acids of the protein associated diseases, minimum support count 3 and 4 were considered and minimum confidence as 90%. Based on the strong association rules, this proposed system was focused on predicting the most dominating amino acids than the other amino acids that cause the disease from the protein data sets. The general work flow of the proposed system is as follows:

a. **<u>Protein Sequence Selection</u>**. As stated earlier, in this work, three misfolded diseases (i.e. *Sickle Cell Anemia*, *Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa 4*) were taken in consideration. Protein sequences (amino acid chain) associated with these diseases were collected from protein data bank named Universal Protein Resource (www.uniprot.org/). Table I shows the experimented human diseases, their associated proteins and their lengths.

TABLE I. DIFFERENT HUMAN DISEASES AND INVOLVED PROTEINS

Disease	**Protein Name**	**Lengths**
Sickle Cell Anemia	Hemoglobin Subunit Beta Entry Code: P68871	147
Nephrogenic Diabetes Insipidus (NDI)	Vasopressin V2 Receptor (V2R) Entry Code: P30518	371
Retinitis Pigmentosa 4 (RP4)	Rhodopsin (Opsin-2) Entry Code: P08100	348

Source: http://www.uniprot.org/

b. **<u>Data set Splitting</u>**. Each of the protein sequences were split into amino acid sub sequences of length 10. For example, Hemoglobin Subunit Beta protein sequence (associated with *Sickle Cell Anemia* disease) contained amino acids of 147 length which was split into 15 sub sequences of length 10 each as shown in Table II.

TABLE II. SUB SEQUENCES OF HEMOGLOBIN SUBUNIT BETA PROTEIN SEQUENCE

10 MVHLTPEEKS	20 AVTALWGKVN	30 VDEVGGEALG
40 RLLVVYPWTQ	50 RFFESFGDLS	60 TPDAVMGNPK
70 VKAHGKKVLG	80 AFSDGLAHLD	90 NLKGTFATLS
100 ELHCDKLHVD	110 PENFRLLGNV	120 LVCVLAHHFG

130	140	147
KEFTPPVQAA	YQKVVAGVAN	ALAHKYH

Source: http://www.uniprot.org/uniprot/P68871

c. **Generating Association Rules**: The sub sequences was then used for associative pattern identification through Apriori Algorithm data mining technique. Association rules were generated based on minimum support count threshold and minimum 90% confidence level. In this work, the lengths of the protein sequences were not uniform and thus to generate and analyse a significant number of association rules, the minimum support count was subjectively selected 3 for Hemoglobin Subunit Beta protein and 4 for Vasopressin V2 Receptor and Rhodopsin proteins.

d. **Measuring Interestingness of Association Rules**. All the association rules generated in the previous step may not be useful. Therefore, the interestingness of these rules were measured and evaluated. Here, objective measuring tools, *Lift* and *Improve* were used to finally sort out the useful association rules.

e. **Pattern Identification**. Based on the strong and useful association rules, this proposed system focused on predicting the most dominating amino acids, and thus the associative patterns among the amino acids were identified for each protein misfolded disease.

In this work, the association rules were generated and measured by using following sequences:

1. firstly, *Support* and *Confidence* threshold was used to filter out frequent itemsets and strong association rules
2. secondly, *Lift* and *Improve* value were calculated
3. then, according to the *Lift* and *Improve* value, useful association rules were sorted out

A) Algorithm

In this work, te algorithm used takes three inputs: (i) the protein sequence of a particular protein misfolded disease, (ii) minimum support count and (iii) minimum threshold confidence. Then the algorithm returns the strong and useful association rules of the most dominating amino acids for the concerned protein misfolded disease (Pseudocode Fig. 3).

```
Input: Protein sequence, Support Count, Confidence, Usefulness measuring parameter (Lift and Improve)
Output: Useful Strong Association Rules
Procedure:
generate_association_rules():
1:  TRANSACATIONS ← each consecutive 10 amino acids of Protein sequence
2:  for each elements∈TRANSACATIONS do
3:      Count frequency of each amino acid
4:  ITEMSETS ← all possible itmsets
5:  Eliminate itemsets having support count below threshold value
6:  for each elements ∈ITEMSETS do
7:      ASSOCRULES ← all association rules
8:  ASSOCRULES ← strong association rules
9:  ASSOCRULES ← useful strong association rules
```

Figure 3. Pseudocode for generating useful strong association rules

Details: Here we have taken a Protein sequence, support count, confidence, lift and improve as input. In line 1, we have generated subsequences called *TRANSACATIONS* by taking 10 consecutive amino acids from the given Protein sequence. Line 2 and 3 counts the frequency of each amino acid for each element of *TRANSACATIONS*. In line number 4, a recursive function is called to generate all possible itemsets of various length (e.g. A, B, H, AB, AH, BH, ABH, etc) and stored into *ITEMSETS*. Line 5, eliminates all itemsets from *ITEMSETS* whose support count is less than the given support count. In line 6 and 7, all possible association rules are generated for each itemset. Line 8, eliminates association rules having confidence below the given threshold and finally useful association rules are sorted out by using *Lift* and *Improve* measures in line 9.

V. Experimental Results

The algorithm of the experiment had been implemented using C++ in a laptop computer with an Intel Core i5-7200U CPU (clock frequency 2.7 GHz and 4 GB RAM). Experimental results were obtained from each of the protein sequences. During the computation, the number of iterations were not fixed. The algorithm was continued till no further successful extensions were found. The work thus followed three basic actions:

a. Frequent itemsets generation
b. Generation of strong association rules
c. Determining interestingness/usefulness of association rules

In doing so, following consideration were made:

a. Support count threshold between 3 to 4 for frequent itemset generation.
b. Minimum 90% confidence level to obtain strong association rules.
c. Using *Lift* and *Improve* as measuring instrument to find useful strong association rules.

A. *Frequent Itemsets Generation*

Frequent itemsets generation means the frequent amino acid sets generation from the transactional protein datasets (sub sequences). For every protein sequences, frequent itemsets were generated. The algorithm maintains list of frequent amino acid sets to further generate strong association rules.

1) ***Disease-1: Sickle Cell Anemia****:* For *Sickle Cell Anemia*, protein sequence *Hemoglobin Subunit Beta* were loaded as input file. Here, 3 was considered as minimum support count. The process continued up to 5th iteration and garnered total 135 itemsets (comprising 1-itemsets to 5-itemsets) of amino acids. A few of the generated frequent itemsets for *Sickle Cell Anemia* is graphically represented in Fig. 4.

ISBN: 978-1-63248-157-3 doi: 10.15224/978-1-63248-157-3-12

2) ***Disease-2: Nephrogenic Diabetes Insipidus***: For *Nephrogenic Diabetes Insipidus (NDI)*disease, protein sequence *Vasopressin V2 Receptor* was loaded as the input file. Here, due to moderate length (371), minimum support count 4 was considered. The process continued up to 5th iteration and generated total 234 itemsets. A few of the generated frequent itemsets for *Nephrogenic Diabetes Insipidus* is shown in Fig. 5.

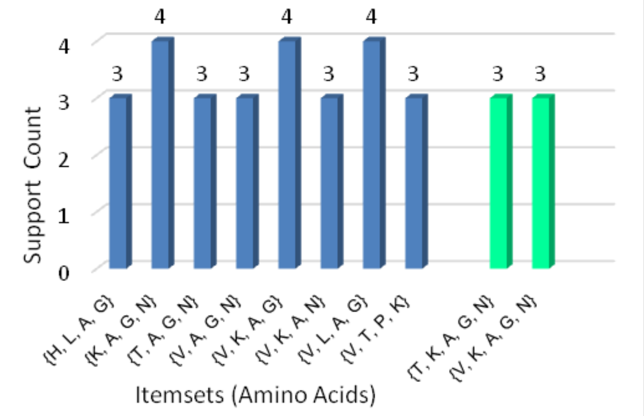

Figure 4. A few frequent 4-itemsets and 5-itemsets obtained from protein sequence for *Sickle Cell Anemia*

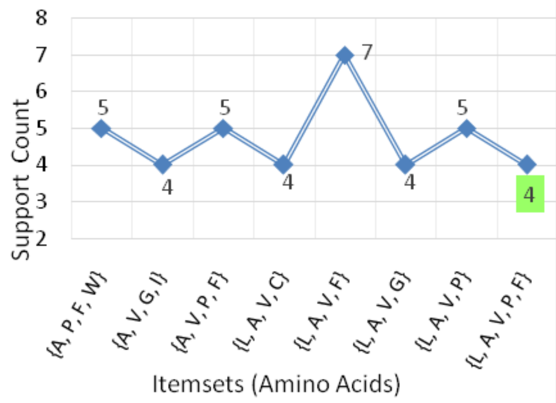

Figure 5. A few frequent 4-itemsets and 5-itemsets obtained from protein sequence for *Nephrogenic Diabetes Insipidus*

3) ***Disease-3: Retinitis Pigmentosa 4***: Protein sequence *Rhodopsin (Opsin-2)* was loaded in the process as input for *Retinitis Pigmentosa 4 (RP4)* disease. Here, 4 was considered as the minimum support count. The process continued up to 5th iteration and generated total 268 itemsets. Few generated frequent itemsets for *Retinitis Pigmentosa 4* is graphically represented in Fig. 6.

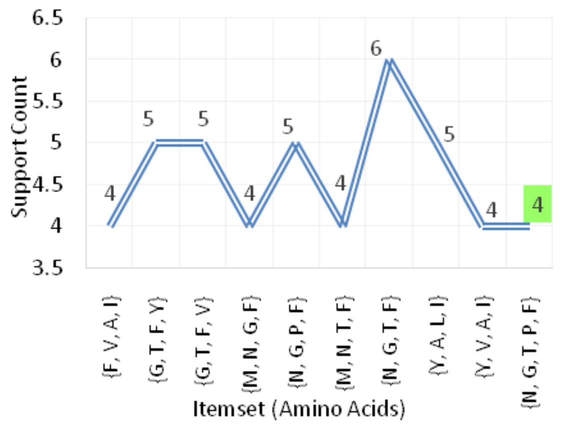

Figure 6. A few frequent 4-itemsets and 5-itemsets obtained from protein sequence for *Retinitis Pigmentosa 4*

B. ***Strong Association Rules Generation***

The algorithm generates strong association rules from the list of frequent itemsets (amino acid sets) for each protein sequence considering 90% confidence threshold.

1) ***Disease-1: Sickle Cell Anemia***: The process generated 698 association rules from 135 frequent itemsets. Among these rules, only 95 rules satisfied the minimum confidence level (90%) and were considered as accepted strong association rules and rest 603 rules were rejected. Examples of few association rules in this phase are shown in Table III.

Table III. Generation of Association Rules for *Sickle Cell Anemia*

Ser	Assoc Rule	Conf.	Result	Ser	Assoc Rule	Conf.	Result
1	A→D	20%	Rejected	492	G→AKT	23%	Rejected
2	D →A	43%	Rejected	493	GK →AT	60%	Rejected
.	.	.	.	494	GKT→A	100%	Accepted
.	.	.	.	495	GT→AK	100%	Accepted
146	G→AK	39%	Rejected	.	.	.	.
147	GK→A	100%	Accepted	694	KNV→AG	100%	Accepted
148	K→AG	46%	Rejected	695	KV→AGN	43%	Rejected
.	.	.	.	696	N→AGKV	50%	Rejected
461	FL→GS	60%	Rejected	697	NV→AGK	75%	Rejected
462	FLS→G	100%	Accepted	698	V→AGKN	16%	Rejected
Assoc = Association, Conf. = Confidence							

2) ***Disease-2: Nephrogenic Diabetes Insipidus***: Here, total 1152 association rules were generated from 234 frequent itemsets. Among these, only 54 rules satisfied the minimum confidence level (90%) and were considered as accepted strong association rules and rest rules were rejected. Few of the accepted rules are shown in Table IV.

TABLE IV. ACCEPTED STRONG ASSOCIATION RULES FOR *NEPHROGENIC DIABETES INSIPIDUS*
(NOT FULL LIST)

Ser	Assoc Rule	Conf.		Ser	Assoc Rule	Conf.
1	K → A	100%		32	AFG → P	100%
2	N → S	100%		33	FG → AP	100%
3	FW → A	100%		.	.	.
.	.	.		.	.	.
16	CV → A	100%		40	FPV → A	100%
17	FV → A	100%		41	GPV → A	100%
.	.	.				
28	DE → P	100%		52	DLP → E	100%
29	FG → P	100%		53	AMT→ L	100%
30	GI → V	100%		54	FLPV→ A	100%

3) ***Disease-3: Retinitis Pigmentosa 4***: Here, total 1252 association rules were generated from 268 frequent itemsets where only 49 satisfied minimum confidence level (90%) and were considered as accepted strong association rules and rest rules are rejected. A few of the accepted rules are shown in Table V.

TABLE V. ACCEPTED STRONG ASSOCIATION RULES FOR *RETINITIS PIGMENTOSA 4* (NOT FULL LIST)

Ser	Assoc Rule	Conf.		Ser	Assoc Rule	Conf.
1	W →A	100%		26	GIT →F	100%
2	W →L	100%		27	FTV →G	100%
.	.	.		31	GPT →F	100%
12	GM →F	100%		.	.	.
13	NY →P	100%		46	ALY →I	100%
.	.	.		47	AVY →I	100%
22	CY →V	100%		48	FNPT →G	100%
23	AFT →G	100%		49	GNPT →F	100%

C. ***Useful Association Rules Identification***

The strong association rules obtained by the previous process are then evaluated by objective measuring tools (*Lift* and *Improve*) and finally only useful rules were sorted out considering following criteria:

- Rules are useful only whose *Lift* >1. The higher the $lift$, the better the rule.
- The higher the *Improve* value, the better the rule.

1) ***Disease-1: Sickle Cell Anemia****: Here, Lift* and *Improve* values of 95 accepted rules (from previous process) were calculated and evaluated. Finally 59 rules were selected as useful strong association rules (Table VI).

TABLE VI: USEFUL STRONG ASSOCIATION RULES FOR *SICKLE CELL ANEMIA*

Ser	Relation	*Lift*	*Improve*	Ser	Relation	*Lift*	*Improve*
1.	GT -> AN	3.75	0.73	31.	AT -> K	1.36	0.27
2.	GT -> KN	3.75	0.73	32.	GT -> K	1.36	0.27
3.	AGT -> KN	3.75	0.73	33.	NT -> K	1.36	0.27
4.	GKT -> AN	3.75	0.73	34.	AGN -> K	1.36	0.27
5.	GT -> AKN	3.75	0.73	35.	AGT -> K	1.36	0.27
6.	AN -> GK	3.00	0.67	36.	ANT -> K	1.36	0.27
7.	GS -> FL	3.00	0.67	37.	GNT -> K	1.36	0.27
8.	NT -> GK	3.00	0.67	38.	ANV -> K	1.36	0.27
9.	KP -> TV	3.00	0.67	39.	ATV -> K	1.36	0.27
10.	ANT -> GK	3.00	0.67	40.	AGNT -> K	1.36	0.27
11.	NT -> AGK	3.00	0.67	41.	AGNV -> K	1.36	0.27
12.	ANV -> GK	3.00	0.67	42.	AD -> G	1.15	0.13
13.	GT -> N	2.50	0.60	43.	AN -> G	1.15	0.13
14.	AGT -> N	2.50	0.60	44.	KN -> G	1.15	0.13
15.	GKT -> N	2.50	0.60	45.	FL -> G	1.15	0.13
16.	AGKT -> N	2.50	0.60	46.	LN -> G	1.15	0.13
17.	KP -> T	2.14	0.53	47.	FS -> G	1.15	0.13
18.	GH -> AL	2.14	0.53	48.	NT -> G	1.15	0.13
19.	GT -> AK	2.14	0.53	49.	NV -> G	1.15	0.13
20.	NT -> AK	2.14	0.53	50.	AKN -> G	1.15	0.13
21.	KPV -> T	2.14	0.53	51.	AFL -> G	1.15	0.13
22.	GNT -> AK	2.14	0.53	52.	FLS -> G	1.15	0.13
23.	GS -> F	1.88	0.47	53.	ANT -> G	1.15	0.13
24.	KN -> AG	1.88	0.47	54.	KNT -> G	1.15	0.13
25.	FS -> GL	1.88	0.47	55.	ANV -> G	1.15	0.13
26.	GLS -> F	1.88	0.47	56.	KNV -> G	1.15	0.13
27.	NT -> AG	1.88	0.47	57.	ALV -> G	1.15	0.13
28.	KNT -> AG	1.88	0.47	58.	AKNT -> G	1.15	0.13
29.	KNV -> AG	1.88	0.47	59.	AKNV -> G	1.15	0.13
30.	AN -> K	1.36	0.27				

2) ***Disease-2: Nephrogenic Diabetes Insipidus***: Similarly, *Lift* and *Improve* values of 54 accepted rules (obtained from previous process) were calculated and evaluated. Finally 14 rules were selected as useful strong association rules as shown in Table VII.

3) ***Disease-3: Retinitis Pigmentosa 4***: Here, *Lift* and *Improve* values of 49 accepted rules (obtained from previous process) were calculated and evaluated. Interestingly all 49 rules were selected as the useful strong association rules as shown in Table VIII.

TABLE VII: USEFUL STRONG ASSOCIATION RULES FOR *NEPHROGENIC DIABETES INSIPIDUS*

Ser	Relation	*Lift*	*Improve*	Ser	Relation	*Lift*	*Improve*
1.	DLP -> E	3.46	0.71	8.	AFG -> P	1.46	0.32
2.	FG -> AP	2.38	0.58	9.	AEL -> P	1.46	0.32
3.	GI -> AV	2.24	0.55	10.	DEL -> P	1.46	0.32
4.	CV -> AL	1.90	0.47	11.	GI -> V	1.27	0.21
5.	AE -> P	1.46	0.32	12.	AGI -> V	1.27	0.21
6.	DE -> P	1.46	0.32	13.	N -> S	1.09	0.08
7.	FG -> P	1.46	0.32	14.	AN -> S	1.09	0.08

TABLE VIII: USEFUL STRONG ASSOCIATION RULES FOR *RETINITIS PIGMENTOSA 4*

Ser	Relation	*Lift*	*Improve*	Ser	Relation	*Lift*	*Improve*
1.	ALS -> W	7.00	0.857	26.	SW -> L	1.21	0.171
2.	W -> AL	3.50	0.714	27.	APW -> L	1.21	0.171
3.	PW -> AL	3.50	0.714	28.	ASW -> L	1.21	0.171
4.	SW -> AL	3.50	0.714	29.	GT -> F	1.17	0.143
5.	QS -> E	2.19	0.543	30.	EM -> F	1.17	0.143
6.	AFP -> S	2.06	0.514	31.	MS -> F	1.17	0.143
7.	NY -> P	1.75	0.429	32.	GM -> F	1.17	0.143
8.	AFS -> P	1.75	0.429	33.	CY -> V	1.17	0.143
9.	AFT -> G	1.59	0.371	34.	AGT -> F	1.17	0.143
10.	FTV -> G	1.59	0.371	35.	GIT -> F	1.17	0.143
11.	FNP -> G	1.59	0.371	36.	GTV -> F	1.17	0.143
12.	FNPT -> G	1.59	0.371	37.	GTY -> F	1.17	0.143
13.	H -> T	1.46	0.314	38.	GPT -> F	1.17	0.143
14.	FH -> T	1.46	0.314	39.	GMN -> F	1.17	0.143
15.	KV -> T	1.46	0.314	40.	MNT -> F	1.17	0.143
16.	QV -> T	1.46	0.314	41.	GNT -> F	1.17	0.143
17.	AY -> I	1.46	0.314	42.	GNPT -> F	1.17	0.143
18.	FGI -> T	1.46	0.314	43.	W -> A	1.09	0.086
19.	FGY -> T	1.46	0.314	44.	LW -> A	1.09	0.086
20.	ALY -> I	1.46	0.314	45.	PW -> A	1.09	0.086
21.	AVY -> I	1.46	0.314	46.	SW -> A	1.09	0.086

22.	W -> L	1.21	0.171	47.	LPW -> A	1.09	0.086
23.	AW -> L	1.21	0.171	48.	LSW -> A	1.09	0.086
24.	CI -> L	1.21	0.171	49.	ILV -> A	1.09	0.086
25.	PW -> L	1.21	0.171				

This work initially identified 698, 1152 and 1252 association rules from *Sickle Cell Anemia*, *Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa-4* diseases respectively. However, after using different tools 95, 54 and 49 were the number of strong association rules and finally the useful rules were found to be only 59, 14 and 49. These final rules indicate the most dominating acids and their patterns for *Sickle Cell Anemia*, *Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa-4* diseases (Fig. 7).

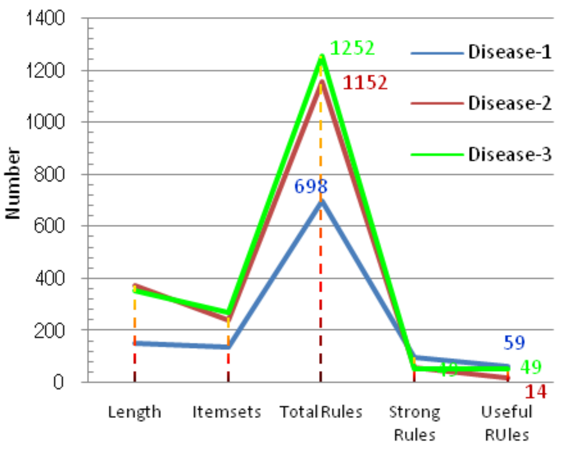

Figure 7. Summary of the Lengths, Itemsets and the Rules for the Protein Sequences of Associated Diseases

VI. Conclusion

Protein, being an integral part of every living organism, if not folded properly may cause critical genetic diseases. As amino acids are the building blocks of protein, relationships among the dominating amino acids and identification of their patterns is an important issue. This work focused to recognize frequent patterns among three complex protein misfolded neurodegenerative human diseases and the relationship of the dominating amino acids using association rule mining. In doing so, itemsets and association rules were generated from the protein sequences. These rules were further evaluated and sorted out with objective measuring tools so that the only strong and interesting patterns are obtained. The patterns acquired from this work are quite impressive which may open up new opportunities in medical science to handle the concerned genetic disorder diseases.

In this work, the experimented disease were associated with relatively small length of protein sequences. However, in future, protein misfolded diseases associated with larger length of protein sequences (e.g. *Cancer*, *Cystic Fibrosis*, etc) may be considered for experimentation. On the other hand, more improved objective measuring tools for usefulness measures of association rules may be applied to obtain more reliable and strong patterns of amino acids.

References

[1] S. Rajasekaran and L. Arockiam, “Frequent Contiguous Pattern Mining Algorithms for Biological Data Sequences”, International Journal of Computer Applications, Vol. 95, No. 14 (2014), 15-20.

[2] G. LakshmiPriya and S. Hariharan, “A Study on Predicting Patterns Over the Protein Sequence Datasets using Association Rule Mining”, Journal of Engineering Science and Technology, Vol. 7, No. 5(2012) 563 – 573.

[3] R. J. Ellis and T. J. Pinheiro, “Danger – misfolding proteins”, Nature, Vol. 416 (2002), 483–484 .

[4] T. K. Chaudhuri and S. Paul, “Protein-misfolding Diseases and Chaperone-based Therapeutic Approaches”, Federation of European Biochemical Societies (FEBS) Journal 273 (2006), 1331–1349.

[5] G. Lakshmi Priya and S. Hariharan, “An Efficient Approach for Generating Frequent Patterns without Candidate Generation”, International Conference on Advances in Computing, Communications and Informatics (ICACCI), *ICACCI’12*, August 3-5, 2012, 1061 – 1067.

[6] S. Dhumale, “Predicting Patterns over Protein Sequences Using Apriori Algorithm”, International Journal of Engineering and Computer Science, Vol. 4 Issue 7, 2015, 13011-13016.

[7] The Chemistry of amino acid. (2003, September 30). Retrieved from www.biology.arizona.edu/ biochemistry/problem_sets/aa/aa.html.

[8] Chemistry of amino acids and protein structure. (n. d.). Retrieved May 12, 2018, from https://www.khanacademy.org/test-prep/mcat/biomolecules/amino-acids-and-proteins1/a/chemistry-of-amino-acids-and-protein-structure

[9] J. Bradbury, “Chaperones: keeping a close eye on protein folding”, The Lancet, Vol. 361, Issue. 9364 (April 2003), 1194–1195.

[10] UniProtKB - P68871 (HBB_HUMAN). Retrieved September 15, 2017, from https://www.uniprot.org/uniprot/P68871

[11] UniProtKB - P30518 (V2R_HUMAN) . Retrieved September 15, 2017, from https://www.uniprot.org/uniprot/P30518

[12] UniProtKB - P08100 (OPSD_HUMAN) . Retrieved September 15, 2017, from, https://www.uniprot.org/uniprot/P08100

[13] M. Gupta and J. Han, “Applications of Pattern Discovery Using Sequential Data Mining”. Retrieved August 28, 2017, from, https://www.microsoft.com/en-us/research/wp-content/uploads/2012/01/gupta11b_apdsdm.pdf

[14] Mining Frequent Itemsets – Apriori Algorithm. (n. d.). Retrieved September 29, 2017, from http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab8-Apriori.pdf

[15] C. Ju, F. Bao, C. Xu and X. Fu, “A Novel Method of Interestingness Measures for Association Rules Mining

ISBN: 978-1-63248-157-3 doi: 10.15224/978-1-63248-157-3-12

Bassed on Profit", Discrete Dynamics in Nature and Society, Vol. 2015, (2015).

[16] T. H. Kang, J. S. Yoo and H. Y. Kim," Mining Frequent Contiguous Sequence Patterns in Biological Sequences", Proceedings of 7th IEEE International Conference on Bioinformatics and Bioengineering (BIBE'08), Athens, Oct 8-10, 2008, pp 723-728

[17] What are proteins made of? (n.d.). Retrieved October 26, 2017, from, http://whoami.sciencemuseum.org.uk/whoami/findout more/yourbody/whatdoyourcellsdo/whatisacellmadeof/whatareproteinsmadeof.