FREQUENT CONTIGUOUS PATTERN MINING OVER BIOLOGICAL SEQUENCES OF PROTEIN MISFOLDED DISEASES

MOHAMMAD SHAHEDUL ISLAM

(BSc Engg., KU)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DEPARTMENT OF COMPUTER ENGINEERING

MILITARY INSTITUTE OF SCIENCE AND TECHNOLOGY

2019

The thesis titled FREQUENT CONTIGUOUS PATTERN MINING OVER BIOLOGICAL SEQUENCES OF PROTEIN MISFOLDED DISEASES Submitted by MOHAMMAD SHAHEDUL ISLAM Roll No: 1014140001 (P) Session: Oct 2014 has been accepted as satisfactory in partial fulfilment of the requirement for the degree of M. Sc. Engineering (CSE)on……

BOARD OF EXAMINERS

1. _____ Chairman

Name: Dr. Md. Abul Kashem Mia
Designation: Professor
Affiliation: Department of CSE
Bangladesh University of Engineering and Technology
Dhaka-1000, Bangladesh

2. _____ Member

Name: Dr Md Mahbubur Rahman, PhD
Designation: Professor
Affiliation: Department of CSE, MIST, Dhaka

3. _____ Member

Name: Air Commodore Md Afzal Hossain, ndc, psc (Ex-officio)
Designation: Senior Instructor / Professor & Head
Affiliation: Department of CSE, MIST, Dhaka
Designation & Address

4. _____ Member

Name: Dr. M. Sohel Rahman (External)
Designation: Professor
Affiliation: Department of CSE
Bangladesh University of Engineering and Technology
Dhaka-1000, Bangladesh

DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

_____

Mohammad Shahedul Islam
13 June 2019

ACKNOWLEDGEMENT

First of all, I would like to thank to almighty Allah for giving me the strength to work on this thesis. At the same time, I would like to express my gratitude to CSE Department of MIST for giving me an opportunity to work for this thesis as a part of the degree.

Special and heartfelt thanks to my supervisor, Dr. Md. Abul Kashem Mia for encouraging me to work with this interesting subject. Despite of busy schedule, his regular supervision, guidance and dedicated supporting role is worth to mention. In preparing the report, he actively supported me and reviewed the report painstakingly. I thankfully acknowledge his contribution to enrich the report.

I would like to acknowledge, with utmost gratitude, the significant contribution of Mr. Swapnil Saha, in data analysis part. Heartfelt thanks to Dr. Gazi Nurun Nahar Sultana, Chief Scientist, Genetic Engineering and Biotechnology Research, Centre for Advanced Research in Sciences (CARS), University of Dhaka for giving domain expert opinions regarding the implications of such research findings. Special thanks to Dr. Md. Shamsur Rahman, Monash University, for his sincere support, suggestion and continuous encouragement in completing the work. I also wish to express my sincerest appreciation to Prof. Dr. Md. Shamsul Arefin, CUET, Assoc Prof Nazrul and Lt Cdr Anis, MIST, Senior Lec Md. Sarwar Kamal, EWU, Doctor Ratan, BSM Medical University and Prof Dr Altaf Hussain, BSMRMU. Their sincere and wholehearted cooperation, consultation and opinions along with supporting information helped me a lot in doing the thesis work. Thanks to my service staffs specially Nimai Chandro for giving clerical support.

And at last, my heartfelt thanks to my wife who gave me constant support and encouragement in the successful completion of this work.

TABLE OF CONTENTS

SUMMARY

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 a unique 3D shape, where a minor error may cause misfolded structure. Genetic disorder diseases i.e. *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 protein misfolded disease i.e. *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 work, we have analyzed the protein sequences associated with more complex protein misfolded diseases by association rule mining technique, where only the useful rules are finally sorted out with the use of interestingness measures.

This work initially generated 135, 1806, 1464, 234 and 268 itemsets from *Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus (NDI),* and *Retinitis Pigmentosa 4 (RP4)* respectively. Then the algorithm generated association rules from those itemsets. The association rules which fall below the threshold Confidence (90%) were pruned as strong association rules. After using objective measuring tools over these strong association rules, the final useful rules were found to be only 59, 19, 35, 14 and 49. These final rules indicate the most dominating amino acids and their patterns for the five diseases. Adopting the quantitative experimental method, this work forms more reliable, useful and strong association rules among the most domination amino acids of corresponding misfolded

proteins and identifies the dominating patterns of amino acid of complex protein misfolded diseases.

Patterns in protein sequences usually have functional, structural or family classification importance. Pattern identification can be used for predicting protein functions, protein fold (structure) recognitions, protein family detection, multiple sequence alignment, etc. The patterns acquired from this work are quite impressive. In addition to the above usual applications, the identified amino acid patterns could be more useful in discovering medicines for concerned protein misfolded diseases and thereby this work may open up new opportunities in medical science to handle genetic disorder diseases.

LIST OF TABLES

LIST OF FIGURES

CHAPTER-1: INTRODUCTION

1.1 Introduction

To survive, all living being need proteins, either in muscles or in cell membrane. Proteins 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. In the folding process, minor error may cause misfolded structure leading to serious consequence. Many cancers and genetic disorder diseases such as Alzheimer's, Parkinson's, Sickle cell anemia, etc are believed to be caused for protein-misfolding. Thus, the relationship between these amino acids is very vital in case of protein misfolded diseases. In pursuant to this, the objective of this research is to identify frequent patterns over biological sequences of protein misfolded diseases, namely *Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa-4*. In this work, protein associated with each disease has been analyzed by association rule mining technique to discover frequent patterns among the amino acids. The association rules were considered to be strong if it satisfied both a minimum support and a confidence threshold. Quantitative experimental study has been conducted to form association rules among the most dominating amino acids for above diseases. This work identified 59, 19, 35, 14 and 49 strong and useful association rules among the most domination amino acids respectively for *Sickle Cell Anemia*, *Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa-4* diseases. Identification of the patterns of most dominating amino acids of the above genetic disorder diseases may open up new opportunities in medical science.

1.2 Problem Definition

Frequent Contiguous Patterns (FCP) 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 FCP [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 that acts as the constituents of all living organisms [2]. Protein misfolding is believed to be 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 [3]. Proteins are made up of smaller building blocks called amino acids, joined together in chains [4]. These chains of amino acids fold up in complex ways, giving each protein a unique 3D shape. 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 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 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 the pattern and association rules of the most dominating amino acids which cause the *Chromaffin Tumor*

disease only. However, predicting the pattern and associations between more complex diseases are yet to be attempted in the literature.

1.3 Application

Patterns in protein sequences usually have functional, structural or family classification importance. It is assumed that these regions are between conserved in evolution and therefore they occur more frequently [7]. Pattern identification can be used for predicting protein functions, protein fold (structure) recognitions, protein family detection, multiple sequence alignment, etc. Protein sequences of the same family typically hold identical patterns and thus if a protein sequence contains patterns common to other protein sequences then it is likely that the protein sequences are biologically related and may belongs to same family. On the other hand, patterns of conserved sequences can often highlight elements that are responsible for structural similarity between proteins and can be used to predict the 3D structure of a protein [7]. Moreover, protein patterns can be used to predict the functions of newly discovered or unknown proteins or to screen genomic databases for other proteins with similar functionality [7].

This thesis work is focused to predict the pattern and association rules of the most dominating amino acids in the protein sequences associated with particular protein misfolded diseases. In addition to the above usual applications, the identified amino acid patterns could be more useful in discovering medicines for concerned protein misfolded diseases and thereby this work may open up new opportunities in medical science to handle genetic disorder diseases.

This study focused on most dominating amino acids change in five genetic diseases resulting from protein misfolding. These dominating amino acids change not only damage the protein formation but also to the structure and biochemical properties with physiological effects ranging from insignificant to severe. Thus identification/reporting of such variant of amino acids for those particular five genetic diseases may have versatile implications. In this regard, Dr. Gazi Nurun Nahar Sultana, Chief Scientist, Genetic Engineering and Biotechnology Research, Centre for Advanced Research in Sciences (CARS), University of Dhaka (personal communication, Jun 23, 2019) highlighted a number of implications of such findings:

- It can be applied for gene study through DNA sequencing, thus particular mutation can be edited through research.
- With the information of such data mining, prenatal diseases can be identified,
- Also disease susceptibility can be predicted through most dominating amino acid changes.
- Overall, such data gives the physicians to take the necessary treatment action as well as genetic counselling.
- Such data can be resource for new drug discovery.

1.3.1 Implications in terms of Medical Science

The findings of this research work has important role in terms of medical science also. In this aspect Dr. Gazi Nurun Nahar Sultana also added her views. “An improved capacity in identifying the relations among the most dominating amino acids in protein sequences related to disease will have an immediate impact on the diagnosis, treatment, and

prevention of genetic disorders. As more population based data are accumulated, amino acids based diagnosis will become more common and the potential for somatic cell gene therapy will increase. Furthermore, the availability of molecular probes for specific gene loci will permit detection of the carriers of disease-associated genes. This ability will enable parents to identify the extent to which their offspring may be at risk for a genetic defect." (G. N. N. Sultana, personal communication, Jun 23, 2019).

1.3.2. Implications in terms of Genetics, Bioinformatics and Biotechnology:

The results of this research work also possess important roles in terms of Genetics, Bioinformatics and Biotechnology. In this regard, Dr. Sultana (personal communication, Jun 23, 2019) pointed out following implications:

- Implication of Amino acids Database is important for studying monogenic and complex genetic disease. Genomic browser allows association between disease phenotype and genetic loci. Also important resource for continuous centralized monitoring of genome and metagenome projects at home and worldwide
- Computational approach can overcome the etiology. Understanding the complex interplay between genes and proteins requires integration of data from a wide variety of sources, i.e. gene expression, genetic linkage, protein interaction, and protein structure among others. Thus, this database can become critical for the integration, representation and visualization of heterogeneous biomedical data.
- Biotechnologically, it might allow development of new drugs for treatment and tools/biomarker for disease diagnosis.

1.3.3 Implications in terms of Protein Sequencing Research

The relationship between the most dominating amino acids in the five genetic diseases resulting from protein misfolding may also have insinuation in terms of Protein Sequencing research. In this aspect Dr. Gazi Nurun Nahar Sultana also gave her opinion. “Identifying the relations among the most dominating amino acids in protein sequences of associated genetic diseases can be implemented by focusing on how a protein leads to the heritable form of the respective disease. Till now, researchers typically have obtained clues into the molecular basis of the disorder. Still, identifying the precise molecular culprit in the cascade of events following the gain of function leading to the associated disease outcome has not been straightforward for researchers, and debates continue. So research on understanding the normal function of genetically associated proteins in such diseases can be marginalized the complex roles of these proteins play in their respective disorders.” (G. N. N. Sultana, personal communication, Jun 23, 2019).

1.4 Literature Review

Frequent Contiguous Patterns (FCP) are small patterns that repeatedly occurs in a database, specially high in bio-sequences. 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 [8]. 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 [1]. Frequent pattern mining was first proposed by Agrawal (1993) for market basket analysis in the form of association rule mining. Based on the concept of strong association rules, Agrawal [9] introduced

association rules for learning uniformities between products of large scale transactions in supermarkets as recorded by their point-of-sale systems. This was termed as market basket analysis. In addition to market basket analysis, association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics [10]. The algorithm developed by Agrawal [9] for association rule mining is called Apriori Algorithm. Here, association rules are developed by studying data for frequent/significant *if – then* patterns and applying the measures of *support* and *confidence* level to establish the most significant relationships.

Apriori algorithm has widely been used for predicting frequent patterns from large biological sequences. Abundant literature has been dedicated to this research and tremendous progress has been made, ranging from efficient and scalable algorithms for frequent itemset mining in transaction databases to numerous research frontiers, such as sequential pattern mining, structured pattern mining, correlation mining, associative classification, and frequent pattern-based clustering, as well as their broad applications [11].

Biological sequences such as DNA and protein sequences consist of long linear chain of chemical components and typically contain a large number of items [12]. These sequences hold contiguous sequences which typically consist of more than hundreds of frequent items. DNA sequences contain four nucleotides namely Adenine (A), Cytosine (C), Guanine (G) and Thymine (T), protein sequences contain 20 amino acids, a gene sequence is a sequence of nucleotides arranged in a specific order and a genome is the complete set of genes of an organism[1]. The challenging task in pattern finding of biological sequences is to find frequent contiguous patterns [1]. Data Mining has recently increased its popularity in

classifying the biological sequences and structures based on their critical features and functions [1].

Tae Ho Kang and his team [12] proposed an algorithm to efficiently find the frequent maximal contiguous sequences from several biological data. The proposed algorithms could accept several values of minimum support threshold and could produce results only by spanning tree search. It could be applied to DNA sequence with a small number of items (dimension) and amino acid sequence with a large number of items. The author [13] focused on protein–DNA bindings between transcription factors (TFs) and transcription factor binding sites (TFBSs) and proposed a framework to discover associated TF–TFBS binding sequence patterns in the most explicit and interpretable form. The framework was based on association rule mining with Apriori algorithm. NN Das and Poonam [14] has studied different sequence mining algorithms and proposed a new algorithm for generating frequent patterns from DNA sequences only. But interestingly the literature [14] did not show any experimental results of finding the frequent patters from NDA sequence. Mutakabbir, Mahin and Hasan [15] proposed two algorithms. The first algorithm indexed the unique sequences of length four using an integer value and the second algorithm discovered the frequency of the frequent patterns of various lengths by searching through the integer values instead of the patterns themselves. All this was done by the use of mapping techniques e.g. Hash Map, where the subsequent nucleotide sequences (any combination of A, T, C and G) of a given size could be identified within a particular DNA sequence.

Rajasekaran and Arockiam [1] analysed different pattern mining algorithms for bio sequence. SP-Index method scans the database for matching with the patterns of existing database and then next level patterns are generated with the help of SP-Index trees. Location based FCP generates pattern table with patterns and their locations for all existing patterns in the DB by scanning the DB and then sort the pattern table by last occurring position [1]. Fast Contiguous FCP creates a spanning tree for base patterns using position information and reduces the search space and time by using hash table and binary search respectively. All of these algorithms have advantages and disadvantages.

Jingsong Zhang and his team [16] proposes an algorithm (Con Sgen) for discovering contiguous sequential generators which adopts n-gram model, called shingles, to generate potential frequent subsequences and leverages several pruning techniques to prune the unpromising parts of search space and then the contiguous sequential generators were identified by using the equivalence class-based lower-closure checking scheme. They experimented the algorithm on both DNA and protein data sets.

Protein is one among the important factors and acts as the constituents of all living organisms [2]. Protein misfolding is believed to be 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 [3]. Proteins are made up of smaller building blocks called amino acids, joined together in chains [4]. These chains of amino acids fold up in complex ways, giving each protein a unique 3D shape. Thus, the relationship between these amino acids is very vital in case of protein misfolded diseases. 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, and 6] have been identified on this issue.

G. Lakshmi Priya and S. Hariharan [5] aimed at extracting the hidden and the most dominating amino acids among the infected protein sequence which causes some infections in human. They tried to predict patterns applying strong association rules over the frequent itemsets of the protein sequence named *Succinate dehydrogenase* (DHSB_HUMAN) which is involved in *chromaffin tumor* disease. The researchers [5] named their system as GENPAT and focused in finding the most dominating amino acids (in *Succinate dehydrogenase* protein) which causes the disease *chromaffin tumor*. The system functioned by generating frequent itemsets from the protein sequence and construct a frequent pattern tree. Thereafter strong association rules were generated based on 90% confidence thresholdto identified the domination amino acids.

G. Lakshmi Priya and S. Hariharan [2] again 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 applied in finding the 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. However, the research could predict relatively larger pattern than the earlier one [5].

Almost similar another work was carried out by S. Dhumale [6] which was focused on finding the most dominating amino acids responsible to cause five diseases that is *Epilepsy*, *Hartnup*, *Cystinuria*, *Alzheimer's* disease and *chromaffin tumor*. This work considered a protein sequence from NCBI (National Center for Biotechnology Information) database and used Apriori algorithm to find the domination amino acid pattern. As experimental result, the author claimed five amino acid patterns (association rules), each to be responsible for above individual diseases. This work suffers serious limitations. First, the protein sequence which was considered here is anonymous. Secondly, all the mentioned diseases might not be associated with a single protein. The author did not give any reliability of the information and my frequent search also could not generate any authenticity in this regard. It is to mention that all diseases are not associated with the protein changes. Some are multifactorial diseases, some are infectious diseases and so on. Thirdly, the author arbitrarily increased the minimum *Support* count value from 2 to 5, generated association rules with confidence threshold 90% and declared set of amino acid pattern (association rule) responsible for each of the disease. But on what basis this deduction was arrived was not at all cleared.

The above three works were focused to predict the pattern and association rules of the most dominating amino acids which causes the *Chromaffin Tumor* disease. However, predicting the pattern and associations between more complex protein misfolded diseases are yet to be attempted in the literature.

1.5 Objectives

This work is grounded in two general research fields i.e. Bioinformatics and Genetics. The outcomes of this research will greatly contribute in these areas as well. Thus, the objectives of this research are:

- To identify frequent patterns over biological sequences of protein misfolded diseases using association rule mining.

- To generate strong association rules for the most dominating amino acids of five protein misfolded diseases, namely *Sickle Cell Anemia*, *Breast Cancer Type 1*, *Cystic Fibrosis*, *Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa 4*.

1.6 Summary of Result

In this work, the biological sequences of five protein misfolded diseases, namely *Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa 4*were experimented to find out the most dominating amino acids and their pattern. In connection to this, five protein sequences as associated with the aforesaid diseases were processed and examined. The work thus generated the following:

a. Frequent itemsets

b. Strong association rules

c. Useful of association rules

The summary of the result is shown in Table 1.1 and Figure 1.1 is its graphical representation.

Table 1.1: Summary of the Result

S No	Disease	Lengths of Associated Protein	Frequent Itemsets	Total Association Rules	Strong Association Rules	Useful Association Rules
1.	Sickle Cell Anemia	147	135	698	95	59
2.	Breast Cancer	1863	1806	20884	80	19
3.	Cystic Fibrosis	1480	1464	14792	96	35
4.	Nephrogenic Diabetes Insipidus (NDI)	371	234	1152	54	14
5.	Retinitis Pigmentosa 4 (RP4)	348	268	1252	49	49

This work initially generated 135, 1806, 1464, 234 and 268 itemsets from *Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus (NDI),* and *Retinitis Pigmentosa 4 (RP4)* respectively. Then the algorithm generated association rules from those itemsets. The association rules which fall below the threshold Confidence (90%) were pruned as strong association rules. After using objective measuring tools over these strong association rules, the final useful rules were found to be only 59, 19, 35, 14 and 49. These final rules indicate the most dominating amino acids and their patterns for *Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus (NDI),* and *Retinitis Pigmentosa 4 (RP4*).

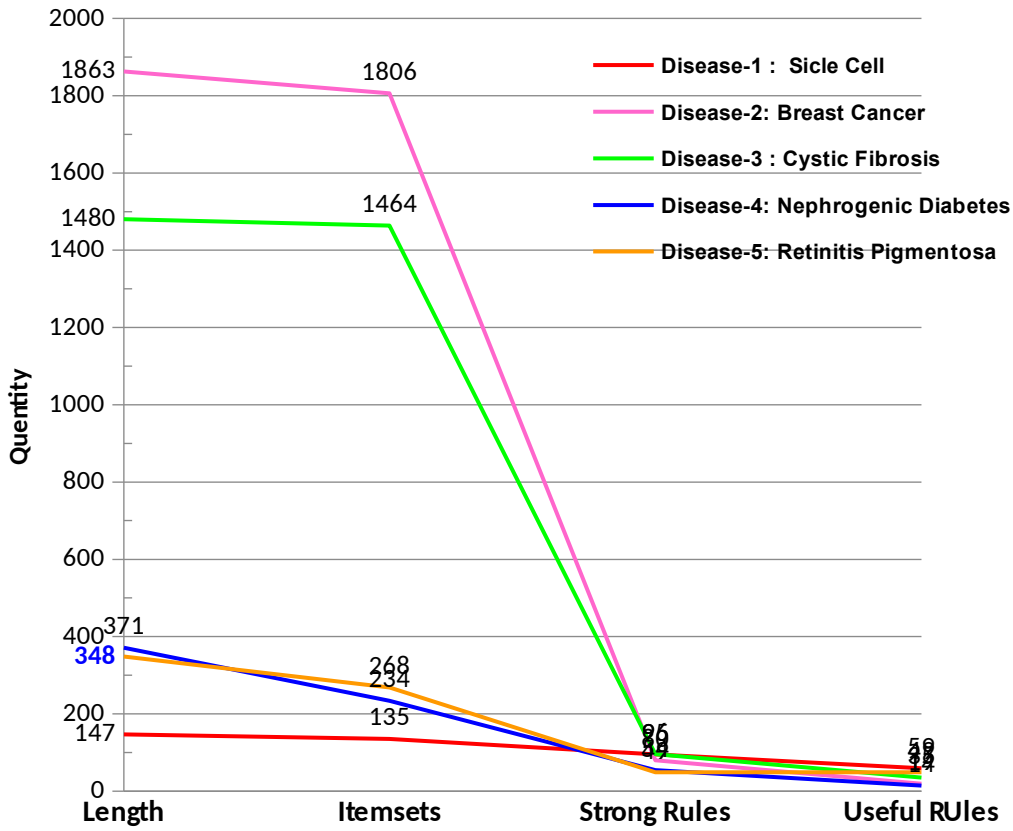


Fig 1.1: Summary of the Result

1.7 Thesis Organization

The thesis is divided into four chapters. Chapter 1 covers the definition of the problem, application of thesis findings, literature review, objective of the thesis and summary of the results. In Chapter 2, the theoretical background of the subject area including different protein misfolded diseases, relevant pattern mining algorithm, usefulness measuring tools, etc are discussed. Chapter 3 is the fundamental section of this thesis. It covers the steps of the experiment, pseudocode of the algorithm used in the system, experimental results and analysis of the result. Lastly the conclusion and future works are covered in Chapter 4.

CHAPTER-2: PRELIMINARIES

2.1 Introduction

Protein is an essential molecule for all living being. Protein is made up of numerous amino acids which determine the biological activities of that protein. Protein folding is an important is an important process for all living creature. Error in the folding process may cause malfunctioning even critical genetic disorder diseases. Thus inferring the relationship among the amino acids is important to analyze the protein misfolding and associated diseases. Identifying the most dominating amino acids and their relationship patterns is essential. Development in Data Mining techniques can play important role to discover these hidden relationships among the most dominating amino acids. In this Chapter, the structure of the protein, diseases associated with protein misfolding are discussed. In sequel of that, association rule mining technique and its measuring tools are also elaborated here.

2.2 Amino Acid

To survive, all living being needs proteins, either used in muscles or even in the cell membrane. Amino acids are used to build these proteins in every living cell. Amino acids play central roles both as building blocks of proteins and as intermediates in metabolism [18]. The biological activity of the protein is determined by the chemical properties of the amino acids. It is important to understand amino acid structure and properties because it is essential to comprehend the structure and properties of the protein. Amino acids are made from carbon, hydrogen, nitrogen, and oxygen. However, all amino acids have five basic parts:

i. a central carbon atom(C),

ii. a hydrogen atom (H),

iii. an amino group (-NH_2) which consist of a nitrogen atom and two hydrogen atoms,

iv. a carboxyl group (-COOH) which consist of a carbon atom, two oxygen atoms, and one hydrogen atom, and

v. an R-group or side chain.

Figure 2.1 shows the basic structure of an amino acid. Amino acid is characterized as unique due to its R-group (side chain). Each of the 20 amino acids has a different side chain structure. Side chain permits an amino acid to react with other amino acids in distinct ways. Side chains contain mainly hydrogen, carbon, and oxygen atoms, whereas some may have sulfur or nitrogen atoms in their R-groups [19].

Each amino acid has a name, abbreviation and side chain structure (Table 2.1).Though more than 50 amino acids have been discovered by the scientists; only 20 are used to make proteins in human body. Nine of those 20 are marked as essential. The other 11 can be synthesized by an adult body. The 20 amino acids that are found within proteins convey a vast array of chemical versatility [18]. Thousands of combinations of those 20 amino acids are used to build all of the proteins in human body.

Table2.1: Representation of Amino Acid in One and Three Letter Codes

Seria	One-letter Code	Three-letter Code	Amino Acid Name

Table2.1: Representation of Amino Acid in One and Three Letter Codes

l			
1	A	Ala	Alanine
2	B	Asx	Aspartic acid or Asparagine
3	C	Cys	Cysteine
4	D	Asp	Aspartic acid
5	E	Glu	Glutamic acid
6	F	Phe	Phenylalanine
7	G	Gly	Glycine
8	H	His	Histidine
9	I	Ile	Isoleucine
10	K	Lys	Lysine
11	L	Leu	Leucine
12	M	Met	Methionine
13	N	Asn	Asparagine
14	O	Pyl	Pyrrolysine
15	P	Pro	Proline
16	Q	Gln	Glutamine
17	R	Arg	Arginine
18	S	Ser	Serine
19	T	Thr	Threonine
20	U	Sec	Selenocysteine
21	V	Val	Valine
22	W	Trp	Tryptophan
23	X	Xaa	Any amino acid
24	Y	Tyr	Tyrosine
25	Z	Glx	Glutamic acid or Glutamine

2.3 Protein

Proteins are complex molecules, made up of hundreds of smaller units called amino acids that are attached to one another by peptide bonds, forming a long chain [20]. Figure 2.2 and Fig. 2.3 show how the amino acids joins together by peptide bonds and make protein. The human body has thousands of different proteins, all of which are necessary for staying alive and healthy [19]. Proteins have a wide array of crucial functions in human bodies such as they store amino acids, function as antibodies, act as hormones, have structural functions, transport important molecules and last but certainly not least, proteins can act as enzymes [21]. The precise amino acid content,

and the sequence of those amino acids, of a specific protein, is determined by the sequence of the bases in the gene that encodes that protein [18].

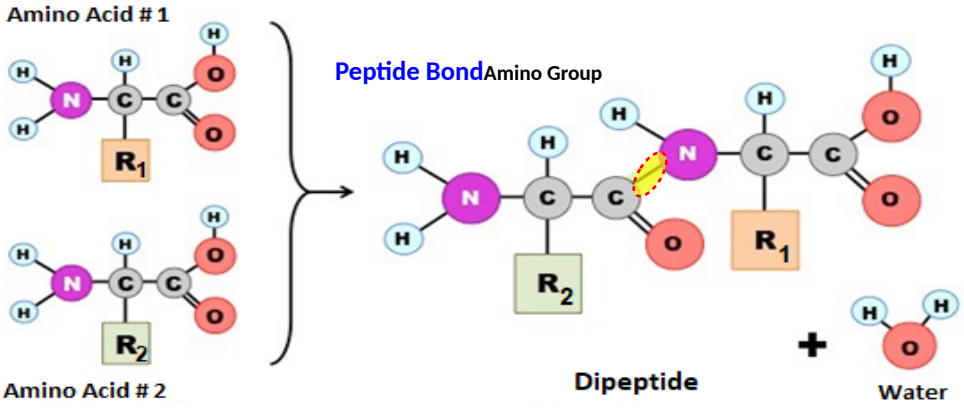


Fig 2.2 : Amino Acids are Joined Together Through Peptide Bonds

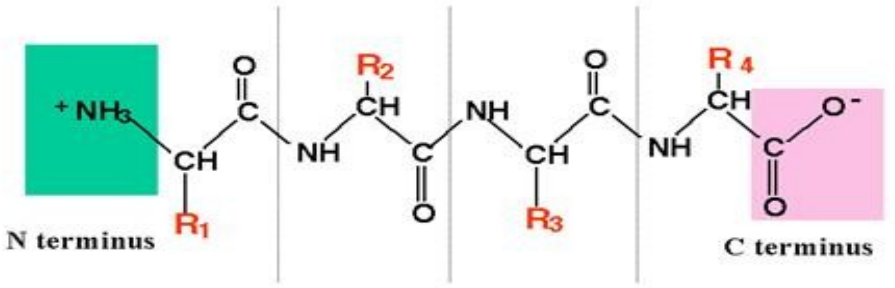


Fig 2.3 : Polypeptide Chain of Amino Acids to Build Primary Structure of a Protein

Proteins are organic compounds of amino acids arranged in a linear chain and folded into a globular form and are called as amino acid polymers [5]. Each protein sequence has four levels of structure:

a. Primary structure. At the initial level protein takes the primary structure which is a straight chain of amino acids i.e. linear polypeptide with amino acids sequence.

b. Secondary structure. The secondary structure comes after the primary structure. Here, the original chain of primary structure begins to fold and twist. The secondary structure is the folded version of the linear polypeptide stabilized by hydrogen bonding [20]. In the chain, each of the amino acids interacts with the others and it twists like a corkscrew (alpha helix) or it takes the shape of a folded sheet (beta sheet).

c. Tertiary structure. Several secondary structures come together and held together by different types of interactions and form the tertiary structure. Hydrogen bonds, hydrophobic interactions, ionic bonds, and disulfide bonds are involved in the stability of tertiary structures [21].

d. Quaternary structure. This is the fourth and final phase in the building process of a protein. In this level, numerous amino acid chains from the tertiary structures fold together in to a globular form.

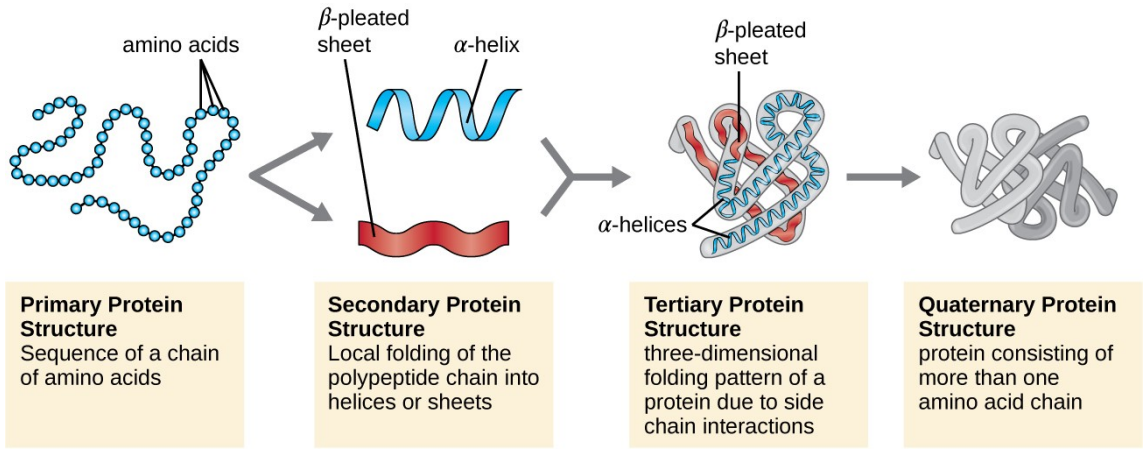


Fig 2.4 : Various Levels of Protein Structure

Amino acids sequences contain the necessary information, basing on which, protein determine how that protein will fold into a three-dimensional structure and the stability of the resulting structure. Protein folding and its stability have become a critically

significant area of research for last two decades and progress is being made in every days.

2.4 Protein Misfolding

A protein can be functionally active when it acquires a unique three-dimensional conformation through the complicated folding of the polypeptide chain (amino acids chain) coded from the nuclear genome. The folding pathway is defined by the core amino acid sequence and the local cellular environment. For any living organism, protein folding is a crucial issue because it adds flesh to the gene skeleton. A small error in the folding process results in a misfolded structure, which can sometimes be lethal [22]. However, it has been observed that many proteins cannot fold properly by themselves within the cellular environment. Changes in the polypeptide chain, either resulting from inherited or acquired gene variations or from abnormal amino acid modifications, may change the folding process and give rise to misfolding of the protein [23].

Proteins that are not able to achieve the native state, due either to an unwanted mutation in their amino acid sequence or simply because of an error in the folding process, are recognized as misfolded and subsequently targeted to a degradation pathway [3]. Due to misfolding, a protein may have adverse effect on its functionality, such as:

a. The protein may lose its usual function. This phenomenon is observed in case of cystic fibrosis (CF) and a1-antitrypsin deficiency diseases.

b. The protein may gain deleterious function. This phenomenon is noticed in many neurodegenerative diseases such as Parkinson's, Alzheimer's and

Huntington's disease, in which protein misfolding results in the formation of harmful amyloid.

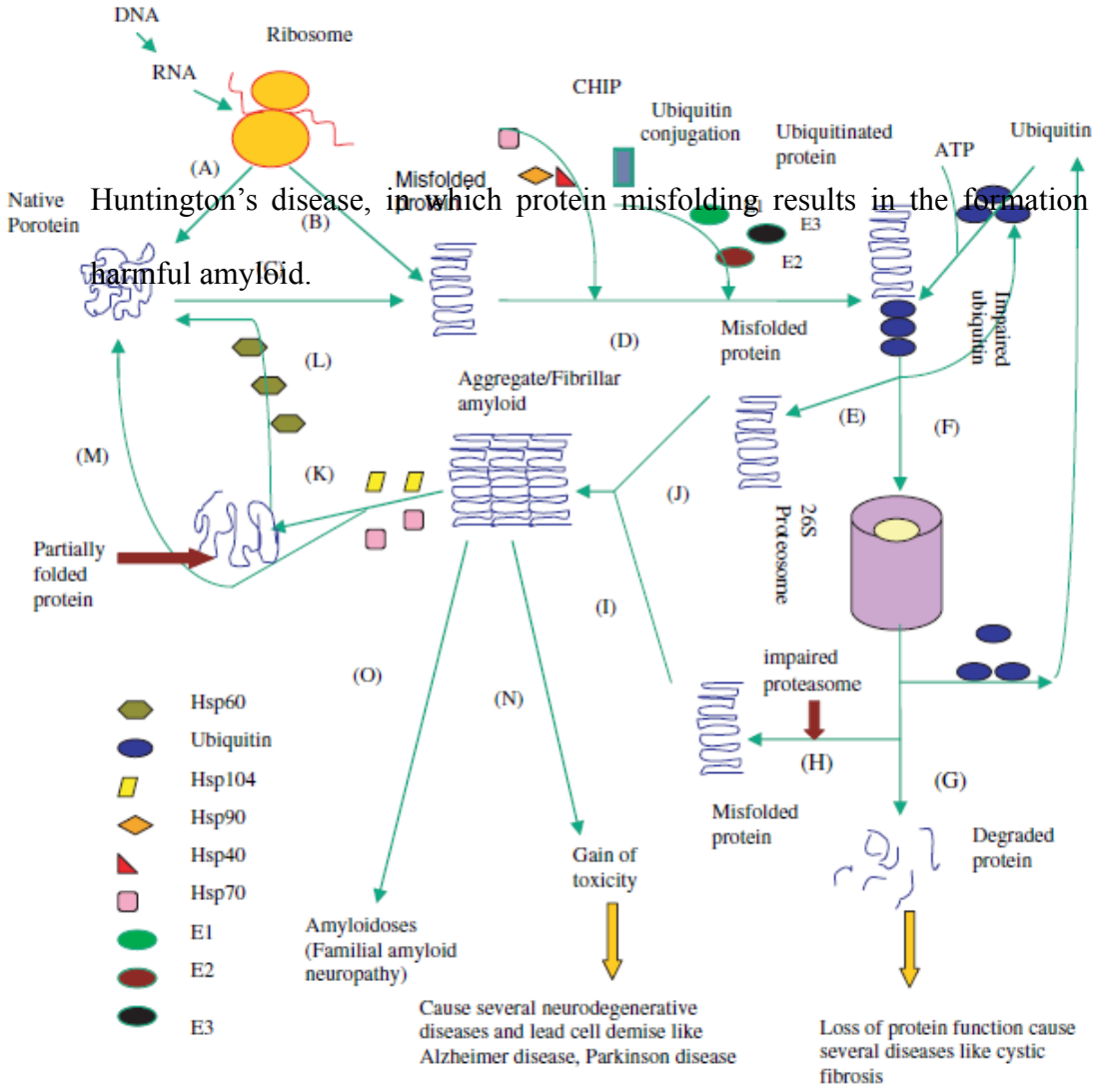


Fig 2.5 : Sequence of Cellular Misfolded Protein [3]

In some instances, the mutations are so acute in nature that they turn the gene product biologically inactive. This happens in case of cystic fibrosis transmembrane regulator (CFTR) protein. In other cases, the mutations are rather minor and the resulting proteins

lose its partial functions of usual activity. However, mutation in the gene (encoding the disease-causing protein) is very common in almost all cases of protein misfolding mediated disorders [3].Over the last two decades, protein misfolding and its pathogenic effect have become a significant area of human bio-molecular research.

2.5 Protein Misfolding Diseases

For the last couple of years, protein misfolding and its effects have become a matter of great concern. According to the prion researcher Susan Lindquist, 'protein misfolding could be involved in up to half of all human diseases' [24]. Many cancers and other protein-misfolding disorders are caused by mutations in proteins. Protein misfolding is believed to be 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 [3]. Table 2.2 shows a list of human diseases caused by protein misfolding, aggregation or trafficking.

Table 2.2: Protein's Misfolding Involved in Different Human Diseases [3]

Proteins	**Disease**
Hemoglobin	Sickle cell anemia
CFTR protein	Cystic fibrosis
Prion protein (PrP)	Creutzfeld Jakob disease
S	Scrapie (Mad Cow Disease)
F	Familial insomnia
Huntingtin	Huntington's disease
b-amyloid protein	Alzheimer's disease
b-glucosidase	Gaucher's disease
a-Synuclein	Parkinson's disease
V2 vasopressin receptor	Nephrogenic diabetes insipidus
Transthyretin	Transthyretin amyloidoses
Rhodopsin	Retinitis pigmentosa
P53	Cancer

In this work, five protein misfolded diseases (i.e. Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus and Retinitis Pigmentosa 4) have been taken in consideration.

2.5.1. **Sickle Cell Anemia (SKCA)** Sickle cell anemia disease is caused by mutations affecting the gene represented in this entry, i. e. misfolding of protein, Hemoglobin Subunit Beta. It is a genetic disorder in which the amino acid valine at the sixth position of the -globin chain is replaced by glutamine [3].

Disease description [25]: The disease is characterized by abnormally shaped red cells resulting in chronic anemia and periodic episodes of pain, serious infections and damage to vital organs. Normal red blood cells are round and flexible and flow easily through blood vessels, but in sickle cell anemia, the abnormal hemoglobin (called Hb S) causes red blood cells to become stiff. They are C-shaped and resemble a sickle. These stiffer red blood cells can led to micro vascular occlusion thus cutting off the blood supply to nearby tissues.

2.5.2. **Breast Cancer (BC)** [26] Disease susceptibility is associated with variations affecting the gene represented in this entry. Mutations in protein, Breast Cancer Type 1 (BRCA1), are thought to be responsible for 45% of inherited breast cancer. Moreover, BRCA1 carriers have a 4-fold increased risk of colon cancer, whereas male carriers face a 3-fold increased risk of prostate cancer. Cells lacking BRCA1 show defects in DNA repair by homologous recombination.

Disease description: A common malignancy originating from breast epithelial tissue. Breast neoplasms can be distinguished by their histologic pattern. Invasive ductal carcinoma is by far the most common type. Breast cancer is etiologically and genetically heterogeneous. Important genetic factors have been indicated by familial occurrence and

bilateral involvement. Mutations at more than one locus can be involved in different families or even in the same case.

2.5.3. Cystic Fibrosis The disease is caused by mutations affecting the gene represented in this entry.

Disease description [27]: A common generalized disorder of the exocrine glands which impairs clearance of secretions in a variety of organs. It is characterized by the triad of chronic broncho pulmonary disease (with recurrent respiratory infections), pancreatic insufficiency (which leads to malabsorption and growth retardation) and elevated sweat electrolytes. It is the most common genetic disease in Caucasians, with a prevalence of about 1 in 2000 live births. Inheritance is autosomal recessive.

2.5.4. Nephrogenic Diabetes Insipidus (NDI) Nephrogenic diabetes insipidus (NDI) is a disorder known to be caused by misfolding of one hormonal protein, antidiuretic hormone, also known as vasopressin, where more than 70 different mutation have been identified [3].

Disease description [28]. Nephrogenic diabetes insipidus (also known as renal diabetes insipidus) is a form of diabetes insipidus primarily due to pathology of the kidney. This is in contrast to central/neurogenic diabetes insipidus, which is caused by insufficient levels of antidiuretic hormone

2.5.5. Retinitis Pigmentosa 4 (RP4) The disease is caused by mutations affecting the gene represented in this entry. More than 100 mutations have been identified in the misfolded protein, rhodopsin [3].

Disease description [29]. A retinal dystrophy belonging to the group of pigmentary retinopathies. Retinitis pigmentosa is characterized by retinal pigment deposits visible on fundus examination and primary loss of rod photoreceptor cells followed by secondary loss of cone photoreceptors. Patients typically have night vision blindness and loss of midperipheral visual field. As their condition progresses, they lose their far peripheral visual field and eventually central vision as well.

2.6 Data Mining

Automated data collection tools and database technology generates tremendous amounts of data and stored those in databases, data warehouses and other information repositories. There is a massive increase in the amount of data recorded and stored on digital media at present days. However, the data which has been stored needs to be converted into information and knowledge to make them more useful. So, it is important to analyse these data and discover relevant and interesting information through reliable ways. Here, come the concepts of data mining. Data mining is the entire process of applying computer based techniques, including new methods for knowledge based discovery from data [14].

Data mining is the non-trivial process of extracting interesting, implicit, valid, novel, previously unknown, potentially useful and understandable patterns or knowledge from huge amount of data. Many people treat data mining as a synonym for another popularly used term, knowledge discovery from data, or KDD, while others view data mining as merely an essential step in the process of knowledge discovery [30].Thus, data mining can be defined as “the process of discovering meaningful new correlations, patterns and trends by digging into large amounts of data stored in warehouses, using artificial

intelligence (AI) and statistical and mathematical techniques" and it may be called Knowledge Discovery in Databases [31].

The functionalities of the data mining techniques are as follows; data characterization, data discrimination, association analysis, classification, prediction, clustering, outliers and association rule mining [5].Data mining itself involves the uses of machine learning, statistics, artificial intelligence, database sets, pattern recognition and visualization [32].The major applications of data mining are in healthcare, market basket analysis, education, manufacturing engineering, customer relationship management, fraud detection, intrusion detection, customer segmentation, financial banking, corporate surveillance, research analysis, criminal investigation, bio informatics, etc.

2.7 Data Mining in Bioinformatics

Bioinformatics is defined as research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioural or health data, including those to acquire, store, organise, archive analyse, or visualise such data [33].It seems that data mining approaches ideally suited for bioinformatics, as of its data-rich. Over recent years the development of technology both computationally, medically and within biology has allowed for data to be developed and accumulated at an extraordinary rate. To interpret this data and drawing conclusions out of this data requires sophisticated computational analysis [34]. Use of data mining can be the most active technique to infer structure and principles of biological datasets and to solve biological problems. Mining biological data helps to extract useful knowledge from massive datasets gathered in biology, and in other related life sciences areas such as medicine and neuroscience [35]. Typical applications of data mining to bioinformatics include protein structure

prediction, gene classification, gene finding, disease diagnosis, disease prediction, protein and gene interaction network reconstruction, data cleansing, protein sub-cellular location prediction, analysis of mutations in cancer and gene expressions, etc.

2.8 Frequent Pattern Mining

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. For example, a set of items, such as amino acid Phenylalanine and Glycine that appear frequently together in a protein data set, is a frequent itemset. A subsequence, such as buying 1st paper, then pencil, and then eraser, if it occurs frequently in a shopping history database then it is a frequent sequential pattern. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data [36]. It also helps in data classification, clustering, indexing and other data mining tasks.

Frequent Contiguous Patterns (FCP) 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 FCP [1].Data Mining has recently increased its popularity in classifying the biological sequences and structures based on their critical features and functions [2]. Pattern mining is useful in the bioinformatics domain for predicting rules for organization of certain elements in genes, for protein function prediction, for gene expression analysis, for protein fold recognition and for motif discovery in DNA sequences [37]. To analyse, predict and manage bulk biological data, numerous computer algorithms and methods are developed. These algorithms help to compare and align biological sequences and predict bio-sequence patterns [1].In this work Apriori algorithm

is used to analyse, predict and identify the desired pattern of domination amino acids in the protein sequences.

2.9 Association Rule Mining

Frequent pattern mining provides the solution for association rules mining [1]. Mining frequent pattern and association rule is one of most important tasks of data mining [6]. Association rule mining is one shorts of pattern mining which is built from frequent itemset mining. Frequent itemset mining is defined as the set of items that appears together frequently in a database. In data mining, association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases [10].

It is mentioned earlier that proteins are the sequences of amino acids joined together in a chain. These chains of amino acids fold up in complex ways, giving each protein a unique 3D shape. Protein misfolding is believed to be the primary cause of genetic disorder diseases. Thus, the relationship between the 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 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 the solution for association rules formation among the most dominating amino acids for different protein misfolded diseases.

Association rules are *If– Then* statements that help to reveal the relationships between apparently distinct data in a relational database or other kind of information warehouse.

An association rule consists two segments: (1) an antecedent (if) and (2) a consequent (then). An antecedent is an item that is available in the data and a consequent is an item that is generated in combination with the antecedent. Association rules are developed by studying data for frequent/significant *if – then* patterns and applying the measures of support and confidence level to establish the most significant relationships.

In data mining, association rule learning is a popular and well-explored researched method for discovering interesting relations between variables in large databases [38]. Based on the concept of strong association rules, Agrawal [9] introduced association rules for learning uniformities between products of large scale transactions in supermarkets as recorded by their point-of-sale systems. For example, the rule {note book, pencil}⇒{eraser} found in the sales data of a supermarket would indicate that if a customer buys note book and pencil together, he/she is likely to also buy eraser.

In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection and bioinformatics [39].

According to the formal definition given by Rakesh Agrawal [40], the problem of association rule is defined as:

> Let $\boldsymbol{I} = \{i_1, i_2, ..., i_n\}$ be a set of *n* binary attributes called *items*. Let $\boldsymbol{D} = \{t_1, t_2, ..., t_n\}$ be a set of transactions called the *database*. Each transaction in $\boldsymbol{D}$ has a unique transaction ID and contains a subset of the items in $\boldsymbol{I}$. A *rule* is defined as an implication of the form $X \rightarrow Y$ where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are subset of $\boldsymbol{I}$ ($\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{I}$) and they have no element in common, i.e. $\boldsymbol{X} \cap \boldsymbol{Y} = \emptyset$. The sets of items (for short

itemsets) ***X*** and ***Y*** are called *antecedent* (left-hand-side) and *consequent* (right-hand-side) of the rule respectively.

To illustrate the concepts, an easy example is taken from the bioinformatics sphere consisting. The example that is considered here is very small. However, in practical application, datasets often contain thousands or millions of transactions and to obtain a statistically significant rule, it needs to support many transactions. The set of amino acid items, ***I*** = {***A***, ***M***, ***D***, ***K***, ***L***} and a small database consisting of ten transactions shown in table 2.3 where 0 represents the absence of an item and 1 the presence in a transaction. An example for a rule in this case could be {***A, D***}⇨{***L***}, which means that if ***A*** and ***D*** are present in a sequence then ***L*** may also be present in that sequence.

Table 2.3: Occurrences of Amino Acid Items in a Database Transaction

Transaction ID	**A**	**M**	**D**	**K**	**L**
t_1	1	0	1	0	0
t_2	1	0	1	0	1
t_3	0	1	0	1	1
t_4	1	0	1	0	1
t_5	1	1	0	1	0
t_6	0	1	1	0	0
t_7	1	0	1	1	1
t_8	1	0	1	0	1
t_9	1	1	1	1	1
t_{10}	1	1	1	0	1

From the set of all possible rules, the most interesting rules can be selected by using constraints on various measures of interest and significance. Some of these useful measures are support, confidence, lift and conviction.

2.9.1 Support

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

$$supp(X) = \frac{No.\ of\ transacctions \in which\ itemset\ X\ appears}{Total\ no.\ of\ transactions}$$

In the above example, total number of transactions is 10. The number of transacctions where the itemset {***A, D, L***} and {***A, D***} is appeared are 6 and 7 respectively.

$$\text{Thus, the} \quad supp(\{A, D, L\}) = \frac{6}{10} = 0.60$$

$$\text{and,} \quad supp(\{A, D\}) = \frac{7}{10} = 0.70$$

2.9.2 Confidence

The *confidence* of a rule is defined as:

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

It indicates the likelihood of item ***Y*** being appeared when item ***X*** is appeared. Thus for the rule {***A, D***} ☜ {***L***}, the confidence will be,

$$\frac{supp(\{A, D, L\})}{supp(\{A, D\})} = \frac{0.60}{0.70} = 0.857$$

This implies that for 85.7% of the transactions containing amino acid ***A*** and ***D***, the rule {***A, D***} ☜ {***L***} is correct.

Confidence can be interpreted as the conditional probability ***P(Y|X)***, the probability of finding the right-hand-side of the rule in transactions under the condition that these transactions also contain the left-hand-side [10].

2.9.3 Lift

The ***lift*** of a rule is defined as:

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

This indicates the likelihood of the itemset ***Y*** being appeared when item ***X*** is appeared while taking into account the popularity of ***Y***. If the value of lift is greater than 1, it means that the itemset ***Y*** is likely to be appeared with itemset ***X***, while a value less than 1 implies that itemset ***Y*** is unlikely to be appeared if the itemset ***X*** is appeared.

Thus the rule **{*A, D*}** ☜ **{*L*}** has the following lift:

$$\frac{supp(\{A, D, L\})}{supp(\{L\}) * supp(\{A, D\})} = \frac{0.60}{0.70 * 0.70} = 1.22$$

2.9.4 Conviction

The *conviction* of a rule can be defined as:

$$conv(X \rightarrow Y) = \frac{1 - supp(Y)}{1 - conf(X \rightarrow Y)}$$

The conviction of the rule ***X***☜***Y*** can be interpreted as the ratio of the expected frequency that ***X*** occurs without ***Y***(that is to say, the frequency that the rule makes an incorrect prediction) if ***X*** and ***Y*** were independent divided by the observed frequency of incorrect predictions [10].

Thus the rule **{*A, D*}** ☜ **{*L*}** has the following conviction:

$$\frac{1 - supp(\{L\})}{1 - conf(\{A, D\} \rightarrow \{L\})} = \frac{1 - 0.70}{1 - 0.857} = \frac{0.30}{0.143} = 2.098$$

2.10 Association Rule Mining Algorithm

There exist many algorithms that are applied for data mining. However, Apriori remains as an important algorithm as it has introduced several key ideas used in many other pattern mining algorithms thereafter [41]. Apriori algorithm proposed by Agrawal is a classical algorithm in data mining. The Apriori algorithm is a popular and foundational member of the correlation based 'Data Mining kernels' used today [42]. It is applied for mining frequent itemsets and significant association rules.

Apriori is devised to operate on databases containing lot of transactions (TIDs), for example, collections of items brought by customers or details of a website frequentation. It is also prominently applied in variety of domain like text mining, pattern mining for bio-sequences, predicting protein sequences, predicting gene organization rules, DNA sequencing, in the field of healthcare for the detection of adverse drug reaction, field of telecommunications, intrusion detection and many more.

According to Rakesh Agrawal, the Apriori principle can be written as follows [43]:

- If an itemset is frequent, then all of its subsets must also be frequent, or
- If an item set is infrequent then all its supersets must also be infrequent

Apriori principle holds due its downward-closure property of the support measure [44]:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

It means, support of an itemset never exceeds the support of its subsets. This property is also called the anti-monotone property of support.

Apriori algorithm is an influential algorithm for mining frequent itemsets for Boolean association rules [17]. Using the above properties, Apriori algorithm can efficiently generate all frequent itemsets. Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time (a step known as *candidate generation*) and groups of candidates are tested against the data [46]. The algorithm terminates when no further successful extensions are found. The basic pseudocode of Apriori algorithm as proposed by Rakesh Agrawalis as follows [40]:

2.10.1 Apriori Algorithm Pseudocode

procedure **Apriori** (T, *minSupport*)
{ //T = database and *minSupport*=Minimum Support
L_1= {frequent 1-items};
for (k= 2; $L_{k-1} \neq \emptyset$; k++)
{
C_k= Candidates generated from L_{k-1}
//that iscartesian product L_{k-1}* L_{k-1} and eliminating any k-1 size itemset that is notfrequent
for each transaction **t** in database
do
{
#increment the count of all candidates in C_k that are contained in **t**
L_k = candidates in C_k with *minSupport*
}//end for each
}//end for
return $U_k L_k$;
}

Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently [10]. From item sets of length $k - 1$, it generates candidate item sets of length k. Then the candidates which have an infrequent subset are pruned. As per the

downward closure property, the candidate set only contains all frequent item sets of length *k*. After that, the algorithm determines frequent item sets among the candidates by scanning the transaction database. The steps of the Apriori algorithm may be defined as follows:

Step-1: Scan the transaction database to get the support of *S* each 1-itemset, compares *S* with *minSupport* and get a support of 1-itemsets, L_1.

Step-2: Use L_{k-1} join L_{k-1} to generate a set of candidate *k*-itemsets and use Apriori property to prune the unfrequented *k*-itemsets from this set.

Step-3: Scan the transaction database to get the support *S* of each candidate *k*-itemset in the given set, compares *S* with *minSupport* and get a set of frequent *k*-itemsets, L_k.

Step-4: If the candidate set is not Null then go to Step2, otherwise go to Step-5

Step-5: Use the frequent item sets to generate Classification rules. For each frequent item set 1, generate all nonempty subsets of 1.

Step-6: Generate Classification rule from frequent items

$$Confidence(X \rightarrow Y) = \frac{Support_{count}(X \cup Y)}{Support_{count}(X)}$$

Step-7: For every frequent item set, X generates all non-empty subsets of X, for every non empty subset S of X, output rules

$$S \rightarrow (X - S)$$
$$if \frac{Support_{count}(X - S)}{Support_{count}(S)} \geq \min\_count$$

2.10.2 Apriori Algorithm Example

How the Apriori algorithm works and the associations rules are generated from the frequent itemsets will be explained by an example. The input will be (1) a transaction database, *D* (2) a *minSupport* as the threshold of minimum support count and (3) threshold of minimum confidence level. The output will be the (1) set of frequent itemsets and (2) association rules.

Consider a small transactional protein dataset *D*, with segmented data samples (S1, S2…., S5) of amino acids (A, M, K, L). Suppose minimum support is 3 and minimum confidence is 60%. First frequent itemsets are to be identified using Apriori algorithm. Then association rules will be generated using minimum confidence from the frequent itemsets.

SequenceID	Subsequence
S1	ALMA
S2	KLAM
S3	KKALM
S4	LLM
S5	AALM

2.10.2.1 Identification of Frequent Itemsets

Convert the attributes into binary flags. If a particular amino acid is present in sample then 1, otherwise 0.

SequenceID	A	M	K	L
S1	2	1	0	1
S2	1	1	1	1
S3	1	1	2	1
S4	0	1	0	2
S5	2	1	0	1

Scan the data set, *D* to count the frequencies, known as supports, of each member item (here, amino acid) separately.

Item	Support Count
{A}	6
{M}	5
{L}	6
{K}	3

C_1

Compare item's support count with *minSupport*

Item	Support Count
{A}	6
{M}	5
{L}	6
{K}	3

L_1

Only those items are significant for which support is greater than or equal to the threshold support. Here, the *minSupport is* 3. So, after first iteration of the algorithm, each amino acid of the 1-item candidate dataset (C_1) is frequent and thus a member of the 1-item frequent dataset (L_1).

Next the algorithm will find the frequent itemsets having 2 items. For this, the algorithm combines each frequent itemsets of size 1 (each single item) to make a set of candidate itemsets of size 2 (having 2 items). This generates two amino acids frequent patterns.

Generate **C_2**

Itemset
{A, K}
{A, M}
{A, L}
{L, M}
{L, K}
{M, K}

C_2

Scan the dataset, *D* for Support for 2 item patterns

Itemset	Support Count
{A, K}	2
{A, M}	4
{A, L}	4
{L, M}	5
{L, K}	2
{M, K}	2

C_2

Compre support with *minSupport* and eliminate infrequent itemsets

Itemset	Support Count
{A, M}	4
{A, L}	4
{L, M}	5

L_2

Here, based on the support value, the algorithm eliminates the infrequent candidate itemsets of size 2 from C_2, here, {A,K}, {L,K}, {M,K} and the frequent itemsets are left as significant itemsets. Thus 2-item frequent dataset are listed as L_2.

Similarly, as next step, the Apriori algorithm will find the frequent itemsets having 3 items. For this, the algorithm combines each frequent itemsets of size 2 to make a set of

candidate itemsets of size 3 (having 3 items). This generates three amino acids frequent patterns.

Generate C_3

Itemset
{A, L, M}
{L, M, L}
C_3

Itemset	Support Count
{A, L, M}	3
{L, M, L}	2
C_3	

Compre support with *minSupport* and eliminate infrequent itemsets

Itemset	Support Count
{A, L, M}	3
L_3	

Here, L_3 contains the only one frequent itemset of size 3. Since, no more candidate itemset can be generated; the Apriori algorithm will end here.

Now apply association rule considering our desirable confidence value.

2.10.2.2 Association Rule Generation

Now, we generate association rules for frequent item sets from L_3. For every item sets of L_3, we generate all nonempty subsets of frequent item sets.

Let, consider N= {A, L, M}, then it's all nonempty subsets are {A}, {L}, {M}, {A, M}, {A, L} and {L, M}.

Classification rule from frequent itemsets can be obtained using the following rule as mentioned earlier: $$Confidence(X \rightarrow Y) = \frac{Support_{count}(X \cup Y)}{Support_{count}(X)}$$

Considering minimum confidence threshold is 60%. The resulting association rules are shown below:

Association Rules	Confidence	Result Status
Rule-1: **A ^ M → L**	$\frac{Support_{count}(\{A, L, M\})}{Support_{count}(\{A, M\})} = \frac{3}{4} = 75$	Selected
Rule-2: **A ^ L → M**	$\frac{Support_{count}(\{A, L, M\})}{Support_{count}(\{A, L\})} = \frac{3}{4} = 75$	Selected

Rule-3: **L ^ M → A**	$\frac{Support_{count}(\{A,L,M\})}{Support_{count}(\{L,M\})}=\frac{3}{5}=60$	Selected
Rule-4: **A → L ^ M**	$\frac{Support_{count}(\{A,L,M\})}{Support_{count}(\{A\})}=\frac{3}{6}=50$	Rejected, (confidence < 60%)

As shown in the above, total four rules were developed. Since the minimum confidence threshold is 60%, rule-4 (confidence=50%) is rejected as it is below the threshold value. However, rule-1, 2 and 3 is accepted as association rules of the example, because their confidence is greater than or equal to the minimum threshold.

2.11 Interestingness Measures for Association Rules Mining

Association rules mining is an important technology in the domain of data mining and hidden knowledge discovering [47]. Association rules mining algorithm can generate a lot of association rules or patters or knowledge, but most of them have redundant information and limited resources. Thus, all of them cannot be used directly for an application. Therefore, it is significant to evaluate the interestingness (or usefulness) of the association rules before the practical use of the frequent patters as discovered using association rules mining technology.

Objective measure and subjective measure are mainly two broad kinds of measures for evaluating the interestingness/usefulness of the rules. Benefit of using objective measures is that they mainly use statistical methods and a quantitative value to determine the interestingness of rules which is reliable, easy to operate and convincing. Objective Measures are *Support*, *Confidence*, *Lift*, *Improve*, *Validity*, *Influence*, *Conviction* and *Bi-lift*, *Bi-improve*, and *Bi-confidence*, for *Lift*, *Improve* and *Confidence*, respectively etc. [48].

Objective measures include *Support*, *Confidence, Lift*, *Validity*, *Conviction* and *Improve*. Subjective interestingness involves the personality characteristics of subject (users) such

as domain knowledge and the hobbies [47. Though the definitions of *Support*, *Confidence, Lift*and*Conviction*has already been discussed but their limitation important to know.

a. Limitation of *Support* and *Confidence*. Due to subjectively selected support threshold value, many infrequent itemsets which have been discarded may have potential value. The rules are called strong association rules if the *Support* and *Confidence* are larger than the respective minimum *support* and minimum *confidence* threshold. But strong association rules are not always effective, some are not what users are interested in, and some are even misleading [47].

b. Limitation of *Lift*. Lift takes events A and B in equivalence position. According to the *Lift*, (A → B) and (B→A) are the same; that means, if we accept rule (A→B), (B→A) should be also accepted, but fact is not like this [47].

2.11.1 Improve

Literature [49] proposed a new interestingness/usefulness measure method of association rules based on the description of the defects of the traditional interestingness measurement method. This is called "*Improve.*" It means that the difference of the conditional probability ($B|A$) and the probability of "B"

$$Improve\ (A \rightarrow B) = [(B|A) - P(B)]$$

Limitation of *Improve*. [49] Firstly, how much improvement of probability can be called improvement? Secondly, the probability of former pieces' occurrence

will seriously affect *Improve* evaluation in such a way that when it is high, the *improve* value will be very small all the time.

To overcome the shortcomings of *Lift*, *Improve* and *Confidence,* literature [47] suggests following corrections to the measures:

2.10.2 Bi-lift [47]

From different researches, it has been evident that *lift* provides good evaluation results. But the problem of *lift* needs to be corrected. The higher the $li(A \rightarrow B)$ is, the better the rule $A \rightarrow B$is, while the higher the $lift(\acute{A} \rightarrow B)$ is, the worse the rule $A \rightarrow B$is. So, the correction of *Bi-lift* measure method, $li(\acute{A} \rightarrow B)$ as denominator, and $lift(A \rightarrow B)$ as numerator, namely, ratio of $lift(A \rightarrow B)$ to $lift(A \rightarrow B)$; *Bi-lift* formula is as follows:

$$Bi-lift(A \rightarrow B) = \frac{lift(A \rightarrow B)}{lift(\acute{A} \rightarrow B)}$$

$$¿ \frac{P(AB)/P(A)P(B)}{P(\acute{A}B)/P(\acute{A})P(B)}$$

$$¿ \frac{P(AB)P(\acute{A})}{P(\acute{A}B)P(A)}$$

Its value range is [0, ∞]. If *Lift* value is larger than 1, it shows that the emergence of "*A*" promotes the emergence of "*B*," and then we call them positive correlation rules. The higher the *Bi-li*$(A \rightarrow B)$ is, the better the rule $A \rightarrow B$ is.

2.11.3 Bi-improve [47]

Because of the defects of *improve*, the paper [47] put forward *Bi-improve*. Because the probability of former pieces' occurrence will seriously affect *Improve* evaluation in such a way that when it is high, the *improve* value will be very small all the time. In order to eliminate the influence, correction was given by multiplying the ratio of the occurrence

possibility of antecedent to the no occurrence probability of antecedent. *Bi-improve* formula is as follows:

$$Bi-improve(A \to B)=\frac{[P(B|A)-P(B)]*P(A)}{P(\acute{A})}$$

¿ $\frac{P(AB)-P(A)P(B)}{P(\acute{A})}$

The higher the *Bi-improve* $(A \to B)$ is, the better the rule $A \to B$ is.

2.11.4 Bi-confidence [47]

Confidence indicates that the appearance of some itemsets will lead to appearance of other itemsets. But the confidence of association rules only thinks about the occurrence possibility of "B" when "A" occurs, but not consider the relationship between "A" and "B" when "A" does not occur. So, it makes a lot of association rules mining invalid. For the above problems of association rules, the description of confidence is not perfect and not enough to show the degree of correlation between itemsets. Putting forward the concept of *Bi-confidence*, and its definition is as follows:

$$Bi-confidence(A \to B)=\frac{P(AB)}{P(A)}-\frac{P(\acute{A}B)}{P(\acute{A})}$$

$$= \frac{P(AB)-P(A)P(B)}{P(A)*[1-P(A)]}$$

The value range of Bi-confidence is [-1,1]. If the *Bi-confidence* value is greater than 0, then "A" and "B" have the positive correlation. If the *Bi-confidence* is equal to 1, then it shows that "A" and "B" in record set appear together or not. If the *Bi-confidence* is equal to 0, then "A" has no relation with "B". If the *Bi-confidence* is less than 0, then it shows that "A" and "B" have the negative correlation. The higher the *Bi-confidence* $(A \to B)$ is, the better the rule $A \to B$is.

Evaluation results of *Bi-lift*, *Bi-improve*, and *Bi-confidence* are almost the same, and their stabilities of evaluations are high [47]. Therefore, combining *Support* and *Confidence* with *Lift*, *Bi-lift*, *Bi-improve* and *Bi-confidence*, a reasonable framework for measuring the interestingness/usefulness of the rules can be developed. In this work, the procedures were followed:

- firstly, *Support* and *Confidence* threshold was used to filter out frequent set
- secondly, *Lift, Bi-lift*, *Bi-improve*, and *Bi-confidence* value were calculated
- then, according to the *Bi-lift*, *Bi-improve* and the *Bi-confidence* value, association rules were evaluated comprehensively

Actually, the final evaluation results of these three kinds of measure methods are very close and give perfect results.

CHAPTER-3 : PATTERN MINING FOR PROTEIN MISFOLDED DISEASES

3.1 Introduction

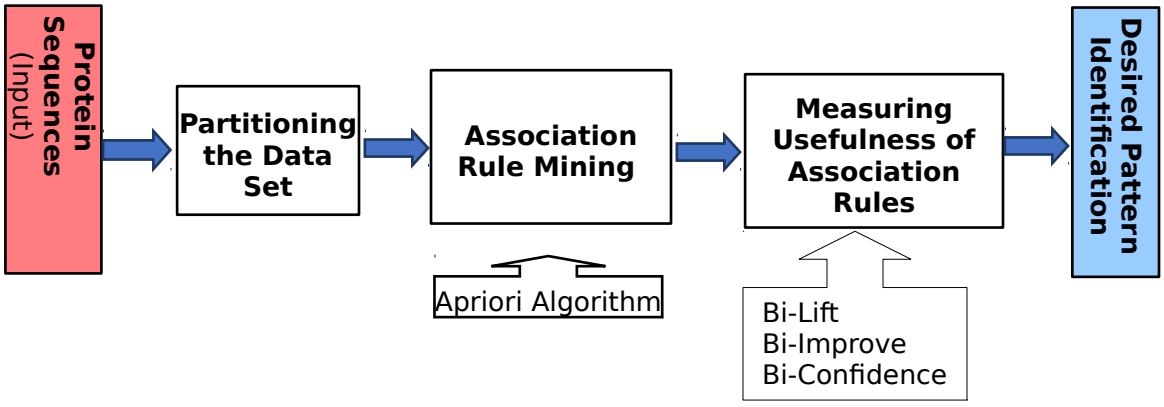
To find the frequent patterns among the most domination amino acids of protein misfolded diseases, in this work, five protein misfolded diseases (i.e. *Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa 4*) was taken in consideration. The protein sequences associated with each of the diseases were collected from protein data bank of UniProt knowledge database (http://www.uniprot.org/). In the first step, frequent itemsets were generated from the protein sequences of respected diseases. Then association rules were generated out of those frequent itemsets. In the next step, the strong association rules identified considering 90% confidence threshold. Thereafter, in the last step, using different measuring tools (*Bi-Lift*, *Bi-Improve* and *Bi-Confidence*), only the useful and interesting association rules were screen out to identify the desired patterns of dominating amino acids for the respective protein associated misfolded diseases.

3.2 Steps of the Pattern Identification

In this study, five 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 (http://www.uniprot.org/). Then the associative patterns among the amino acids were identified using a data mining technique. To generate the strong association rules from the amino acids which cause the disease, the confidence level was considered as 90% and/or above and support count was ranged between 3 to5. 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 architecture of the system is shown in the following figure:



Step-1: **Selection of Protein Sequence**. As mentioned earlier, in this work, five protein misfolded diseases (i.e. *Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa 4*) was taken in consideration. The protein sequences (amino acid chain) associated with each of the diseases was collected from a well-recognised protein data bank named Universal Protein Resource (UniProt) (http://www.uniprot.org/) in FASTA form. It is to note that the UniProt is a comprehensive resource for protein sequence and annotation data. UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), UK, the SIB Swiss Institute of Bioinformatics, Switzerland and the Protein Information Resource (PIR), USA. The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. Due to its world-wide acceptance and high degree of reliability, protein sequences were

collected from UniProt protein knowledgebase. The protein sequences (amino acid chains) for each of the concerned diseases are shown in Appendix-A.

Table 3.1 shows the human diseases and the name of the protein which is involved for the corresponding diseases. Here Breast Cancer Type 1 susceptibility protein possesses the highest length of total 1863 amino acids which is involved for Breast Cancer disease. On the other hand, the protein involved for *Sickle Cell Anemia* disease is Hemoglobin Subunit Beta which is a binding block of only 147 amino acids.

Table 3.1 :Different Human Diseases and Involved Proteins

Ser	**Disease**	**Protein Name**	**Lengths**	**Web link**
1.	Sickle Cell Anemia	Hemoglobin Subunit Beta Entry Code: P68871	147	www.uniprot.org/uniprot/P68871
2.	Breast Cancer	Breast Cancer Type 1 susceptibility protein Entry Code: P38398	1863	www.uniprot.org/uniprot/P38398
3.	Cystic Fibrosis	Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Entry Code: P13569	1480	www.uniprot.org/uniprot/P13569
4.	Nephrogenic Diabetes Insipidus (NDI)	Vasopressin V2 Receptor (V2R) Entry Code: P30518	371	www.uniprot.org/uniprot/P30518
5.	Retinitis Pigmentosa 4 (RP4)	Rhodopsin (Opsin-2) Entry Code: P08100	348	www.uniprot.org/uniprot/P08100

From different studies it has been revealed that for *Sickle Cell Anemia* and *Retinitis Pigmentosa 4 (RP4)* diseases, only Hemoglobin and Rhodopsin protein are involved respectively. On the other hand, for *Breast Cancer* disease, TP53 (Tumor protein), Phosphatase and tensin homolog (PTEN) protein, BRCA1 and BRCA2 etc. are also involved. In case of *Cystic Fibrosis* disease, Mitochondrial genes and Mucin 1 genes might be involved indirectly. Whereas, for *Nephrogenic Diabetes Insipidus (NDI)* disease, another protein known as aquaporin-2 (AQP2) is activated to serve as a passageway or water channel

through which water crosses the cell membrane. However, proteins as mentioned in table 3.1 against each disease are primarily involved for the corresponding diseases [3] and that's why this work focused to analyse these proteins to find the relation between the most dominating amino acids of the experimented diseases.

Step-2: **Partitioning the Data Set**. The FASTA form of each of the protein sequences (amino acid chain) was subdivided into amino acid subsets where each subset had the length of 10. However, where the total length of the input sequence was not divisible by 10, the reminder set of amino acids formed the last subset. For example, the length of Hemoglobin Subunit Beta protein (which was responsible for the *Sickle Cell Anemia* disease) was 147, that mean this protein sequence contained the amino acid chain of 147 lengths. Here, 20 amino acids combining with each other and formed the sequence of 147 lengths. This sequence was then partitioned into amino acid sub sequences of length 10. Thus total 14 sub sequences were of length 10 and the rest 7 amino acids combination formed the last sub sequence (Table 3.2).

Table 3.2 : Sub Sequences of Hemoglobin Subunit Beta Protein

10	20	30	40	50
MVHLTPEEKS	AVTALWGKVN	VDEVGGEALG	RLLVVYPWTQ	RFFESFGDLS
60	70	80	90	100
TPDAVMGNPK	VKAHGKKVLG	AFSDGLAHLD	NLKGTFATLS	ELHCDKLHVD
110	120	130	140	147
PENFRLLGNV	LVCVLAHHFG	KEFTPPVQAA	YQKVVAGVAN	ALAHKYH

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

The length of the protein sequences involved with the diseases is different. Thus, some sequences have very less sub sequences of length 10 and some have many. Table 3.3 highlights the number of sub sequences of each of the protein sequences.

Table 3.3 : No. of Sub Sequences of each Protein Sequences

Name of the Protein	Total Length	Length of Each Sub Sequences	No. of Sub Sequences	Length of Last Sub Sequences
Hemoglobin Subunit Beta	147	10	15	7
Breast Cancer Type 1 susceptibility protein	1863	10	187	3
Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)	1480	10	148	0
Vasopressin V2 Receptor (V2R)	371	10	38	1
Rhodopsin (Opsin-2)	348	10	35	8

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

Step-3: **Association Rule Mining**: The subsets of amino acids was then used for associative pattern identification using a mining technique. Here, Apriori Algorithm was applied for data mining. For mining frequent item sets, Apriori is treated as an influential algorithm for Boolean association rules [17]. Association rules were obtained based on predetermined minimum support count and minimum confidence level. In this work, minimum 90% confidence level and minimum support count between 3 and 5 (depending on the length of the protein sequence) was considered to obtain strong association rules. It is to mention that the value of the minimum support count is usually subjectively decided by the researchers. The higher the minimum support count, the lesser and stronger the association rules for a particular confidence level. However, if the support count is too high then many interesting association rules may be discarded. In this work, lengths of protein sequences are not uniform. Some have shorter length and some have longer. So, considering above issues, to generate and analyse a significant number of association rules, the minimum support count was also subjectively varied between 3, 4 and 5 depending on length of protein (Table 3.4).

Table 3.4 :Minimum Support Count and Confidence Level Considered for Each Protein Sequences to Obtained Association Rules

Name of the Protein	**Total Length**	**Minimum Support Count**	**Minimum Confidence Level**
Hemoglobin Subunit Beta	147	3	90%
Breast Cancer susceptibility protein	1863	5	90%
Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)	1480	5	90%
Vasopressin V2 Receptor (V2R)	371	4	90%
Rhodopsin (Opsin-2)	348	4	90%

Step-4: **Measuring Usefulness of Association Rules**. In the previous steps, association rule algorithm would generate a significant number of rules. Rules generated by this way, most of those have redundant information and thereby, all those rules cannot be useful. Thus, it is necessary to evaluate the usefulness of those rules. This evaluation may be conducted by objective or subjective measures. There are different objective measures for association rules, such as, *Lift*, *Validity*, *Conviction*, *Improve*, *Chi-Square Analysis*, *Certainty Factor*, etc. According to the evaluation results and the performance analysis for measure method, the *Lift*, *Validity and Conviction* is not effective and sometimes it even appears as essential mistake [47]. On the other hand, though *Improve* and *Chi-square analysis* do not have major fault but the stability of their evaluation is not good. However, evaluation results of *Bi-lift*, *Bi-improve*, and *Bi-confidence* are almost the same, and their stabilities of evaluations are high [47]. Considering this, improved objective measuring tools (*Bi-lift, Bi-improve* and *Bi-confidence*) were used to evaluate the association rules comprehensively. As such, *Bi-lift, Bi-improve* and *Bi-confidence* value of each of the association rules were calculated to finally prune the useful association rules.

Step-5: **Identification of Pattern**. The association among amino acid subsets is known by value of calculated confidence. The association among the amino acid subsets is strong if their calculated confidence is equal to or greater than threshold confidence (i.e 90%). 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. Hence, amino acids of those patterns which satisfied given support count, confidence and usefulness measures are responsible for causing the diseases.

3.3 Algorithm for Generating Association Rules

The algorithm used in this work takes three inputs: (i) the whole protein sequence of a particular protein misfolded disease, (ii) minimum support count and (iii) the threshold confidence level. Then the algorithm returns the Strong association rules of the most dominating amino acids for the concerned protein misfolded disease.

Input:

- ***Protein_Sequence***, Protein sequence of protein misfolded disease
- ***Support_Count***, Minimum Support Count
- ***Confidence***, Threshold Confidence

Output:

- ***Rules***,Strongassociation rules for the most dominating amino acids of protein misfolded disease

Procedure:

generate_association_rules()
1: *Dataset*= generate_subsequence_dataset(*Protein_Sequence*);
2: *L1*= find_frequent_itemset_of_length_1(*Dataset*);
3: **for**(*i*= 2; *Li*-1 ≠ Ø; i++) **do**

```
4:    Li←find_frequent_itemset(Dataset, Li-1);
5:    Rules← Ø ;
6: for( i = 2; Li ≠ Ø; i++ ) do
7:    Rules← find_association_rules(Li);
8:    M_Rules←find_association_measures(Rules);
9: return M_Rules;

generate_subsequence_dataset(Protein_Sequence)
1: Dataset←Ø;
2: len= length(Protein_Sequence);
3: for(i = 1; i<= len; i+=10 ) do
4:   if (i+9 <= len) then
5:     Dataset← Protein_Sequence.subsequence(i, i+9);
6:   else
7:     Dataset← Protein_Sequence.subsequence(i, len);
8: return Dataset;

find_frequent_itemset(Dataset, A)
1: B←Ø;
2: for( i = 1; i<length(A); i++ ) do
3:    for( j = i+1; j<= length(A); j++ ) do
4:      k=length(A[i]);
5:      if(A[i][1]=A[j][1]^A[i][2]=A[j][2]^.....^A[i][k1]=A[j][k-1])then
6:        Temporary =A[i]⋈A[j];
7:        if(is_frequent(Dataset, Temporary) ) then
8:          B← Temporary;
9: return B;

is_frequent(Dataset, Temporary):
1: count= Ø;
2: for( i = 1; i<= length(Dataset); i++ ) do
3:    if(Temporary ∈ Dataset[i]) then
4:      count=count + 1;
5: if( count>= Support_Count ) then
6:   return true;
7: else
8:   return false;

find_association_rules(L):
1: R←Ø;
2: for( i = 1; i<= length(L); i++ ) do
3:    for(j = 1; j <length(L[i]);j++) do
4:      left=L[i].subset(1, j);
5:      right=L[i].subset(j+1, length(L[i]));
6:      var= (support_count(L[i])/support_count(left))*100;
```

```
7: if(var>= Confidence ) then
8:   R←make_rules(left, right);
9: return R;

find_association_measures(Rules):
1: R← Ø ;
2: for( i = 1; i <= length(Rules); i++ ) do
3:   T.left=A =Pairs[i].left;
4:   T.right =B = Pairs[i].right;
5:   T.bi_lift =(p(AB)*p(A´))/(p(A´B)*p(A));
6:   T.bi_confidence = (p(AB)-(p(A)*p(B)))/(p(A)*(1-p(A)));
7:   T.bi_improve = (p(AB)-(p(A)*p(B)))/p(A´);
8: R←T;
9: return R;
```

The procedure starts with the method *generate_association_rules()*.

Step-1*:** In this step, the ***Dataset is generated by calling the method named *generate_subsequence_dataset*(***Protein_Sequence***). Here, ***Protein_Sequence*** is the protein sequence of protein misfolded disease. This method splits the protein sequence after each 10 elements of the given misfolded protein sequence and insert them into the ***Dataset*** and return it.

Step-2*:** In this step, $\boldsymbol{L_1}$ is generated which denotes the frequent itemset of length 1 by calling the method named *find_frequent_itemset_of_length_1*(Dataset***).

Step-3, 4*:** In this step, a loop runs until $\boldsymbol{L_{i-1}}$ becomes empty. Here, $\boldsymbol{L_i}$ denotes the $\boldsymbol{i}^{th}$ frequent itemset. $\boldsymbol{L_i}$ is generated by calling *find_frequent_itemset*(Dataset***, $\boldsymbol{L_{i-1}}$). This procedure generates the i^{th} frequent itemset from the $(\boldsymbol{i}\text{-}1)^{th}$ frequent itemset. It runs a nested loop where it takes each two item from $(\boldsymbol{i}\text{-}1)^{th}$ frequent itemset and if it matches all the protein except the last one between that two itemset, then it joins that two itemset and check if the itemset is frequent or not. If the itemset is frequent, then it insert that itemset into the $\boldsymbol{i}^{th}$

frequent itemset. After completing this procedure this method returns the i^{th} frequent itemset.

Step-6, 7*:** In this step, a loop runs until L_{i-1} becomes empty starting from L_2 and find the association rules by calling the method named *find_association_rules*(L***). In each iteration of the loop inside this method it takes an item from the i^{th} frequent itemset and splits it into two parts from first to last. Then it calculates the confidence and insert the rules having confidence above the given confidence and returns the set of rules. Finally, the association rules are stored in ***Rules***.

Step-8: In this step, a loop runs over all items of ***Rules*** by calling the method name*d* *find_association_measures*(***Rules***). Then it calculates ***bi_lift***, ***bi_confidence*** and ***bi_improve*** for each of the items of ***Rules***. Finally, the rules with metrics for association rules measuring are stored in ***R***.

3.4 Experimental Results

To conduct the experiments, the algorithm had been implemented using C++. The computations were performed in a laptop computer with an Intel Core i5-7200U CPU having a clock frequency of 2.7 GHz and 4 GB of RAM. Experimental results were obtained from each of the protein sequences (amino acid chain) which were subdivided into amino acid sub sequences of length 10. The sub sequences were treated as transaction protein subsets/datasets. During the computation, the number of iterations was not fixed. The algorithm was continued till no further successful extensions were found.

It is already mentioned that the biological sequences of five protein misfolded diseases, namely *Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa 4* were considered to be experimented to find out the most dominating amino acids and their pattern. In connection to this, five protein sequences as shown in table 3.1 were processed and examined as input to the system. The work thus follows three basic actions:

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

In doing so, following considerations were made:

a. Support count threshold between 3 and5 (depending on the length of the protein sequence) for frequent itemset generation.
b. Minimum 90% confidence level to obtain strong association rules.
c. Using Bi-lift, Bi-improve and Bi-confidence as measuring instrument to prune the useful strong association rules.

To start the process, at the beginning, the FESTA format of the protein chain sequence (e.g. *Hemoglobin Subunit Beta*) for each disease (e.g. *Sickle Cell Anemia*) was loaded as the input file. The work performed four tasks on the input: (i) dividing the protein sequence into transaction protein subsets/datasets of amino acid of length 10 (ii) generating valid frequent amino acid itemsets of minimum support count (iii) generating strong association rules

considering minimum confidence level, and (iv) prune useful strong association rules using usefulness measuring instrument.

3.4.1 Frequent Itemsets Generation

Frequent itemsets generation means the frequent amino acid sets generation from the transactional protein datasets. These itemsets were generated using the Apriori process with predetermined minimum support count between 3 and 5 as specified in table 3.4. For every corresponding protein sequences of protein misfolded diseases, frequent itemsets were generated. The generated frequent amino acid sets for the diseases can be viewed as reports as shown under each of the diseases. The Apriori process maintains list of frequent amino acid sets to further generate strong association rules.

Disease-1: Sickle Cell Anemia (Protein: Hemoglobin Subunit Beta)

For ***Sickle Cell Anemia*** disease, protein chain sequence ***Hemoglobin Subunit Beta*** was loaded in the process as input file. This protein sequence was consisted of total 147 amino acids. The sequence was subdivided into 15 transaction protein subsets/sub-sequences of amino acid of length 10. Here, 3 was considered as the minimum support count. The process garnered total 135 itemsets of amino acids which satisfied the minimum support count3. Among this, frequent 1-itemsets were16 in number, frequent 2-itemsets were 50, frequent 3-itemsets were 50, frequent 4-itemsets were 17 and frequent 5-itemsets were only 2. The process satisfies the threshold support count unto 5^{th}iteration and thus ends there. A concise list of frequent itemsets (amino acid sets) generated for this disease are shown in Fig 3.2.

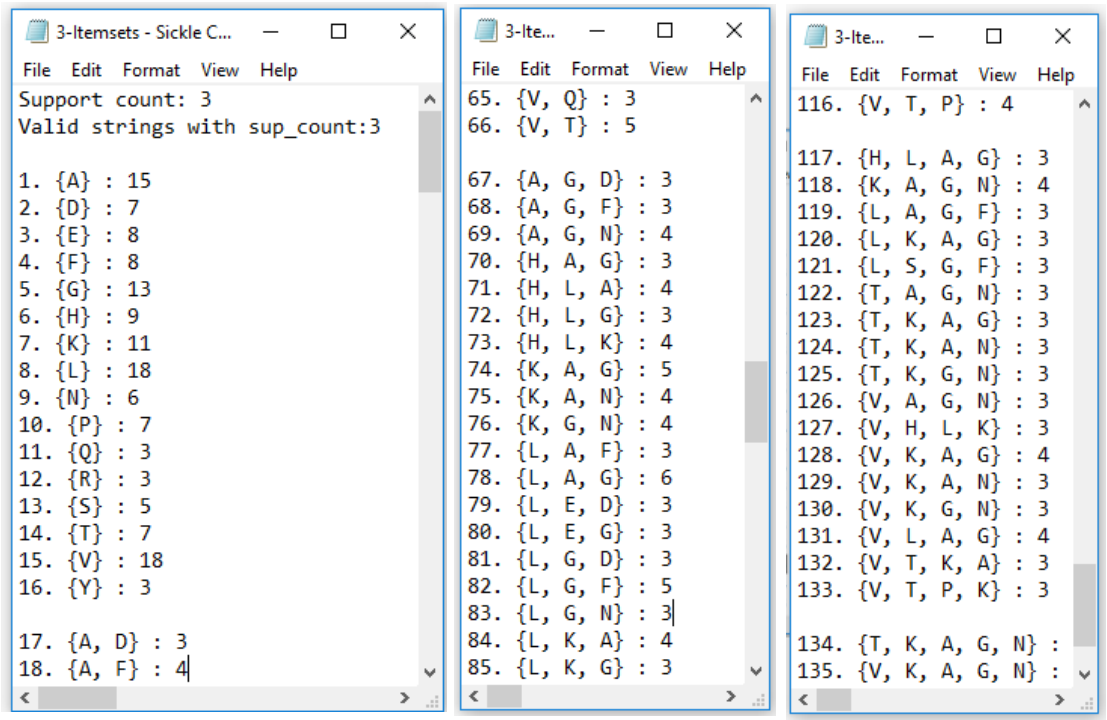


Fig-3.2: List (concise) of Frequent Itemsets (amino acid sets) obtained from Protein Sequence for Sickle Cell An

Frequent Itemsets generated from protein sequence for *Sickle Cell Anemia* disease is also graphically represented in Fig 3.3 to Fig 3.5. Fig 3.3 shows that itemset {L} and {V} are the most frequent single itemset with highest support count 18. Itemset {A}, {G}, and {K}

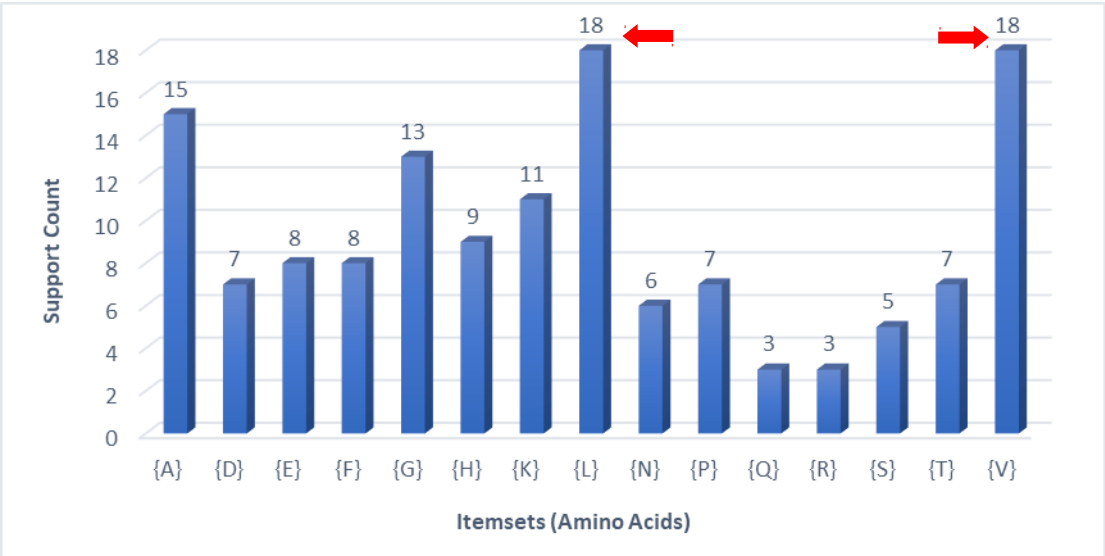


Fig 3.3: Frequent 1-itemsets (L1) obtained from Protein Sequence for Sickle Cell Anemia

are next consecutive frequent itemset with support count 15, 13 and 11 respectively. Fig 3.4

indicates that the most dominating 3-itemsets are {L, A, G} and {V, A, G} with highest support count 6. In 4thiteration, 17 frequent 4-itemsets were generated. Among those three frequent itemsets possess support count 4 which is the highest and rest 14 possess 3which is lowest (Fig 3.5). Fig 3.5 also shows frequent 5- itemsets for *Sickle Cell Anemia* which comprises only two itemsets {T, K, A, G, N} and {V, K, A, G, N} with support count 3.

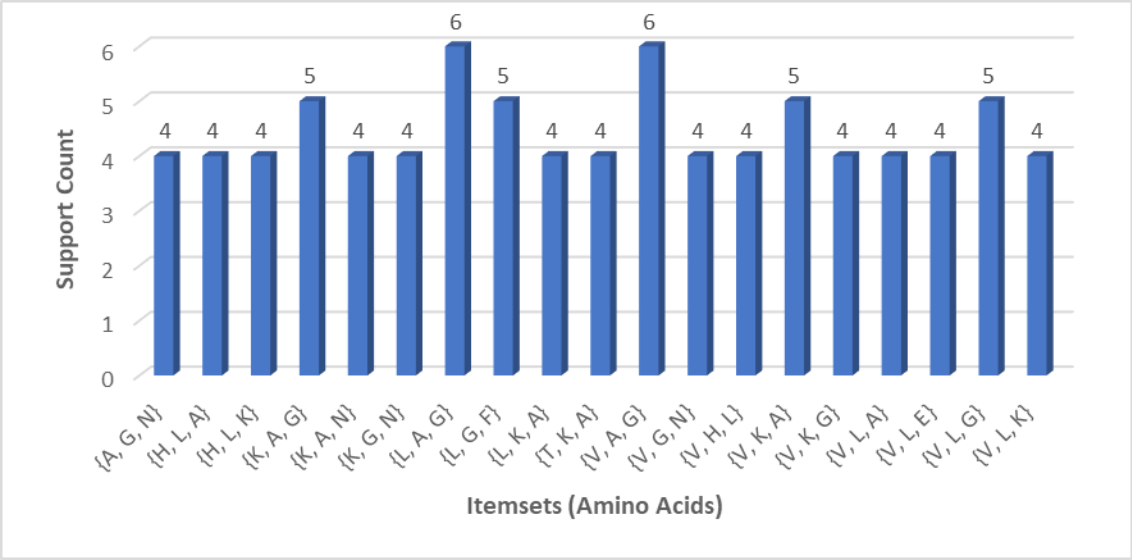


Fig 3.4: Frequent 3-itemsets (L3) obtained from Protein Sequence for Sickle Cell Anemia

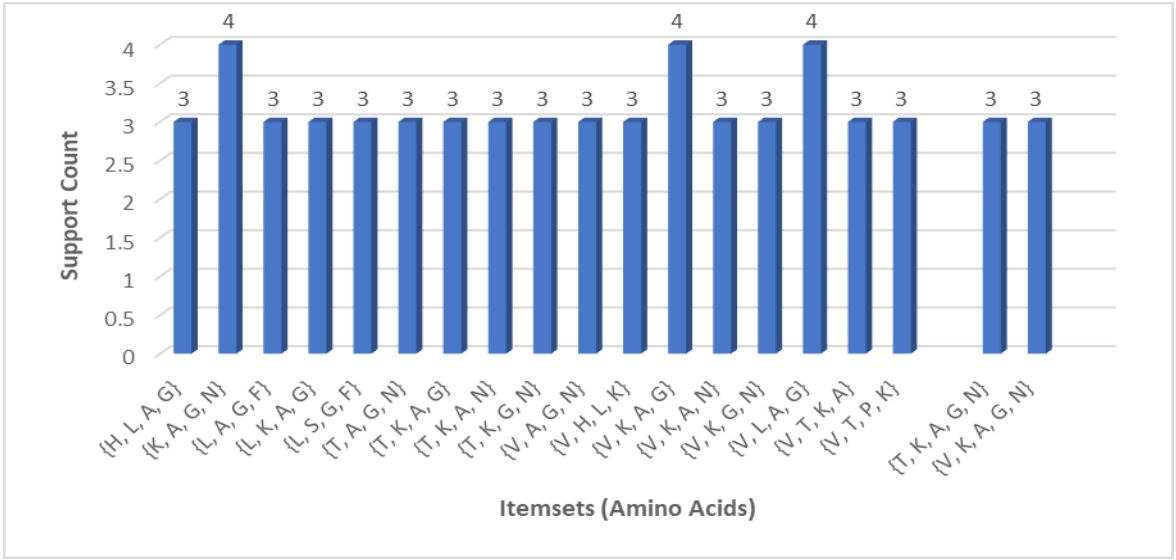


Fig 3.5: Frequent 4-itemsets (L4) and 5-itemsets (L5) for Sickle Cell Anemia

Disease-2: Breast Cancer (Protein: Breast Cancer Type 1 Susceptibility Protein)

For ***Breast Cancer*** disease, protein chain sequence ***Breast Cancer Type 1 Susceptibility Protein*** was loaded in the process as the input file. This protein chain sequence was consisted of total 1863amino acids. The sequence was subdivided into 187 transaction protein subsets/sub-sequences of amino acid of length 10. Here, due to the long length, 5 was considered as the minimum support count. The process generated total 1806 itemsets of amino acids which satisfied the minimum support count5 (full list is shown at Appendix-B). Among this, frequent 1-itemsets were 20 in number, frequent 2-itemsets were 176, frequent 3-itemsets were 669, frequent 4-itemsets were 744, frequent 5-itemsets were 191 and frequent 6-itemsets were 6 (Fig-3.6).

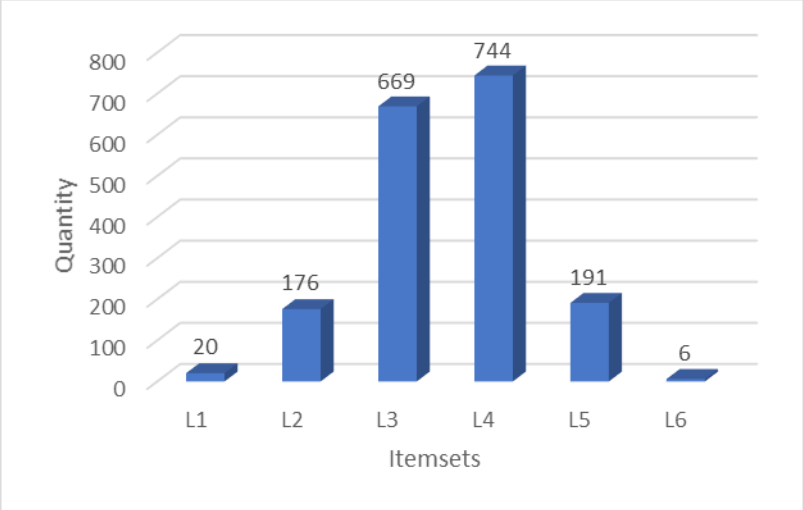


Fig 3.6: Number of Different Frequent Itemsets obtained from Protein Sequence for Breast Cancer

The process satisfies the threshold support count unto 6th iteration and thus ended there. A concise list of frequent itemsets (amino acid sets) generated for this disease are shown in Fig 3.7.

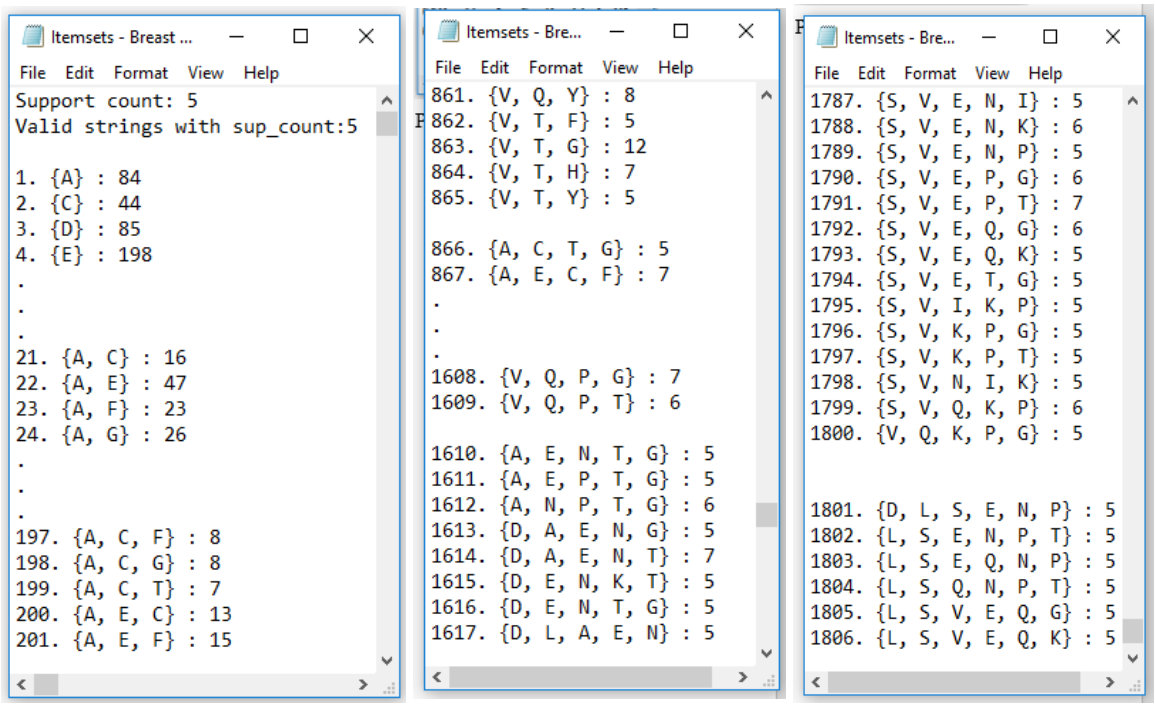


Fig-3.7: List (concise) of Frequent Itemsets (amino acid sets) obtained from Protein Fig 3.1: Architecture of t

Frequent Itemsets (L1 to L6) were generated from protein sequence of *Breast Cancer* disease and Fig 3.8and 3.9 are the graphical representation of itemsets L_5 and L_6

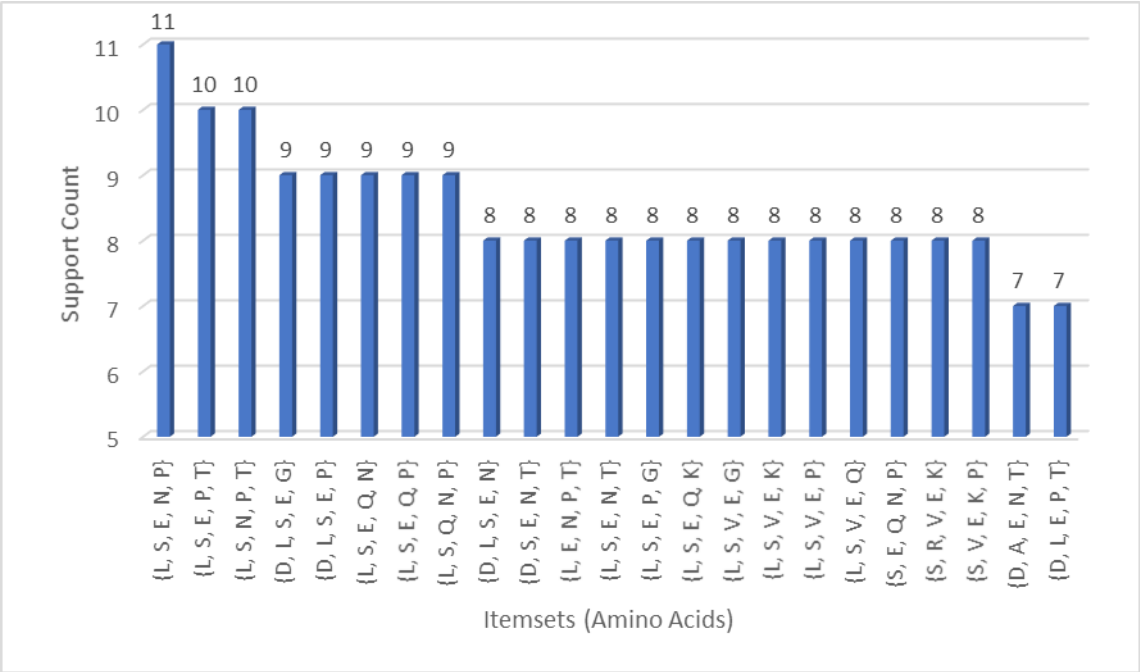


Fig 3.8: Top Frequent 5-itemsets (L5) obtained from Protein Sequence for Breast Cancer

respectively. From Fig 3.8, it is observed that the most significant 5-itemsets is {L, S, E, N, P} with highest support count 11. In this iteration, the next highest frequent 5-itemsets {L, S, E, P, T} and {L, S, N, P, T} were generated with equal support count 10.Lastly, the frequent 6-itemsets are shown in Fig 3.9. It is observed that total six itemsets are generated here with equal support count of 5.

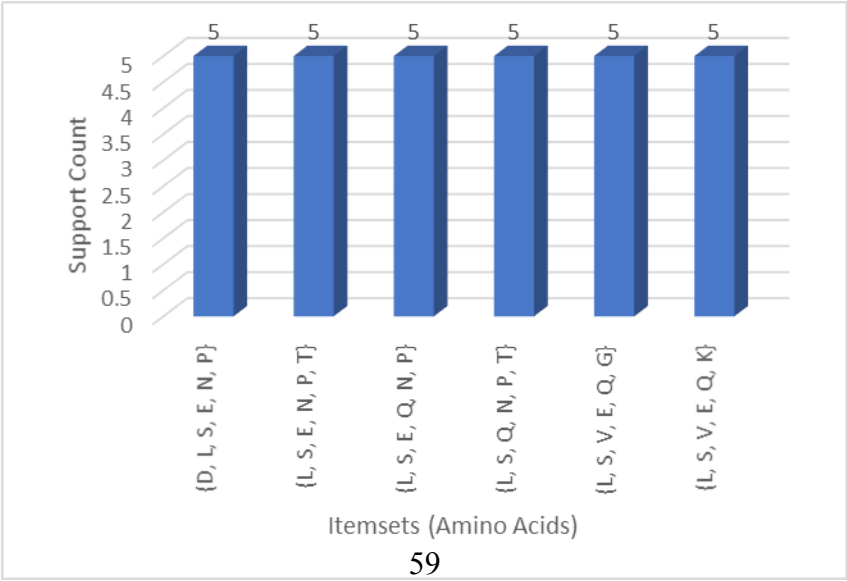


Fig 3.9: Frequent 6-itemsets (L6) obtained from Protein Sequence for Brest Cancer

Disease-3: Cystic Fibrosis (Protein: Cystic Fibrosis Transmembrane Conductance Regulator)

For ***Cystic Fibrosis*** disease, protein chain sequence ***Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)*** was loaded in the process as the input file. This protein chain sequence was consisted of total 1480 amino acids. The sequence was subdivided into 148 transaction protein subsets/sub-sequences of amino acid of length 10. Here, due to the long length, minimum support count 5 was considered. The process generated total 1464 itemsets of amino acids which satisfied the minimum support count 5 (full list is shown at Appendix-C). Among this, frequent 1-itemsets were 20 in number, frequent 2-itemsets were 178, frequent 3-itemsets were 607, frequent 4-itemsets were 563, frequent 5-itemsets were 95 and frequent 6-itemsets were only 1 (Fig 3.10).

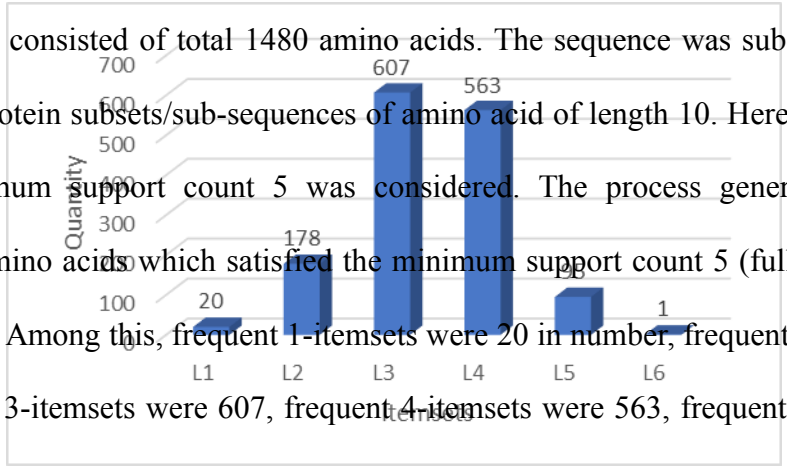


Fig 3.10: Number of Different Frequent Itemsets Generated from Protein Sequence for Cystic Fibrosis

The process satisfied the threshold support count unto 6th iteration and thus ends there. A concise list of frequent itemsets generated for this disease is shown in Fig 3.11.

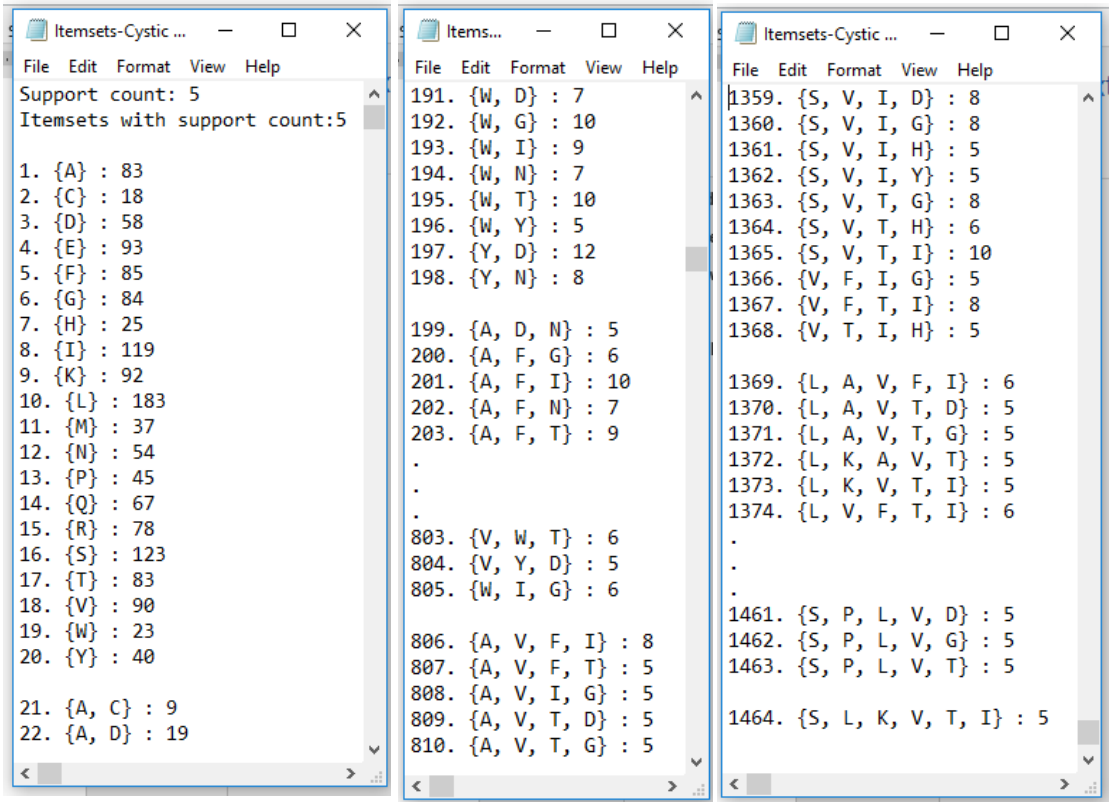


Fig-3.11: List (concise) of Frequent Itemsets (amino acid sets) obtained from Protein Sequence for Cystic Fibro

Frequent Itemsets ($L_1$to L_6) were generated from protein sequence of ***Cystic Fibrosis*** disease. It is mentioned earlier that the number of frequent 4-itemsets and frequent 5-itemsets were 563 and 95 respectively. However due to space limitation, few top frequent 4-itemsets and frequent 5-itemsets are shown in Fig 3.12 and Fig 3.13 respectively.

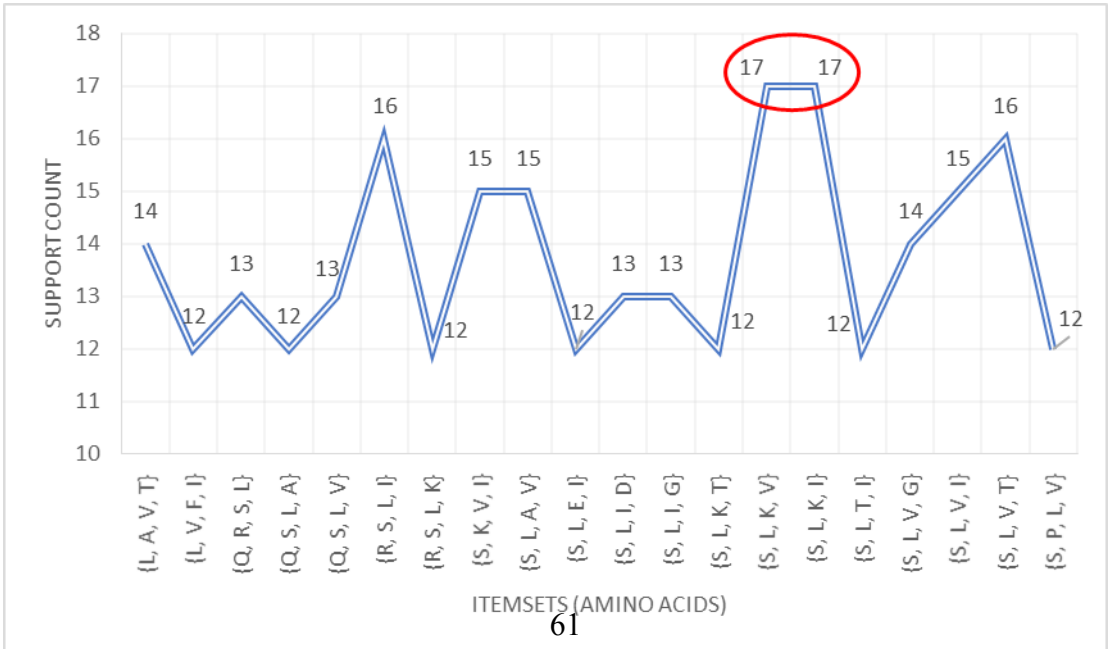


Fig 3.12: Frequent 4-itemsets (L4) obtained from Protein Sequence for Cystic Fibrosis

In fourth iteration, the highest frequent 4-itemsets {S, L, K, I} and {S, L, K, V} were generated with support count 17 (Fig 3.12). In fifth iteration, among the 95 itemset, {S, L, K, V, T} was generated as the highest frequent 5-itemset with support count 8(Fig 3.13).

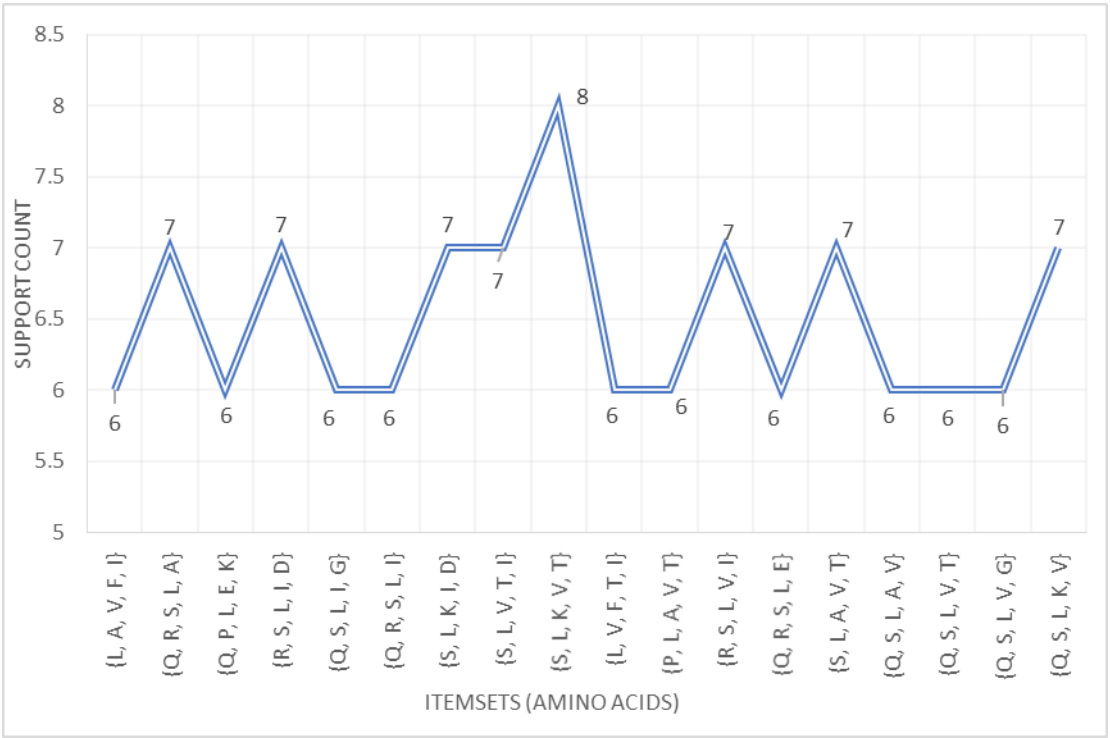


Fig 3.13: Frequent 5-itemsets (L5) obtained from Protein Sequence for Cystic Fibrosis

Disease-4: **Nephrogenic Diabetes Insipidus** ***(Protein: Vasopressin V2 Receptor)***

For ***Nephrogenic Diabetes Insipidus (NDI)*** disease, protein chain sequence ***Vasopressin V2 Receptor (V2R)*** was loaded in the process as the input file. This protein chain sequence was consisted of total 371 amino acids. The sequence was subdivided into 38 transaction protein subsets/sub-sequences of amino acid of length 10. Here, due to moderate length, minimum support count 4 was considered. The process generated total 234 itemsets of amino acids which satisfied the minimum support count 4 (full list is shown at Appendix-D). Among this, frequent 1-itemsets were 20 in number, frequent 2-itemsets were 89, frequent 3-itemsets were 99, frequent 4-itemsets were 25 and frequent 5-itemsets were only1 (Fig 3.14).The process satisfied threshold support count up to 5th iteration and thus ends there. Concise list of frequent itemsets generated for this disease are shown in fig 3.15.

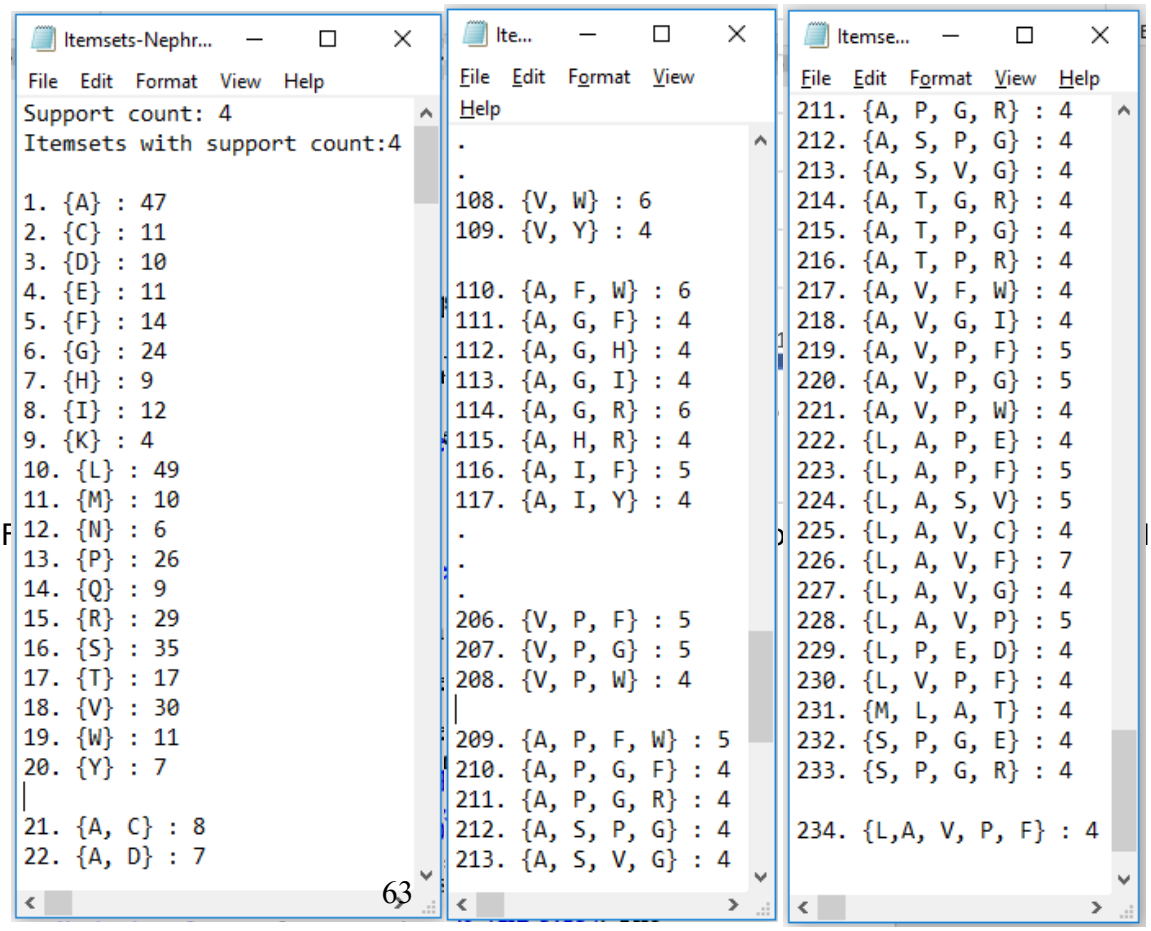


Fig 3.14: Number of Different F) Ins

Fig-3.15: List (concise) of Frequent Itemsets (amino acid sets) obtained from Protein Sequence for Nephrogeni

Fig 3.16 and Fig 3.17 are the graphical representation of frequent 3-itemsets and frequent 4-itemsetsrespectively.

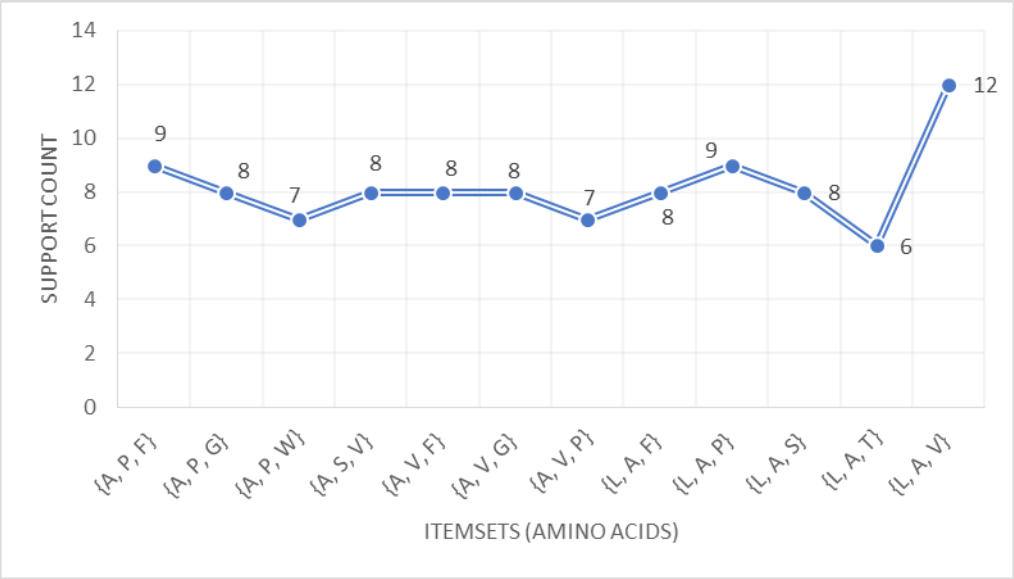


Fig 3.16: Frequent 3-itemsets (L3) obtained from Protein Sequence for Nephrogenic Diabetes Insipidus (NDI)

Frequent Itemsets ($L_1$to L_5) were generated from protein sequence of ***Nephrogenic Diabetes Insipidus (NDI)*** disease. In L_1, itemset {L} is the most frequent 1-itemset with highest support count 49 (Fig 3.15). The itemset {A} and {S} are next consecutive frequent itemset with support count 47 and 35 respectively. The most significant 2-itemsets is {L, A} with highest support count 20. It is mentioned earlier that the number of frequent 3-itemsets and frequent 4-itemsets were 99 and 25 respectively. Few top frequent 3-itemsets and frequent 4-itemsets are shown in Fig 3.16 and Fig 3.17 respectively. Itemsets {L, A, V} is generated as the highest frequent 3-itemsets having support count 12 (Fig 3.16).In fourth iteration, the highest frequent 4-itemsets {L, A, V, F} was generated with support count 7 (Fig 3.17).

Disease-5: Retinitis Pigmentosa 4 (Protein: Rhodopsin)

For ***Retinitis Pigmentosa 4 (RP4)*** disease, protein chain sequence ***Rhodopsin (Opsin-2)*** was loaded in the process as the input file. This protein chain sequence was consisted of total 348 amino acids. The sequence was subdivided into 35 transaction protein subsets/sub-

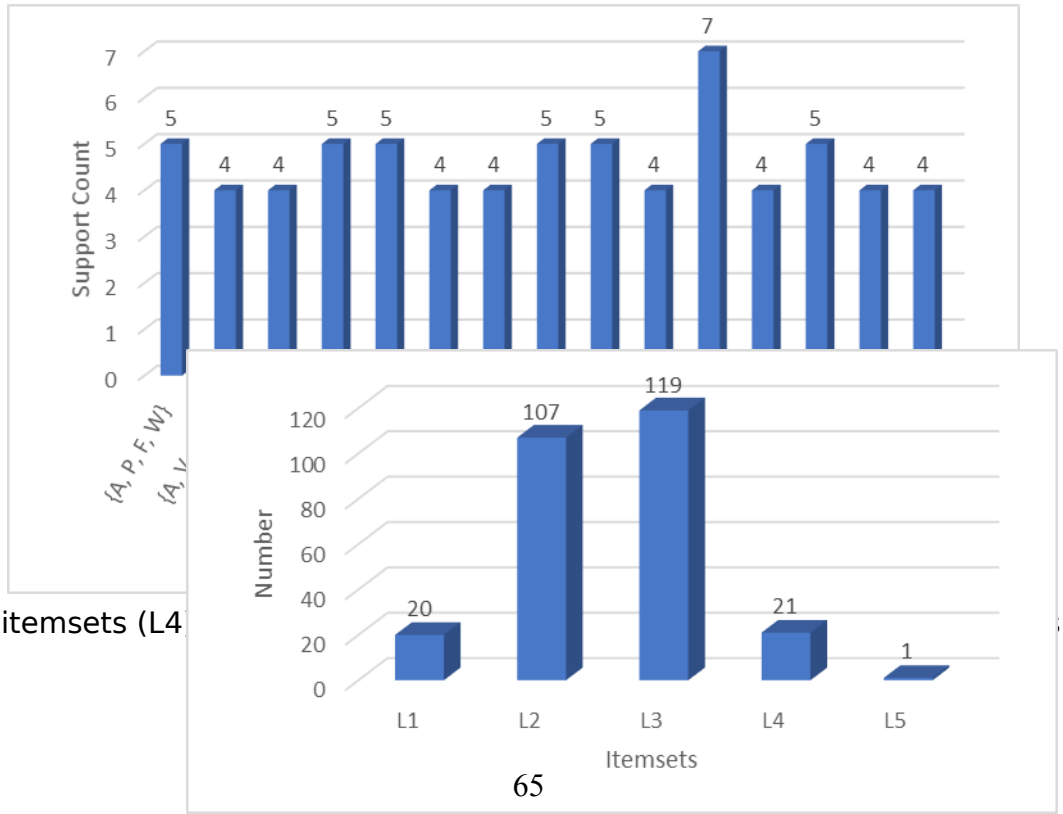


Fig 3.17: Frequent 4-itemsets (L4 Insipidus (NDI)

Fig 3.18: Number of Different Frequent Itemsets obtained from Protein Sequence for Retinitis Pigmentosa 4

sequences of amino acid of length 10. Here, due to moderate length of the protein sequence, 4 was considered as the minimum support count. The process generated total 268 itemsets of amino acids which satisfied the minimum support count 4 (full list is shown at Appendix-E). Among this, frequent 1-itemsets were 20 in number, frequent 2-itemsets were 107, frequent 3-itemsets were 119, frequent 4-itemsets were 21 and frequent 5-itemsets was only one (Fig 3.18).

The process satisfied the threshold support count unto 5th iteration and thus ends there. The concise list of frequent itemsets generated for this disease is shown in fig 3.19.

Fig 3.20 and fig 3.21 are the graphical representation of frequent 3-itemsets and frequent 4-itemsets respectively.

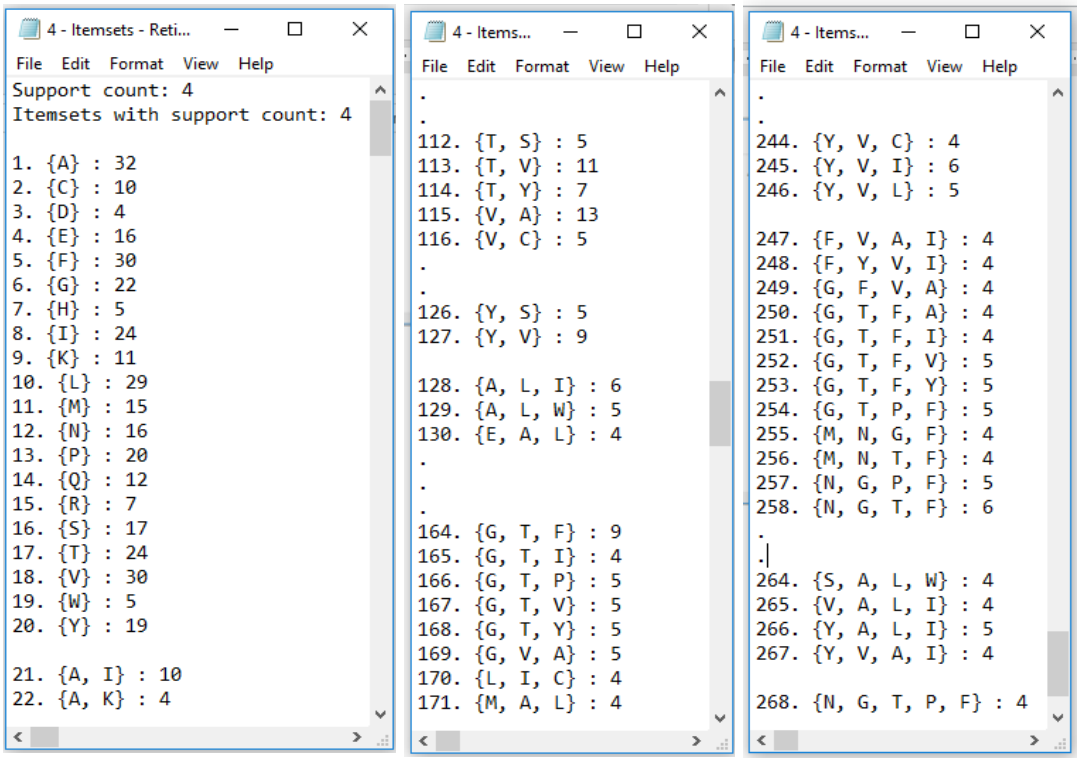


Fig-3.19: List (concise) of Frequent Itemsets (amino acid sets) Obtained from Protein Sequence for Retinitis Pig

Frequent Itemsets ($L_1$to L_5) were generated from protein sequence for Retinitis Pigmentosa 4 (RP4) disease. In L_1, itemset {A} is the most frequent 1-itemset with highest support count 32 (Fig 3.19). The itemset {F} and {V} are next highest frequent 1-itemset with support count 30. The most significant 2-itemsets is {V, A} with highest support count 13. The nest frequent 2-itemsets is {T, F} having support count 12. It is mentioned earlier that the number of frequent 3-itemsets and frequent 4-itemsets were 119 and 21 respectively.

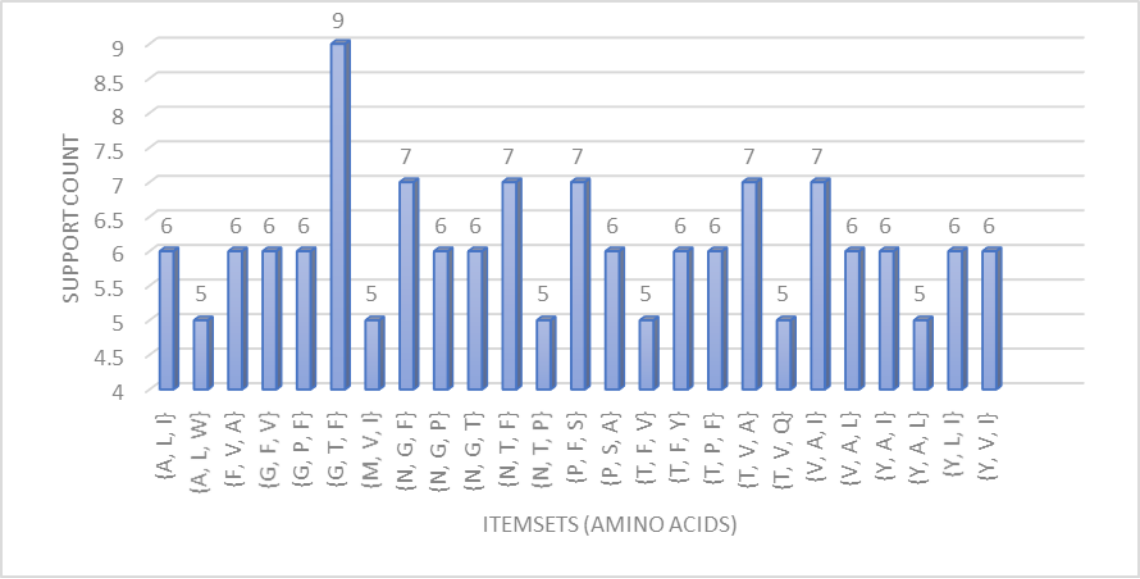


Fig 3.20: Frequent 3-itemsets (L3) obtained from Protein Sequence for Retinitis Pigmentosa 4

However due to space limitation, few top frequent 3-itemsets and frequent 4-itemsets are shown in Fig 3.20 and Fig 3.21 respectively. Itemsets {G, T,F} was generated as the highest frequent 3-itemsets having support count 9 (Fig 3.20). In fourth iteration, the highest frequent 4-itemsets {N, G, T, F} was generated with support count 6 (Fig 3.21).

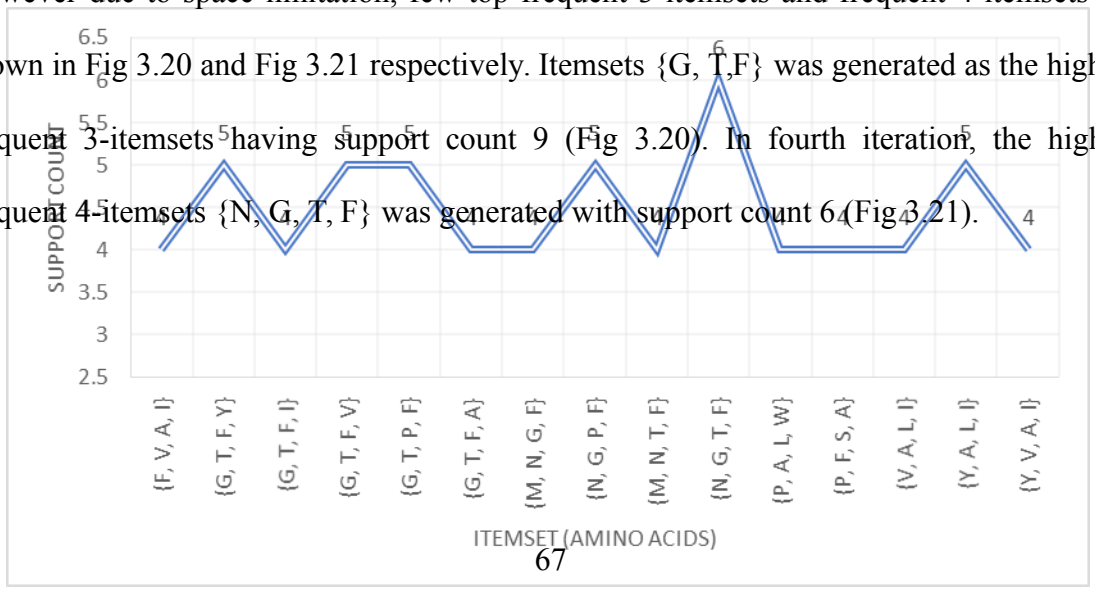


Fig 3.21: Frequent 4-itemsets (L4) obtained from Protein Sequence for Retinitis Pigmentosa 4

3.4.2 Generation of Strong Association Rules

The Apriori process maintains list of frequent itemsets (amino acid sets) and from this list strong association rules are generated. Association rules were obtained based on predetermined minimum confidence level. The association among the amino acid subsets is strong if their calculated confidence is equal to or greater than the threshold confidence (in this work, 90%). Based on the strong 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 doing so, the confidence is measured using the following equation (as mentioned in chapler-2):

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

The association rules were generated considering the above equation and following criteria:

- For each frequent itemsets, L, all nonempty subsets of L are generated
- For each nonempty subset ***S*** of L, the rule is ***S = (1 – S)***

<u>Disease-1: Sickle Cell Anemia (Protein: Hemoglobin Subunit Beta)</u>

Apriori process generates frequent itemsets. In case of ***Sickle Cell Anemia***, the Apriori algorithm handled the protein sequence of ***Hemoglobin Subunit Beta*** protein and generated total 135 frequent itemsets of amino acids. Association rules were generated from each of those frequent itemsets. Here, total 698 association rules were generated from 135 frequent itemsets. Among this 698 association rules, only 95 rules satisfied the minimum confidence level (90%) and considered as accepted strong association rules and rest 603 rules are rejected. For example, two associations rules generation measures are as follows:

a. Firstly, consider frequent itemset {*A, D*} with support count 8. Let *X*={*A, D*}. Now, all nonempty subsets of *X* are as follows:

X = {{A}, {D}}

Thus, $A \rightarrow D = \frac{Support_{Count}(\{A, D\})}{Support_{Count}(\{A\})} = \frac{3}{15} = 20$

Here, the measured confidence (20%) is less than minimum confidence threshold (90%). Thus, this association rule is *Rejected* (Table 3.5, Ser. 1).

b. Secondly, consider frequent itemset {*K, A, G*} with support count 5.

Let *Y* = {*K, A, G*}. Now, all nonempty subsets of *Y*are as follows:

Y = {{K}, {A}, {G}, {K, A}, {K, G}, {A, G}}

Thus, $GK \rightarrow A = \frac{Support_{Count}(\{K, A, G\})}{Support_{Count}(\{G, K\})} = \frac{5}{5} = 100$

Here, the measured confidence (100%) is greater than minimum confidence threshold (90%). Thus, this association rule is *Accepted* (Table 3.5, Ser. 147).

Table 3.5 shows few results and interpretation of the association rules generated from the protein sequence of *Sickle Cell Anemia*

Table 3.5 : Generation of Association Rules for *Sickle Cell Anemia*

Ser	Association Rule	Confidence	Result	Ser	Association Rule	Confidence	Result
1	A -> D	20.00%	Rejected	480	**GNT -> A**	**100.00%**	**Accepted**
2	D -> A	42.90%	Rejected	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	489	AK -> GT	42.90%	Rejected
53	L -> R	16.70%	Rejected	490	AKT -> G	75.00%	Rejected
54	**R -> L**	**100.00%**	**Accepted**	491	AT -> GK	75.00%	Rejected
55	L -> S	22.20%	Rejected	492	G -> AKT	23.10%	Rejected
.	.	.	.	493	GK -> AT	60.00%	Rejected
.	.	.	.	494	**GKT -> A**	**100.00%**	**Accepted**
146	G -> AK	38.50%	Rejected	495	**GT -> AK**	**100.00%**	**Accepted**
147	**GK -> A**	**100.00%**	**Accepted**	496	K -> AGT	27.30%	Rejected
148	K -> AG	45.50%	Rejected	497	KT -> AG	60.00%	Rejected
149	A -> KN	26.70%	Rejected	.	.	.	.
150	AK -> N	57.10%	Rejected	.	.	.	.
.	.	.	.	518	**GNT -> K**	**100.00%**	**Accepted**
.	.	.	.	519	**GT -> KN**	**100.00%**	**Accepted**
331	KV -> N	42.90%	Rejected	520	K -> GNT	27.30%	Rejected
332	N -> KV	50.00%	Rejected	521	KN -> GT	75.00%	Rejected
333	NV -> K	75.00%	Rejected	522	**KNT -> G**	**100.00%**	**Accepted**
334	V -> KN	16.70%	Rejected	.	.	.	.
335	A -> LV	26.70%	Rejected	.	.	.	.
.	.	.	.	681	AN -> GKV	75.00%	Rejected
.	.	.	.	682	**ANV -> GK**	**100.00%**	**Accepted**
459	FGL -> S	60.00%	Rejected	683	AV -> GKN	42.90%	Rejected
460	**FGS -> L**	**100.00%**	**Accepted**	684	G -> AKNV	23.10%	Rejected
461	FL -> GS	60.00%	Rejected	685	GK -> ANV	60.00%	Rejected
462	**FLS -> G**	**100.00%**	**Accepted**	.	.	.	.
463	**FS -> GL**	**100.00%**	**Accepted**	.	.	.	.
.	.	.	.	694	**KNV -> AG**	**100.00%**	**Accepted**
.	.	.	.	695	KV -> AGN	42.90%	Rejected
477	AT -> GN	75.00%	Rejected	696	N -> AGKV	50.00%	Rejected
478	G -> ANT	23.10%	Rejected	697	NV -> AGK	75.00%	Rejected
479	GN -> AT	60.00%	Rejected	698	V -> AGKN	16.70%	Rejected

Disease-2: Breast Cancer (Protein: Breast Cancer Type 1 Susceptibility Protein)

In case of ***Breast Cancer***, the Apriori algorithm handled the protein sequence of ***Breast Cancer Type 1 Susceptibility*** protein and generated total 1806 frequent itemsets of amino acids considering minimum support count 5. Association rules were generated from each of those frequent itemsets. Here, total 20,884 association rules were generated from 1806 frequent itemsets. Among these association rules, only 80 rules satisfied the minimum confidence level (90%). Hence, these rules are considered as accepted strong association rules and rest rules are rejected. Few of these accepted rules are shown in Table 3.6 (full list is at Appendix-F)

Table 3.6 : Accepted Strong Association Rules for *Breast Cancer*

Ser	Association Rule	Confidence		Ser	Association Rule	Confidence
1	AD →E	100.00%		56	GKLN →P	100.00%
2	DH →E	90.00%		.	.	.
3	MS →E	93.30%		.	.	.
.	.	.		62	GQRS →L	100.00%
.	.	.		63	NQRS →L	100.00%
25	DRS →E	90.90%		64	LRSV →E	100.00%
26	DSV →E	100.00%		65	EKQV →L	100.00%
27	DNV →E	100.00%		.	.	.
28	FLN →P	100.00%		.	.	.
.	.	.		.	.	.
.	.	.		78	LNQST →P	100.00%
40	IKR →S	100.00%		79	EGLQV →S	100.00%
41	FKV→S	90.00%		80	EKQSV →L	100.00%

Disease-3: Cystic Fibrosis (Cystic Fibrosis Transmembrane Conductance Regulator)

In case of ***Cystic Fibrosis***, the Apriori algorithm handled the protein sequence of ***Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)*** protein and generated total 1464 frequent itemsets of amino acids considering minimum support count 5. Association

rules were generated from each of those frequent itemsets. Here, total 14,792 association rules were generated from 1464 frequent itemsets. Among these association rules, only 96 rules satisfied the minimum confidence level (90%). Hence, these rules are considered as accepted strong association rules and rest rules are rejected. Few of these accepted rules are shown in Table 3.7 (full list is at Appendix-G)

Table 3.7 : Accepted Strong Association Rules for *Cystic Fibrosis*

Ser	Association Rule	Confidence		Ser	Association Rule	Confidence
1	AG →L	90.00%		70	EKPQ →L	100.00%
2	DT →L	91.70%		71	LPQR → K	100.00%
3	HV →L	90.90%		72	ALQR → S	100.00%
4	NW →L	100.00%		.	.	.
5	TW → L	90.00%		.	.	.
6	AM → L	92.90%		82	HIKV → S	100.00%
7	PY → L	100.00%		83	HISV → K	100.00%
8	QY → L	91.70%		84	AGIS → L	100.00%
.	.	.		.	.	.
.	.	.		.	.	.
24	DTV → L	100.00%		94	APSV → L	100.00%
25	HTV → L	100.00%		95	LPST → V	100.00%
26	AIM → L	100.00%		96	IKLTV → S	100.00%

Disease-4: Nephrogenic Diabetes Insipidus (Vasopressin V2 Receptor)

In case of ***Nephrogenic Diabetes Insipidus (NDI)***, the Apriori algorithm handled the protein sequence of ***Vasopressin V2 Receptor (V2R)*** protein and generated total 234 frequent itemsets of amino acids considering minimum support count 4. Association rules were generated from each of those frequent itemsets. Here, total 1152 association rules were generated from 234 frequent itemsets. Among these association rules, only 54 rules satisfied the minimum confidence level (90%). Hence, these rules are considered as accepted strong association rules and rest rules are rejected. Few of these accepted rules are shown in Table 3.8 (full list is at Appendix-H)

Table 3.8 : Accepted Strong Association Rules for *Nephrogenic Diabetes Insipidus*

Ser	Association Rule	Confidence		Ser	Association Rule	Confidence
1	K → A	100.00%		32	AFG → P	100.00%
2	N → S	100.00%		33	FG → AP	100.00%
3	FW → A	100.00%		.	.	.
4	FG → A	100.00%		40	FPV → A	100.00%
5	GI → A	100.00%		41	GPV → A	100.00%
.	.	.		42	PVW → A	100.00%
.	.	.		43	AEL → P	100.00%
16	CV → A	100.00%		.	.	.
17	FV → A	100.00%		.	.	.
18	HV → A	100.00%		49	GLV → A	100.00%
19	PV → A	100.00%		50	LPV → A	100.00%
.	.	.		51	DEL → P	100.00%
28	DE → P	100.00%		52	DLP → E	100.00%
29	FG → P	100.00%		53	AMT → L	100.00%
30	GI → V	100.00%		54	FLPV → A	100.00%

Disease-5: Retinitis Pigmentosa 4 (Rhodopsin)

In case of ***Retinitis Pigmentosa 4***, the algorithm handled the protein sequence of ***Rhodopsin (Opsin-2)*** protein and generated total 268 frequent itemsets of amino acids considering minimum support count 4. Association rules were generated from each of those frequent itemsets. Here, total 1252 association rules were generated from 268 frequent itemsets. Among these, only 49 rules satisfied the minimum confidence level (90%). Hence, these rules are considered as accepted strong association rules and rest rules are rejected. Few of these accepted rules are shown in Table 3.9 (full list is at Appendix-I)

Table 3.9 : Accepted Strong Association Rules for *Retinitis Pigmentosa 4*

Ser	Association Rule	Confidence		Ser	Association Rule	Confidence
1.	W →A	100.00%		26.	GIT →F	100.00%
2.	W →L	100.00%		27.	FTV →G	100.00%
3.	H →T	100.00%		28.	GTV →F	100.00%
4.	AW →L	100.00%		29.	FGY →T	100.00%
5.	LW -> A	100.00%		30.	GTY →F	100.00%
.	.	.		31.	GPT →F	100.00%

.	.	.		.	.	.
12.	GM →F	100.00%		.	.	.
13.	NY →P	100.00%		41.	ALS →W	100.00%
14.	PW →A	100.00%		42.	ASW →L	100.00%
15.	PW →L	100.00%		43.	LSW →A	100.00%
16.	SW →A	100.00%		44.	SW →AL	100.00%
.	.	.		45.	ILV →A	100.00%
.	.	.		46.	ALY →I	100.00%
22.	CY →V	100.00%		47.	AVY →I	100.00%
23.	AFT →G	100.00%		48.	FNPT →G	100.00%
24.	AGT →F	100.00%		49.	GNPT →F	100.00%

3.4.3 Identification of Usefulness of Association Rules

The association rules obtained by minimum support count and minimum confidence threshold, are called strong association rules. But interestingly all strong association are not always effective [13]. It might be that some rules are not what the users are interested in. On the other hand, some rules might be misleading. Therefore, the measuring of interestingness or usefulness of the strong association rule are important. As mentioned earlier, improved objective measuring tools (*Bi-lift, Bi-improve* and *Bi-confidence*) were used to evaluate the association rules comprehensively. As such, *Bi-lift, Bi-improve* and *Bi-confidence* value of each of the association rules were calculated to finally prune the useful association rules.

Lift, Bi-lift, Bi-improve and *Bi-confidence* value of each of the association rules were calculated and the rules were pruned based on the following criteria:

- The rule ($A \rightarrow B$) will be considered as positively correlated rule (emergence of "A" promotes the emergence of "B,") if its *Lift* value is greater than 1. Thus, those rules are useful only whose *Lift* value is greater than 1. The higher the ($A \rightarrow B$)is, the better the rule ($A \rightarrow B$) is, while the higher the ($\bar{A} \rightarrow B$) is, the worse the rule ($A \rightarrow B$) is.

- The higher the *Bi-*($A \rightarrow B$) is, the better the rule ($A \rightarrow B$) is.
- The higher the *Bi-improv*($A \rightarrow B$) is, the better the rule ($A \rightarrow B$) is.
- If the *Bi-confidence* value is greater than 0, then $P(AB)>P(A)P(B)$, which shows that "*A*" and "*B*" have the positive correlation. Thus, those rules are useful only whose *Bi-confidence* value is greater than 0. The higher the *Bi-confidence* ($A \rightarrow B$) is, the better the rule $A \rightarrow B$ is.

Disease-1: Sickle Cell Anemia (Protein: Hemoglobin Subunit Beta)

In case of ***Sickle Cell Anemia***, the Apriori algorithm handled the protein sequence of ***Hemoglobin Subunit Beta*** protein and total 698 association rules were generated. Among this 698 association rules, only 95 rules satisfied the minimum confidence level (90%) and considered as accepted strong association rules. Now these 95 rules were further evaluated to determine their usefulness. In doing so, *Lift, Bi-lift, Bi-improve* and *Bi-confidence* values of each of these association rules were calculated and the rules were pruned based on the criteria stated in the earlier paragraph.

From Table 3.10, it is evident that 59 rules satisfy the *Lift, Bi-lift, Bi-improve* and *Bi-confidence* value and sorted as positive strong association rules (i.e. these 59 rules are useful or effective rules). Rest 36 rules are redundant or might be misleading and thus not effective. Full list of useful strong association rules are shown at Appendix-J.

Table 3.10 :Usefulness Measures of Association Rules for *Sickle Cell Anemia*

Ser	Rules	Lift	Bi-lift	Bi-Improve	Bi-confidence	
1	GT →AN	3.75	12	0.183	0.917	Useful Strong
2	GT →KN	3.75	12	0.183	0.917	
3	AGT →KN	3.75	12	0.183	0.917	
4	GKT →AN	3.75	12	0.183	0.917	
5	GT →AKN	3.75	12	0.183	0.917	

6	AN →GK	3	11	0.242	0.909	Association Rules
7	GS →FL	3	6	0.167	0.833	
.	.	.	.	.	.	
.	.	.	.	.	.	
41	AGNV →K	1.364	1.5	0.067	0.333	
42	FL →G	1.154	1.25	0.067	0.2	
43	AN →G	1.154	1.222	0.048	0.182	
44	KN →G	1.154	1.222	0.048	0.182	
.	.	.	.	.	.	
.	.	.	.	.	.	
58	AKNT →G	1.154	1.2	0.033	0.167	
59	AKNV →G	1.154	1.2	0.033	0.167	
60	GH →A	1	1	0	0	Redundant Rules
61	GK →A	1	1	0	0	
62	KN →A	1	1	0	0	
.	.	.	.	.	.	
.	.	.	.	.	.	
94	PT →V	0.833	0.786	-0.073	-0.273	
95	FG →L	0.833	0.769	-0.1	-0.3	

In the above protein misfolded disease, the first accepted useful association rule is GT →AN because it satisfies the required criteria as shown below:

Criteria-1: *Lift* value should be greater than 1.

Test: Here, *lift* (GT →AN) = 3.75, which is greater than 1. So, criteria-1 satisfies.

Criteria-2: The higher the *Bi-*($A \rightarrow B$) is, the better the rule ($A \rightarrow B$) is.

Test: Here, *Bi-lift* (GT →AN) = 12, which is a positive higher value. So, criteria-2 satisfies.

Criteria-3: The higher the *Bi-improv*($A \rightarrow B$) is, the better the rule ($A \rightarrow B$) is.

Test: Here, *Bi-improve* (GT →AN) = 0.183,which is positive value. So, criteria-3 satisfies.

Criteria-4: *Bi-confidence* value is greater than 0.

Test: Here, *Bi-confidence* (GT→AN) = 0.917, which is greater than 0. So, criteria-4 satisfies.

Disease-2: Breast Cancer (Protein: Breast Cancer Type 1 Susceptibility Protein)

In case of ***Breast Cancer***, the Apriori algorithm handled the protein sequence of ***Breast Cancer Type 1 Susceptibility*** protein and 20,884 association rules were generated. Among these association rules, only 80 rules satisfied the minimum confidence level (90%) and hence, considered as accepted strong association rules. To determine their effectiveness, these 80 rules were further evaluated by corresponding *Lift, Bi-lift, Bi-improve* and *Bi-confidence* values.

From Table 3.11, it is evident that 19 rules satisfy the *Lift, Bi-lift, Bi-improve* and *Bi-confidence* values and sorted as positive strong association rules (i.e. these 19 rules are useful or effective rules). Rest 61rules are redundant or might be misleading and thus not effective (details in Appendix-K).

Table 3.11 :Usefulness Measures of Association Rules for *Breast Cancer*

Ser	Rules	Lift	Bi-lift	Bi-Improve	Bi-confidence	
1	ANPT →G	2.149	2.235	0.018	0.552	Useful Strong Association Rules
2	NQST →P	1.948	2.011	0.016	0.503	
3	FLN →P	1.948	2.0	0.013	0.5	
4	GKLN →P	1.948	2.0	0.013	0.5	
5	GLNT →P	1.948	2.0	0.013	0.5	
6	LNQST →P	1.948	2.0	0.013	0.5	
7	ILQS →N	1.545	1.569	0.01	0.363	
8	IPSV →K	1.365	1.379	0.007	0.275	
9	EKQV →L	1.199	1.208	0.006	0.172	
10	DHP →L	1.199	1.207	0.005	0.171	
11	QRT →L	1.199	1.207	0.005	0.171	
12	GPST →L	1.199	1.207	0.005	0.171	
13	GQRS →L	1.199	1.207	0.005	0.171	
14	NQRS →L	1.199	1.207	0.005	0.171	
15	DPY →L	1.199	1.205	0.005	0.17	
16	DEHP →L	1.199	1.205	0.005	0.17	
17	FPST →L	1.199	1.205	0.005	0.17	
18	EKQSV →L	1.199	1.205	0.005	0.17	

19	NQR →L	1.079	1.084	0.004	0.069	
20	ADR →E	0.944	0.943	-0.002	-0.06	Redundant Rules
.	.	.	.	.	.	
.	.	.	.	.	.	
78	EGKV →S	0.835	0.829	-0.008	-0.206	
79	EQR →S	0.751	0.741	-0.017	-0.315	
80	FKV →S	0.751	0.741	-0.017	-0.315	

Disease-3: Cystic Fibrosis (Cystic Fibrosis Transmembrane Conductance Regulator)

In case of ***Cystic Fibrosis***, the Apriori algorithm handled the protein sequence of ***Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)*** protein and generated total 14,792 association rules. Among these association rules, only 96 rules satisfied the minimum confidence level (90%) and hence, considered as accepted strong association rules. To determine their effectiveness, these 96 rules were further evaluated by corresponding *Lift, Bi-lift, Bi-improve* and *Bi-confidence* values.

From Table 3.12, it is evident that 35 rules satisfy the *Lift, Bi-lift, Bi-improve* and *Bi-confidence* values and sorted as positive strong association rules (i.e. these 35 rules are useful or effective rules). Rest 61 rules are redundant or might be misleading and thus not effective. Full list of useful strong association rules are shown at Appendix-L.

Table 3.12 :Usefulness Measures of Association Rules for *Cystic Fibrosis*

Ser	Rules	Lift	Bi-lift	Bi-Improve	Bi-confidence	
1	EKLP →Q	2.209	2.328	0.023	0.57	Useful Strong Association
2	PVW →A	1.783	1.833	0.015	0.455	
3	CLR →A	1.783	1.833	0.015	0.455	
4	HILV →T	1.783	1.833	0.015	0.455	
5	HILS →T	1.783	1.833	0.015	0.455	
6	FPR →V	1.644	1.707	0.022	0.414	
7	FIPR →V	1.644	1.69	0.017	0.408	
8	APW →V	1.644	1.682	0.014	0.406	
.	.	.	.	.	.	

.	.	.	.	.	.	Rules
31	HIKV →S	1.203	1.212	0.006	0.175	
32	HKLV →S	1.203	1.212	0.006	0.175	
33	IKLTV →S	1.203	1.212	0.006	0.175	
34	IKLV →S	1.094	1.102	0.006	0.084	
35	DIR →S	1.083	1.089	0.005	0.074	
36	ANW →L	0.809	0.803	-0.008	-0.245	Redundant Rules
37	DET →L	0.809	0.803	-0.008	-0.245	
.	.	.	.	.	.	
.	.	.	.	.	.	
94	EQR →L	0.728	0.714	-0.024	-0.361	
95	APS →L	0.728	0.714	-0.024	-0.361	
96	AG →L	0.728	0.698	-0.053	-0.389	

Disease-4: Nephrogenic Diabetes Insipidus (Vasopressin V2 Receptor)

In case of ***Nephrogenic Diabetes Insipidus (NDI)***, the Apriori algorithm handled the protein sequence of ***Vasopressin V2 Receptor (V2R)*** protein and generated total 1152 association rules. Among these association rules, only 54 rules satisfied the minimum confidence level (90%). Hence, these rules are considered as accepted strong association rules. To determine their effectiveness, these 54 rules were further evaluated by corresponding *Lift, Bi-lift, Bi-improve* and *Bi-confidence* values.

From Table 3.13, it is evident that 14 rules satisfy the *Lift, Bi-lift, Bi-improve* and *Bi-confidence* values and sorted as positive strong association rules (i.e. these 14 ruses are useful or effective rules). Rest 40 rules are redundant or might be misleading and thus not effective (details are shown at Appendix-M).

Table 3.13 :Usefulness Measures of Association Rules for *Nephrogenic Diabetes Insipidus*

Ser	Rules	Lift	Bi-lift	Bi-Improve	Bi-confidence	
1	DLP →E	3.455	4.857	0.084	0.794	Usefu
2	FG →AP	2.375	2.833	0.068	0.647	

3	GI →AV	2.235	2.615	0.065	0.618	1 Strong Association Rules
4	CV →AL	1.9	2.125	0.056	0.529	
5	AE →P	1.462	1.6	0.059	0.375	
6	DE →P	1.462	1.571	0.048	0.364	
7	FG →P	1.462	1.545	0.037	0.353	
8	AFG →P	1.462	1.545	0.037	0.353	
9	AEL →P	1.462	1.545	0.037	0.353	
10	DEL →P	1.462	1.545	0.037	0.353	
11	GI →V	1.267	1.308	0.025	0.235	
12	AGI →V	1.267	1.308	0.025	0.235	
13	N →S	1.086	1.103	0.015	0.094	
14	AN →S	1.086	1.097	0.009	0.088	
15	K →A	0.809	0.791	-0.028	-0.265	Redundant Rules
16	FG →A	0.809	0.791	-0.028	-0.265	
17	GI →A	0.809	0.791	-0.028	-0.265	
.	.	.	.	.	.	
.	.	.	.	.	.	
53	PQ →L	0.776	0.75	-0.044	-0.333	
54	MT →L	0.776	0.75	-0.044	-0.333	

Disease-5: Retinitis Pigmentosa 4 (Rhodopsin)

In case of ***Retinitis Pigmentosa 4***, the Apriori algorithm handled the protein sequence of ***Rhodopsin (Opsin-2)*** protein and total 1252 association rules were generated. Among these association rules, only 49 rules satisfied the minimum confidence level (90%). Hence, these rules are considered as strong association rules. To determine their effectiveness, these 49 rules were further evaluated by corresponding *Lift, Bi-lift, Bi-improve* and *Bi-confidence* values.

From Table 3.14, it is evident that al 49 rules satisfy the *Lift, Bi-lift, Bi-improve* and *Bi-confidence* values and sorted as positive strong association rules (i.e. these rules are useful or effective rules). Full list of useful strong association rules are shown at Appendix-N.

Table 3.14 : Usefulness Measures of Association Rules for *Retinitis Pigmentosa 4*

Ser	Rules	Lift	Bi-lift	Bi-Improve	Bi-confidence	
1	ALS -> W	7	31	0.111	0.968	
2	W -> AL	3.5	6	0.119	0.833	
3	PW -> AL	3.5	5.167	0.092	0.806	
4	SW -> AL	3.5	5.167	0.092	0.806	
5	QS -> E	2.188	2.727	0.09	0.633	
6	AFP -> S	2.059	2.385	0.066	0.581	
.	.	.	.	.	.	
.	.	.	.	.	.	
21	AVY -> I	1.458	1.55	0.041	0.355	
22	W -> L	1.207	1.25	0.029	0.2	
23	AW -> L	1.207	1.25	0.029	0.2	
24	CI -> L	1.207	1.24	0.022	0.194	
25	PW -> L	1.207	1.24	0.022	0.194	
.	.	.	.	.	.	
.	.	.	.	.	.	
34	GPT -> F	1.167	1.2	0.024	0.167	
35	EM -> F	1.167	1.192	0.018	0.161	
36	MS -> F	1.167	1.192	0.018	0.161	
37	CY -> V	1.167	1.192	0.018	0.161	
.	.	.	.	.	.	
.	.	.	.	.	.	
48	LSW -> A	1.094	1.107	0.011	0.097	
49	ILV -> A	1.094	1.107	0.011	0.097	

Useful Strong Association Rules

3.5 Analysis

It is evident that frequent pattern mining can provide the solution for association rules formation among the most dominating amino acids for different protein misfolded diseases. Three studies [2, 5 and 6] have been identified on this issue as stated in the literature review. The major limitations of these studies are as follows:

a. All these studies were focused to predict the pattern and association rules of the most dominating amino acids which causes the *Chromaffin Tumor* disease only.

b. Predicting the pattern and associations between different protein misfolded diseases were not attempted.

c. Support threshold were considered relatively high which might by pass many interesting rules generation.

d. Association rules mining algorithm can generate a lot of association rules or patters or knowledge, among which all rules may not contain useful information. Therefore, it is needed to evaluate the interestingness (or usefulness) of the association rules. Unfortunately none of the studies as stated earlier predicted the patterns and association rules of amino acids with due measures of interestingness.

Considering the limitation of earlier studies, this work designed a uniform method to predict the patterns and association rules of the most dominating amino acids for different protein misfolded diseases. The support thresholds were kept relatively low to examine large amount of frequent patterns and their association rules. And the rules were then tested using improved objective measuring tools (*Bi-lift, Bi-improve* and *Bi-confidence*) to evaluate the association rules comprehensively. Finally following patterns and useful strong association rules of the most dominating amino acids for the experimented protein misfolded diseases were found as outcome:

Disease-1: Sickle Cell Anemia				
GT -> AN	GT -> KN	AGT -> KN	GKT -> AN	GT -> AKN
AN -> GK	GS -> FL	NT -> GK	KP -> TV	ANT -> GK
NT -> AGK	ANV -> GK	GT -> N	AGT -> N	GKT -> N
AGKT -> N	KP -> T	GH -> AL	GT -> AK	NT -> AK
KPV -> T	GNT -> AK	KN -> AG	GS -> F	FS -> GL
GLS -> F	NT -> AG	KNT -> AG	KNV -> AG	AN -> K
AT -> K	AGN -> K	GT -> K	NT -> K	AGT -> K
ANT -> K	GNT -> K	ANV -> K	ATV -> K	AGNT -> K
AGNV -> K	FL -> G	AN -> G	KN -> G	NV -> G
AKN -> G	ALV -> G	AD -> G	LN -> G	FS -> G
NT -> G	AFL -> G	FLS -> G	ANT -> G	KNT -> G
ANV -> G	KNV -> G	AKNT -> G	AKNV -> G	

Disease-2: Breast Cancer						
ANPT -> G	NQST -> P	FLN -> P	GKLN -> P	GLNT -> P	LNQST -> P	ILQS -> N
IPSV -> K	EKQV -> L	DHP -> L	QRT -> L	GPST -> L	GQRS -> L	NQRS -> L
DPY -> L	DEHP -> L	FPST -> L	EKQSV -> L	NQR -> L		

Disease-3: Cystic Fibrosis						
EKLP -> Q	PVW -> A	CLR -> A	HILV -> T	HILS -> T	FPR -> V	FIPR -> V
APW -> V	AQW -> V	PRT -> V	FILP -> V	HKLS -> V	LPST -> V	FIN -> K
LPQR -> K	HISV -> K	DLRS -> I	AFLV -> I	DKSV -> I	FMR -> I	FGQ -> I
ADKS -> I	ALQR -> S	HKV -> S	DIKV -> S	DIM -> S	HKR -> S	ADN -> S
AIKN -> S	HIKT -> S	HIKV -> S	HKLV -> S	IKLTV -> S	IKLV -> S	DIR -> S

Disease-4: Nephrogenic Diabetes Insipidus						
DLP -> E	FG -> AP	GI -> AV	CV -> AL	AE -> P	DE -> P	FG -> P
AFG -> P	AEL -> P	DEL -> P	GI -> V	AGI -> V	N -> S	AN -> S

Disease-5: Retinitis Pigmentosa 4							
ALS -> W	W -> AL	PW -> AL	SW -> AL	QS -> E	AFP -> S	NY -> P	AFS -> P
FTV -> G	FNP -> G	AFT -> G	FNPT -> G	AY -> I	H -> T	QV -> T	FGY -> T
ALY -> I	FH -> T	KV -> T	FGI -> T	W -> L	AVY -> I	AW -> L	CI -> L
PW -> L	SW -> L	APW -> L	ASW -> L	GT -> F	GNT -> F	GM -> F	GTV -> F
GTY -> F	GPT -> F	EM -> F	MS -> F	CY -> V	AGT -> F	GIT -> F	GMN -> F
MNT -> F	W -> A	GNPT-> F	LW -> A	PW -> A	SW -> A	LPW -> A	LSW -> A
ILV -> A							

It has been already mentioned that all the previous studies, in this aspect, were focused to predict the pattern and association rules of the most dominating amino acids which causes the *Chromaffin Tumor* disease only. From the literature [2, 5 and 6], following are the accepted strong association rules as generated for *Chromaffin Tumor* disease:

- $PN \rightarrow L$ ..... [2]
- $PI \rightarrow K$ ..... [2, 6]
- $I \rightarrow K$ ..... [5]
- $V \rightarrow L$ ..... [5]

In this work, the same protein sequence (involved with *Chromaffin Tumor* disease) was tested and the result is shown in table 3.15. From this table is evident that $PN \rightarrow L$ and $PI \rightarrow K$ rules as generated by the literature [2, 5 and 6] are useful strong association rules and $I \rightarrow K$ and $V \rightarrow L$ are redundant and should be thus rejected. On the other hand $F \rightarrow D$,

DN $\rightarrow$*L* and *KLY*$\rightarrow$*P* are useful strong association rules which were discarded by the literature.

Table 3.15 : Useful Strong Association Rules for *Chromaffin Tumor* disease

Ser	Rules	Min Support Count	Confidence	Lift	Bi-lift	Bi-Improve	Bi-confidence
1.	F -> D	5	100%	1.75	2.091	0.093	0.522
2.	DN -> L	5	100%	1.12	1.15	0.023	0.130
3.	PN -> L	5	100%	1.12	1.15	0.023	0.130
4.	PI -> K	5	100%	1.12	1.15	0.023	0.130
5.	KLY -> P	5	100%	2	2.556	0.109	0.609

CHAPTER-4 : CONCLUSION AND FUTURE WORK

4.1 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 are an important issue. This work focused to recognize frequent patterns among five complex protein misfolded neurodegenerative human diseases and the relationship of the dominating amino acids. The diseases are*Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus* and *Retinitis Pigmentosa 4.*Association rule mining technique was used for pattern identification. In doing so, itemsets and association rules were generated from the protein sequences considering the minimum support count between 3 and 5 and minimum confidence level as 90%. By this way, a huge number of association rules were generated. As all these rules are not useful and reliable, these rules were further evaluated and sorted out with objective measuring tools. These measuring methods identified only the strong and interesting association rules as associated with the concerned protein misfolded diseases.

The final useful rules association were found to be only 59, 19, 35, 14 and 49 which indicate the most dominating acids and their patterns for *Sickle Cell Anemia, Breast Cancer, Cystic Fibrosis, Nephrogenic Diabetes Insipidus (NDI),* and *Retinitis Pigmentosa 4 (RP4*).

Patterns in protein sequences usually have functional, structural or family classification importance. Pattern identification can be used for predicting protein functions, protein fold (structure) recognitions, protein family detection, multiple sequence alignment, etc. This thesis work is focused to predict the pattern of the most dominating amino acids in the protein sequences associated with particular protein misfolded diseases. The patterns acquired from this work are quite impressive. In addition to the above usual applications, the identified amino acid patterns could be more useful in discovering medicines for concerned protein misfolded diseases and thereby this work may open up new opportunities in medical science to handle genetic disorder diseases.

4.2 Future Work

In this work, only five protein misfolded diseases were experimented. Again protein sequence length of some of the diseases was relatively small. However, in future, more complex protein misfolded diseases and associated with larger length of protein sequences may be considered for experimentation. On the other hand, in this work Apriori algorithm was used as a pattern mining technique for association rule mining. However, as a newer method, Fuzzy Association rule mining technique may be adopted to generate more reliable association rules and test accordingly.

In this work, the protein sequences were partitioned into subsequences of length 10. If the length of the subsequences is changed, the generated rules may also be changed. As such, in future, frequent itemsets and rules can be generated considering the lengths of the subsequences as 10, 15, 20, ….. and thereafter only the common rules between each list can be sorted out. Generating rules in this way may have better potentiality and validity.

REFERENCE

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

[2] Priya, G. L. and Hariharan, S. (2012), "A Study on Predicting Patterns Over the Protein Sequence Datasets using Association Rule Mining", *Journal of Engineering Science and Technology*, vol. 7, No. 5, pp. 563 – 573.

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

[4] Science Museum website, "what are proteins made of", Retrieved from http://whoami.sciencemuseum.org.uk , visited on 10 Jan 2018.

[5] Priya, G. L. and Hariharan, S. (2012), "An Efficient Approach for Generating Frequent Patterns without Candidate Generation", *Proceedings of the International Conference on Advances in Computing, Communications and Informatics 2012 (ICACCI'12)*, pp. 1061-1067, DOI: 10.1145/2345396.2345566

[6] Dhumale, S. (2015), "Predicting Patterns over Protein Sequences Using Apriori Algorithm", *International Journal of Engineering and Computer Science*, vol. 4 Issue 7, 13011-13016.

[7] UoW (2000), "Finding Patterns in Biological Sequences", Technical report CS-2000-22 University of Waterloo, December 2000, Ontario, Canada

[8] Wikipedia contributors. (2017, December 18). Association rule learning. In Wikipedia, The Free Encyclopedia. Retrieved 19:49, December 4, 2017, from https://en.wikipedia.org/w/index.php?title=Association_rule_learning&oldid=890442892

[9] Agrawal, R. T., Imielinski, and Swami, A. (1993), "Mining Association Rules between Sets of Items in Large Databases", In Proc. of the ACM SIGMOD Conference on Management of Data, Washington, D.C., May 1993.

[10] Mining Frequent Itemsets – Apriori Algorithm, Retrieved from http://software.ucv.ro/~cmihaescu/ro/teaching/AIR/docs/Lab8-Apriori.pdf., visited on 16 October 2017

[11] Jiawei Han, Hong Cheng, Dong Xin and Xifeng Yan, (2007), "Frequent pattern mining: current status and future directions", Data Mining Knowledge Discovery 15:55 -86.

[12] Kang, Ho, T., Yoo, J. S. and Kim, H. Y. (2007), "Mining Frequent Contiguous Sequence Patterns in Biological Sequences", *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, pp. 723-728.

[13] Leung, K. T., Wong, K. C., Chan, T. M., Wong, M., Lee, K. H., Lau, C. P. and Stephen, K. W. T. (2010), "Discovering Protein–DNA Binding Sequence Patterns using Association Rule Mining", *Nucleic acids research*.

[14] Das, N. S, and Poonam. (2013), "Brief Survey on DNA Sequence Mining", *International Journal of Computer Science and Mobile Computing,* Vol.2 Issue. 11, pg. 129-134

[15] Mutakabbir, Mahbub, K., Mahin, S. S. and Hasan, M. A. (2014), "Mining Frequent Pattern within a Genetic Sequence using Unique Pattern Indexing and Mapping Techniques", *International Conference on Informatics, Electronics & Vision (ICIEV), pp.* 1-5.

[16] Zhang, Jingsong, Wang, Y., Zhang, C. and Shi, Y. (2016), "Mining Contiguous Sequential Generators in Biological Sequences", *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 13, pp. 855-867.

[17] ShikhaMaheshwari and Pooja Jain. (2013), "Novel Method of Apriori Algorithm using Top Down Approach", *International Journal of Computer Applications,* vol. 77 serial 10, pp.18-21, DOI:10.5120/13430-1126

[18] The biology project, Biochemistry. *The Chemistry of amino acid*. Retrieved from http://www.biology.arizona.edu/ biochemistry/problem_sets/aa/aa.html, visited on 13 Oct 2017.

[19] Angela Lynn Swafford, What Are Amino Acids? - Definition & Structure Retrieved from https://study.com/academy/lesson/what-are-amino-acids-definition-structure-quiz.html, visited on 13 Oct 2017.

[20] Khab Academy, Chemistry of amino acids and protein structure, Retrieved from https://www.khanacademy.org/test-prep/mcat/biomolecules/amino-acids-and-proteins1/a/chemistry-of-amino-acids-and-protein-structure, visited on 13 Oct 2017.

[21] Samuel, L., The Basics of Protein Structure and Function Retrieved from http://www.interactive-biology.com/6711/the-basics-of-protein-structure-and-function/, visited on 13 Oct 2017.

[22] Ellis, R. J. and Pinheiro, T. J. (2002), Danger – Misfolding Proteins. Nature, Apr 4;416(6880):483-4. DOI: 10.1038/416483a

[23] Gregersen, N., Bross, P., Vang, S. and Christensen, J. H. (2006), Protein Misfolding and Human Disease, Annual Review of Genomics and Human Genetics, 7:1, pp.103-124

[24] Bradbury, J. (2003), Chaperones: keeping a close eye on protein folding. Lancet 361 (9364), pp. 1194–1195.

[25] The UniProt Consortium, UniProt: The Universal Protein Knowledgebase atwww.uniprot.org/ uniprot/P68871, visited on 12 Feb 2017.

[26] The UniProt Consortium, UniProt: The Universal Protein Knowledgebase atwww.uniprot.org/ uniprot/P38398, visited on 12 Feb 2017.

[27] The UniProt Consortium, UniProt: The Universal Protein Knowledgebase atwww.uniprot.org/ uniprot/P13569, visited on 12 Feb 2017.

[28] The UniProt Consortium, UniProt: The Universal Protein Knowledgebase atwww.uniprot.org/ uniprot/P30518, visited on 12 Feb 2017.

[29] The UniProt Consortium, UniProt: The Universal Protein Knowledgebase atwww.uniprot.org/ uniprot/P08100, visited on 12 Feb 2017.

[30] Han, J., Kamber, M. and Pei, J. (2012), "*Data Mining Concepts and Techniques*", Third Edition. ISBN 978-0-12-381479-1, Morgan Kaufmann Publishers, USA

[31] Berson, A. and Smith, S. J. (2008), "*Data Warehousing, Data Mining, & Olap*", Tata McGraw-Hill Edition, thirteen reprint, p-122

[32] Li, X. (2011), "*Biological Data Mining and Its applications in Healthcare*" 1st ed. World Scientific Publishing Company.

[33] Fogel, G., Corne, D. and Pan, Y. (2008), "*Computational Intelligence in Bioinformatics*", 1st ed. IEE Press Series on Computational Intelligence.

[34] Raza, K.. (2010), "Application of Data Mining in Bioinformatics", *Indian Journal of Computer Science and Engineering*, vol 1 No 2, pp. 114-118

[35] Rajkumar, P. (2014), "14-useful-applications-of-data-mining", Retrieved from http://bigdata-madesimple.com/14-useful-applications-of-data-mining, visited on 21 Dec 2017

[36] Xifeng, Y., "frequent-pattern-mining" Retrieved from www.kdd.org/kdd2016/topics/view/frequent-pattern-mining, visited on 16 July 2017.

[37] Gupta, M. and Han, J. (2011), "Applications of Pattern Discovery Using Sequential Data Mining", in *Pattern Discovery Using Sequence Data Mining: Applications and Studies*. IGI Global, 2011, pp. 1-23. https://doi.org/10.4018/978-1-61350-056-9.ch001

[38] Shrivastava, A. and Jain, R. C. (2013), "Performance Analysis of Modified Algorithm for Finding Multilevel Association Rules", *Computer Science & Engineering: An International Journal (CSEIJ)*, Vol. 3, No. 4.

[39] Create Association Rules, *Rapid Miner Documentation*. Retrieved from https://docs.rapidminer.com/, visited on 22 Aug 2017

[40] Ramal, R. A. and Srikant, R. (1994), "Fast Algorithms for Mining Association Rules", *20th VLDB Conference*, pp. 487-499, Santiago, Chile.

[41] Fournier-Viger, P., "Introduction to the Apriori algorithm (with Java code)", *The Data Mining Blog*. Retrieved from http://data-mining.philippe-fournier-viger.com, visited on 23 September 2017

[42] Baker, Z. K. and Prasanna, V. (2005), "Efficient Parallel Data Mining with the Apriori Algorithm on FPGAs", Field-Programmable Custom Computing Machines, FCCM 2005. DOI:10.1109/FCCM.2005.31

[43] Hsu, C. L. "Frequent Itemset Generation Using Apriori Algorithm, An Explorer of Things. Retrieved from https://chih-ling-hsu.github.io/2017/03/25/apriori, visited on 20 November 2017.

[44] Tan, Pang-Ning; Michael, Steinbach, Kumar and Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms" (PDF), *Introduction to Data Mining*. Addison-Wesley. ISBN 978-0-321-32136-7.

[45] Prof Wasilewska, A., Lecture Note on "Apriori Algorithm", Retrieved from www3.cs.stonybrook.edu/~cse634/lecture_notes/07apriori.pdf visited on 15 Aug 2017

[46] Suresh, J. and Ramanjaneyulu, T. (2013), "Mining Frequent Itemsets Using Apriori Algorithm", *International Journal of Computer Trends and Technology (IJCTT)*, vol. 4, Issue. 4, pp. 760-764, ISSN: 2231-2803 http://www.ijcttjournal.org

[47] Ju, C., Bao, F., Xu, C. and Fu, X. (2015), "A Novel Method of Interestingness Measures for Association Rules Mining Based on Profit," *Discrete Dynamics in Nature and Society*, vol. 2015.

[48] Bhargava, N. and Shukla, M. (2016), "Survey of Interestingness Measures for Association Rules Mining: Data Mining", Data Science for Business Perspective, *IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS)*, ISSN: 2249-9555, vol.6, no.2, pp.74-80

[49] Li, Y., Wu, C. and Wang, K. (2011), "A new interestingness measures for Ming association rules," *Journal of the China Society for Scientific and Technical Information*, vol. 30, no. 5, pp. 503–507.

Appendix-A

Protein Sequences of Human Diseases

Disease-1: Sickle Cell Anemia

Involved Protein: *Hemoglobin Subunit Beta*

Entry Code: P68871

Length: 147

URL: www.uniprot.org/ uniprot/P68871

FASTA Form

```
MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
KEFTPPVQAAYQKVVAGVANALAHKYH
```

Sub Sequences

```
        10         20         30         40         50
MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS
        60         70         80         90        100
TPDAVMGNPK VKAHGKKVLG AFSDGLAHLD NLKGTFATLS ELHCDKLHVD
       110        120        130        140
PENFRLLGNV LVCVLAHHFG KEFTPPVQAA YQKVVAGVAN ALAHKYH
```

Disease-2: Breast Cancer

Involved Protein: *Breast Cancer Type 1 susceptibility protein*

Entry Code: P38398

Length: 1863

URL: www.uniprot.org/ uniprot/P38398

FASTA Form

```
MDLSALRVEEVQNVINAMQKILECPICLELIKEPVSTKCDHIFCKFCMLKLLNQKKGPSQ
CPLCKNDITKRSLQESTRFSQLVEELLKIICAFQLDTGLEYANSYNFAKKENNSPEHLKD
EVSIIQSMGYRNRAKRLLQSEPENPSLQETSLSVQLSNLGTVRTLRTKQRIQPQKTSVYI
ELGSDSSEDTVNKATYCSVGDQELLQITPQGTRDEISLDSAKKAACEFSETDVTNTEHHQ
PSNNDLNTTEKRAAERHPEKYQGSSVSNLHVEPCGTNTHASSLQHENSSLLLTKDRMNVE
KAEFCNKSKQPGLARSQHNRWAGSKETCNDRRTPSTEKKVDLNADPLCERKEWNKQKLPC
SENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGESESNAKVADVLDVLNEVD
EYSGSSEKIDLLASDPHEALICKSERVHSKSVESNIEDKIFGKTYRKKASLPNLSHVTEN
LIIGAFVTEPQIIQERPLTNKLKRKRRPTSGLHPEDFIKKADLAVQKTPEMINQGTNQTE
QNGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNI
HNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPV
RHSRNLQLMEGKEPATGAKKSNKPNEQTSKRHDSDTFPELKLTNAPGSFTKCSNTSELKE
FVNPSLPREEKEEKLETVKVSNNAEDPKDLMLSGERVLQTERSVESSSISLVPGTDYGTQ
ESISLLEVSTLGKAKTEPNKCVSQCAAFENPKGLIHGCSKDNRNDTEGFKYPLGHEVNHS
RETSIEMEESELDAQYLQNTFKVSKRQSFAPFSNPGNAEEECATFSAHSGSLKKQSPKVT
FECEQKEENQGKNESNIKPVQTVNITAGFPVVGQKDKPVDNAKCSIKGGSRFCLSSQFRG
```

NETGLITPNKHGLLQNPYRIPPLFPIKSFVKTKCKKNLLEENFEEHSMSPEREMGNENIP
STVSTISRNNIRENVFKEASSSNINEVGSSTNEVGSSINEIGSSDENIQAELGRNRGPKL
NAMLRLGVLQPEVYKQSLPGSNCKHPEIKKQEYEEVVQTVNTDFSPYLISDNLEQPMGSS
HASQVCSETPDDLLDDGEIKEDTSFAENDIKESSAVFSKSVQKGELSRSPSPFTHTHLAQ
GYRRGAKKLESSEENLSSEDEELPCFQHLLFGKVNNIPSQSTRHSTVATECLSKNTEENL
LSLKNSLNDCSNQVILAKASQEHHLSEETKCSASLFSSQCSELEDLTANTNTQDPFLIGS
SKQMRHQSESQGVGLSDKELVSDDEERGTGLEENNQEEQSMDSNLGEAASGCESETSVSE
DCSGLSSQSDILTTQQRDTMQHNLIKLQQEMAELEAVLEQHGSQPSNSYPSIISDSSALE
DLRNPEQSTSEKAVLTSQKSSEYPISQNPEGLSADKFEVSADSSTSKNKEPGVERSSPSK
CPSLDDRWYMHSCSGSLQNRNYPSQEELIKVVDVEEQQLEESGPHDLTETSYLPRQDLEG
TPYLESGISLFSDDPESDPSEDRAPESARVGNIPSSTSALKVPQLKVAESAQSPAAAHTT
DTAGYNAMEESVSREKPELTASTERVNKRMSMVVSGLTPEEFMLVYKFARKHHITLTNLI
TEETTHVVMKTDAEFVCERTLKYFLGIAGGKWVVSYFWVTQSIKERKMLNEHDFEVRGDV
VNGRNHQGPKRARESQDRKIFRGLEICCYGPFTNMPTDQLEWMVQLCGASVVKELSSFTL
GTGVHPIVVVQPDAWTEDNGFHAIGQMCEAPVVTREWVLDSVALYQCQELDTYLIPQIPH
SHY

Sub Sequences

```
        10         20         30         40         50
MDLSALRVEE VQNVINAMQK ILECPICLEL IKEPVSTKCD HIFCKFCMLK
        60         70         80         90        100
LLNQKKGPSQ CPLCKNDITK RSLQESTRFS QLVEELLKII CAFQLDTGLE
       110        120        130        140        150
YANSYNFAKK ENNSPEHLKD EVSIIQSMGY RNRAKRLLQS EPENPSLQET
       160        170        180        190        200
SLSVQLSNLG TVRTLRTKQR IQPQKTSVYI ELGSDSSEDT VNKATYCSVG
       210        220        230        240        250
DQELLQITPQ GTRDEISLDS AKKAACEFSE TDVTNTEHHQ PSNNDLNTTE
       260        270        280        290        300
KRAAERHPEK YQGSSVSNLH VEPCGTNTHA SSLQHENSSL LLTKDRMNVE
       310        320        330        340        350
KAEFCNKSKQ PGLARSQHNR WAGSKETCND RRTPSTEKKV DLNADPLCER
       360        370        380        390        400
KEWNKQKLPC SENPRDTEDV PWITLNSSIQ KVNEWFSRSD ELLGSDDSHD
       410        420        430        440        450
GESESNAKVA DVLDVLNEVD EYSGSSEKID LLASDPHEAL ICKSERVHSK
       460        470        480        490        500
SVESNIEDKI FGKTYRKKAS LPNLSHVTEN LIIGAFVTEP QIIQERPLTN
       510        520        530        540        550
KLKRKRRPTS GLHPEDFIKK ADLAVQKTPE MINQGTNQTE QNGQVMNITN
       560        570        580        590        600
SGHENKTKGD SIQNEKNPNP IESLEKESAF KTKAEPISSS ISNMELELNI
       610        620        630        640        650
HNSKAPKKNR LRRKSSTRHI HALELVVSRN LSPPNCTELQ IDSCSSSEEI
       660        670        680        690        700
KKKKYNQMPV RHSRNLQLME GKEPATGAKK SNKPNEQTSK RHDSDTFPEL
       710        720        730        740        750
KLTNAPGSFT KCSNTSELKE FVNPSLPREE KEEKLETVKV SNNAEDPKDL
       760        770        780        790        800
MLSGERVLQT ERSVESSSIS LVPGTDYGTQ ESISLLEVST LGKAKTEPNK
       810        820        830        840        850
CVSQCAAFEN PKGLIHGCSK DNRNDTEGFK YPLGHEVNHS RETSIEMEES
       860        870        880        890        900
ELDAQYLQNT FKVSKRQSFA PFSNPGNAEE ECATFSAHSG SLKKQSPKVT
       910        920        930        940        950
FECEQKEENQ GKNESNIKPV QTVNITAGFP VVGQKDKPVD NAKCSIKGGS
       960        970        980        990       1000
```

```
RFCLSSQFRG NETGLITPNK HGLLQNPYRI PPLFPIKSFV KTKCKKNLLE
      1010       1020       1030       1040       1050
ENFEEHSMSP EREMGNENIP STVSTISRNN IRENVFKEAS SSNINEVGSS
      1060       1070       1080       1090       1100
TNEVGSSINE IGSSDENIQA ELGRNRGPKL NAMLRLGVLQ PEVYKQSLPG
      1110       1120       1130       1140       1150
SNCKHPEIKK QEYEEVVQTV NTDFSPYLIS DNLEQPMGSS HASQVCSETP
      1160       1170       1180       1190       1200
DDLLDDGEIK EDTSFAENDI KESSAVFSKS VQKGELSRSP SPFTHTHLAQ
      1210       1220       1230       1240       1250
GYRRGAKKLE SSEENLSSED EELPCFQHLL FGKVNNIPSQ STRHSTVATE
      1260       1270       1280       1290       1300
CLSKNTEENL LSLKNSLNDC SNQVILAKAS QEHHLSEETK CSASLFSSQC
      1310       1320       1330       1340       1350
SELEDLTANT NTQDPFLIGS SKQMRHQSES QGVGLSDKEL VSDDEERGTG
      1360       1370       1380       1390       1400
LEENNQEEQS MDSNLGEAAS GCESETSVSE DCSGLSSQSD ILTTQQRDTM
      1410       1420       1430       1440       1450
QHNLIKLQQE MAELEAVLEQ HGSQPSNSYP SIISDSSALE DLRNPEQSTS
      1460       1470       1480       1490       1500
EKAVLTSQKS SEYPISQNPE GLSADKFEVS ADSSTSKNKE PGVERSSPSK
      1510       1520       1530       1540       1550
CPSLDDRWYM HSCSGSLQNR NYPSQEELIK VVDVEEQQLE ESGPHDLTET
      1560       1570       1580       1590       1600
SYLPRQDLEG TPYLESGISL FSDDPESDPS EDRAPESARV GNIPSSTSAL
      1610       1620       1630       1640       1650
KVPQLKVAES AQSPAAAHTT DTAGYNAMEE SVSREKPELT ASTERVNKRM
      1660       1670       1680       1690       1700
SMVVSGLTPE EFMLVYKFAR KHHITLTNLI TEETTHVVMK TDAEFVCERT
      1710       1720       1730       1740       1750
LKYFLGIAGG KWVVSYFWVT QSIKERKMLN EHDFEVRGDV VNGRNHQGPK
      1760       1770       1780       1790       1800
RARESQDRKI FRGLEICCYG PFTNMPTDQL EWMVQLCGAS VVKELSSFTL
      1810       1820       1830       1840       1850
GTGVHPIVVV QPDAWTEDNG FHAIGQMCEA PVVTREWVLD SVALYQCQEL
      1860
DTYLIPQIPH SHY
```

Disease-3: Cystic Fibrosis

Involved Protein: *Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)*

Entry Code: P13569

Length: 1480

URL: www.uniprot.org/ uniprot/P13569

FASTA Form

```
MQRSPLEKASVVSKLFFSWTRPILRKGYRQRLELSDIYQIPSVDSADNLSEKLEREWDRE
LASKKNPKLINALRRCFFWRFMFYGIFLYLGEVTKAVQPLLLGRIIASYDPDNKEERSIA
IYLGIGLCLLFIVRTLLLHPAIFGLHHIGMQMRIAMFSLIYKKTLKLSSRVLDKISIGQL
VSLLSNNLNKFDEGLALAHFVWIAPLQVALLMGLIWELLQASAFCGLGFLIVLALFQAGL
GRMMMKYRDQRAGKISERLVITSEMIENIQSVKAYCWEEAMEKMIENLRQTELKLTRKAA
YVRYFNSSAFFFSGFFVVFLSVLPYALIKGIILRKIFTTISFCIVLRMAVTRQFPWAVQT
WYDSLGAINKIQDFLQKQEYKTLEYNLTTTEVVMENVTAFWEEGFGELFEKAKQNNNNRK
TSNGDDSLFFSNFSLLGTPVLKDINFKIERGQLLAVAGSTGAGKTSLLMVIMGELEPSEG
KIKHSGRISFCSQFSWIMPGTIKENIIFGVSYDEYRYRSVIKACQLEEDISKFAEKDNIV
```

```
LGEGGITLSGGQRARISLARAVYKDADLYLLDSPFGYLDVLTEKEIFESCVCKLMANKTR
ILVTSKMEHLKKADKILILHEGSSYFYGTFSELQNLQPDFSSKLMGCDSFDQFSAERRNS
ILTETLHRFSLEGDAPVSWTETKKQSFKQTGEFGEKRKNSILNPINSIRKFSIVQKTPLQ
MNGIEEDSDEPLERRLSLVPDSEQGEAILPRISVISTGPTLQARRRQSVLNLMTHSVNQG
QNIHRKTTASTRKVSLAPQANLTELDIYSRRLSQETGLEISEEINEEDLKECFFDDMESI
PAVTTWNTYLRYITVHKSLIFVLIWCLVIFLAEVAASLVVLWLLGNTPLQDKGNSTHSRN
NSYAVIITSTSSYYVFYIYVGVADTLLAMGFFRGLPLVHTLITVSKILHHKMLHSVLQAP
MSTLNTLKAGGILNRFSKDIAILDDLLPLTIFDFIQLLLIVIGAIAVVAVLQPYIFVATV
PVIVAFIMLRAYFLQTSQQLKQLESEGRSPIFTHLVTSLKGLWTLRAFGRQPYFETLFHK
ALNLHTANWFLYLSTLRWFQMRIEMIFVIFFIAVTFISILTTGEGEGRVGIILTLAMNIM
STLQWAVNSSIDVDSLMRSVSRVFKFIDMPTEGKPTKSTKPYKNGQLSKVMIIENSHVKK
DDIWPSGGQMTVKDLTAKYTEGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLRL
LNTEGEIQIDGVSWDSITLQQWRKAFGVIPQKVFIFSGTFRKNLDPYEQWSDQEIWKVAD
EVGLRSVIEQFPGKLDFVLVDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPVT
YQIIRRTLKQAFADCTVILCEHRIEAMLECQQFLVIEENKVRQYDSIQKLLNERSLFRQA
ISPSDRVKLFPHRNSSKCKSKPQIAALKEETEEEVQDTRL
```

Sub Sequences

```
      10         20         30         40         50
MQRSPLEKAS VVSKLFFSWT RPILRKGYRQ RLELSDIYQI PSVDSADNLS
        60         70         80         90        100
EKLEREWDRE LASKKNPKLI NALRRCFFWR FMFYGIFLYL GEVTKAVQPL
       110        120        130        140        150
LLGRIIASYD PDNKEERSIA IYLGIGLCLL FIVRTLLLHP AIFGLHHIGM
       160        170        180        190        200
QMRIAMFSLI YKKTLKLSSR VLDKISIGQL VSLLSNNLNK FDEGLALAHF
       210        220        230        240        250
VWIAPLQVAL LMGLIWELLQ ASAFCGLGFL IVLALFQAGL GRMMMKYRDQ
       260        270        280        290        300
RAGKISERLV ITSEMIENIQ SVKAYCWEEA MEKMIENLRQ TELKLTRKAA
       310        320        330        340        350
YVRYFNSSAF FFSGFFVVFL SVLPYALIKG IILRKIFTTI SFCIVLRMAV
       360        370        380        390        400
TRQFPWAVQT WYDSLGAINK IQDFLQKQEY KTLEYNLTTT EVVMENVTAF
       410        420        430        440        450
WEEGFGELFE KAKQNNNNRK TSNGDDSLFF SNFSLLGTPV LKDINFKIER
       460        470        480        490        500
GQLLAVAGST GAGKTSLLMV IMGELEPSEG KIKHSGRISF CSQFSWIMPG
       510        520        530        540        550
TIKENIIFGV SYDEYRYRSV IKACQLEEDI SKFAEKDNIV LGEGGITLSG
       560        570        580        590        600
GQRARISLAR AVYKDADLYL LDSPFGYLDV LTEKEIFESC VCKLMANKTR
       610        620        630        640        650
ILVTSKMEHL KKADKILILH EGSSYFYGTF SELQNLQPDF SSKLMGCDSF
       660        670        680        690        700
DQFSAERRNS ILTETLHRFS LEGDAPVSWT ETKKQSFKQT GEFGEKRKNS
       710        720        730        740        750
ILNPINSIRK FSIVQKTPLQ MNGIEEDSDE PLERRLSLVP DSEQGEAILP
       760        770        780        790        800
RISVISTGPT LQARRRQSVL NLMTHSVNQG QNIHRKTTAS TRKVSLAPQA
       810        820        830        840        850
NLTELDIYSR RLSQETGLEI SEEINEEDLK ECFFDDMESI PAVTTWNTYL
       860        870        880        890        900
RYITVHKSLI FVLIWCLVIF LAEVAASLVV LWLLGNTPLQ DKGNSTHSRN
       910        920        930        940        950
NSYAVIITST SSYYVFYIYV GVADTLLAMG FFRGLPLVHT LITVSKILHH
       960        970        980        990       1000
KMLHSVLQAP MSTLNTLKAG GILNRFSKDI AILDDLLPLT IFDFIQLLLI
```

```
      1010       1020       1030       1040       1050
VIGAIAVVAV LQPYIFVATV PVIVAFIMLR AYFLQTSQQL KQLESEGRSP
      1060       1070       1080       1090       1100
IFTHLVTSLK GLWTLRAFGR QPYFETLFHK ALNLHTANWF LYLSTLRWFQ
      1110       1120       1130       1140       1150
MRIEMIFVIF FIAVTFISIL TTGEGEGRVG IILTLAMNIM STLQWAVNSS
      1160       1170       1180       1190       1200
IDVDSLMRSV SRVFKFIDMP TEGKPTKSTK PYKNGQLSKV MIIENSHVKK
      1210       1220       1230       1240       1250
DDIWPSGGQM TVKDLTAKYT EGGNAILENI SFSISPGQRV GLLGRTGSGK
      1260       1270       1280       1290       1300
STLLSAFLRL LNTEGEIQID GVSWDSITLQ QWRKAFGVIP QKVFIFSGTF
      1310       1320       1330       1340       1350
RKNLDPYEQW SDQEIWKVAD EVGLRSVIEQ FPGKLDFVLV DGGCVLSHGH
      1360       1370       1380       1390       1400
KQLMCLARSV LSKAKILLLD EPSAHLDPVT YQIIRRTLKQ AFADCTVILC
      1410       1420       1430       1440       1450
EHRIEAMLEC QQFLVIEENK VRQYDSIQKL LNERSLFRQA ISPSDRVKLF
      1460       1470       1480
PHRNSSKCKS KPQIAALKEE TEEEVQDTRL
```

Disease-4: Nephrogenic Diabetes Insipidus (NDI)

Involved Protein: Vasopressin V2 Receptor (V2R)

Entry Code: P30518

Length: 371

URL: www.uniprot.org/ uniprot/P30518

FASTA Form

```
MLMASTTSAVPGHPSLPSLPSNSSQERPLDTRDPLLARAELALLSIVFVAVALSNGLVLA
ALARRGRRGHWAPIHVFIGHLCLADLAVALFQVLPQLAWKATDRFRGPDALCRAVKYLQM
VGMYASSYMILAMTLDRHRAICRPMLAYRHGSGAHWNRPVLVAWAFSLLLSLPQLFIFAQ
RNVEGGSGVTDCWACFAEPWGRRTYVTWIALMVFVAPTLGIAACQVLIFREIHASLVPGP
SERPGGRRRGRRTGSPGEGAHVSAAVAKTVRMTLVIVVVYVLCWAPFFLVQLWAAWDPEA
PLEGAPFVLLMLLASLNSCTNPWIYASFSSSVSSELRSLLCCARGRTPPSLGPQDESCTT
ASSSLAKDTSS
```

Sub Sequences

```
        10         20         30         40         50
MLMASTTSAV PGHPSLPSLP SNSSQERPLD TRDPLLARAE LALLSIVFVA
        60         70         80         90        100
VALSNGLVLA ALARRGRRGH WAPIHVFIGH LCLADLAVAL FQVLPQLAWK
       110        120        130        140        150
ATDRFRGPDA LCRAVKYLQM VGMYASSYMI LAMTLDRHRA ICRPMLAYRH
       160        170        180        190        200
GSGAHWNRPV LVAWAFSLLL SLPQLFIFAQ RNVEGGSGVT DCWACFAEPW
       210        220        230        240        250
GRRTYVTWIA LMVFVAPTLG IAACQVLIFR EIHASLVPGP SERPGGRRRG
       260        270        280        290        300
RRTGSPGEGA HVSAAVAKTV RMTLVIVVVY VLCWAPFFLV QLWAAWDPEA
       310        320        330        340        350
PLEGAPFVLL MLLASLNSCT NPWIYASFSS SVSSELRSLL CCARGRTPPS
       360        370
LGPQDESCTT ASSSLAKDTS S
```

Disease-5: Retinitis Pigmentosa 4 (RP4)

Involved Protein: Rhodopsin (Opsin-2)

Entry Code: P08100

Length: 348

URL: www.uniprot.org/ uniprot/P08100

FASTA Form

```
MNGTEGPNFYVPFSNATGVVRSPFEYPQYYLAEPWQFSMLAAYMFLLIVLGFPINFLTLY
VTVQHKKLRTPLNYILLNLAVADLFMVLGGFTSTLYTSLHGYFVFGPTGCNLEGFFATLG
GEIALWSLVVLAIERYVVVCKPMSNFRFGENHAIMGVAFTWVMALACAAPPLAGWSRYIP
EGLQCSCGIDYYTLKPEVNNESFVIYMFVVHFTIPMIIIFFCYGQLVFTVKEAAAQQQES
ATTQKAEKEVTRMVIIMVIAFLICWVPYASVAFYIFTHQGSNFGPIFMTIPAFFAKSAAI
YNPVIYIMMNKQFRNCMLTTICCGKNPLGDDEASATVSKTETSQVAPA
```

Sub Sequences

```
        10         20         30         40         50
MNGTEGPNFY VPFSNATGVV RSPFEYPQYY LAEPWQFSML AAYMFLLIVL
        60         70         80         90        100
GFPINFLTLY VTVQHKKLRT PLNYILLNLA VADLFMVLGG FTSTLYTSLH
       110        120        130        140        150
GYFVFGPTGC NLEGFFATLG GEIALWSLVV LAIERYVVVC KPMSNFRFGE
       160        170        180        190        200
NHAIMGVAFT WVMALACAAP PLAGWSRYIP EGLQCSCGID YYTLKPEVNN
       210        220        230        240        250
ESFVIYMFVV HFTIPMIIIF FCYGQLVFTV KEAAAQQQES ATTQKAEKEV
       260        270        280        290        300
TRMVIIMVIA FLICWVPYAS VAFYIFTHQG SNFGPIFMTI PAFFAKSAAI
       310        320        330        340
YNPVIYIMMN KQFRNCMLTT ICCGKNPLGD DEASATVSKT ETSQVAPA
```

Appendix-B

Valid Itemsets Generation

Disease-2: Breast Cancer
(Protein: Breast Cancer Type 1 Susceptibility Protein)

Minimum Support Count Considered: 5

In the following tables, some itemsets of total 1806 are shown.

Ser	Itemsets	Suport Count
1	{A}	84
2	{C}	44
3	{D}	85
4	{E}	198
5	{F}	49
6	{G}	87
7	{H}	49
8	{I}	77
9	{K}	137
10	{L}	156
11	{M}	30
12	{N}	121
13	{P}	96
14	{Q}	97
15	{R}	76
16	{S}	224
17	{T}	111
18	{V}	101
19	{W}	10
20	{Y}	31

Ser	Itemsets	Suport Count
21	{A, C}	16
22	{A, E}	47
23	{A, F}	23
24	{A, G}	26
25	{A, H}	12
26	{A, I}	15
27	{A, K}	32
28	{A, N}	32
29	{A, P}	22
30	{A, Q}	25
31	{A, R}	18
32	{A, T}	27
33	{A, V}	27
34	{A, Y}	9
35	{C, F}	13
36	{C, G}	14
37	{C, H}	10
38	{C, P}	12
39	{C, T}	14
40	{D, A}	20

Ser	Itemsets	Suport Count
177	{S, V}	51
178	{S, W}	6
179	{S, Y}	19
180	{T, F}	17
181	{T, G}	31
182	{T, H}	17
183	{T, W}	5
184	{T, Y}	11
185	{V, C}	10
186	{V, E}	56
187	{V, F}	16
188	{V, G}	30
189	{V, H}	13
190	{V, I}	21
191	{V, K}	38
192	{V, N}	30
193	{V, P}	29
194	{V, Q}	33
195	{V, T}	36
196	{V, Y}	12

Ser	Itemsets	Suport Count
197	{A, C, F}	8
198	{A, C, G}	8
199	{A, C, T}	7
200	{A, E, C}	13
201	{A, E, F}	15

Ser	Itemsets	Suport Count
651	{Q, K, G}	7
652	{Q, K, P}	15
653	{Q, K, T}	7
654	{Q, N, C}	6
655	{Q, N, F}	7

Ser	Itemsets	Suport Count
846	{V, N, I}	11
847	{V, N, K}	13
848	{V, N, P}	10
849	{V, N, T}	11
850	{V, P, F}	5

202	{A, E, G}	17
203	{A, E, H}	8
204	{A, E, I}	9
205	{A, E, K}	21
206	{A, E, N}	20
207	{A, E, P}	15
208	{A, E, Q}	15
209	{A, E, T}	20
210	{A, E, Y}	5
211	{A, F, G}	10
212	{A, G, Y}	5
213	{A, I, F}	7
214	{A, I, G}	7
215	{A, I, K}	8
216	{A, I, T}	5

656	{Q, N, G}	17
657	{Q, N, H}	10
658	{Q, N, I}	16
659	{Q, N, K}	15
660	{Q, N, P}	22
661	{Q, N, T}	14
662	{Q, N, Y}	7
663	{Q, P, F}	6
664	{Q, P, G}	15
665	{Q, P, H}	9
666	{Q, P, T}	19
667	{Q, P, Y}	10
668	{Q, T, F}	6
669	{Q, T, G}	8
670	{Q, T, H}	6

851	{V, P, G}	14
852	{V, P, H}	6
853	{V, P, T}	16
854	{V, P, Y}	5
855	{V, Q, G}	15
856	{V, Q, I}	8
857	{V, Q, K}	17
858	{V, Q, N}	12
859	{V, Q, P}	13
860	{V, Q, T}	12
861	{V, Q, Y}	8
862	{V, T, F}	5
863	{V, T, G}	12
864	{V, T, H}	7
865	{V, T, Y}	5

Ser	Itemsets	Suport Count
866	{A, C, T, G}	5
867	{A, E, C, F}	7
868	{A, E, C, G}	6
869	{A, E, C, T}	6
870	{A, E, F, G}	6
871	{A, E, I, F}	5
872	{A, E, K, F}	7
873	{A, E, K, G}	6
874	{A, E, K, P}	7
875	{A, E, K, T}	8
876	{A, E, N, C}	5
877	{A, E, N, F}	5
878	{A, E, N, G}	9
879	{A, E, N, K}	8
880	{A, E, N, P}	6
881	{A, E, N, T}	10
882	{A, E, P, G}	6
883	{A, E, P, T}	8
884	{A, E, Q, C}	7
885	{A, E, Q, G}	5

Ser	Itemsets	Suport Count
1000	{D, R, E, G}	5
1001	{D, R, E, N}	6
1002	{D, R, E, P}	7
1003	{D, R, E, T}	9
1004	{D, R, V, E}	9
1005	{D, R, V, T}	5
1006	{D, S, A, E}	13
1007	{D, S, A, K}	5
1008	{D, S, A, N}	7
1009	{D, S, E, F}	5
1010	{D, S, E, G}	14
1011	{D, S, E, H}	6
1012	{D, S, E, I}	9
1013	{D, S, E, K}	12
1014	{D, S, E, N}	16
1015	{D, S, E, P}	13
1016	{D, S, E, Q}	6
1017	{D, S, E, T}	14
1018	{D, S, I, T}	5
1019	{D, S, K, G}	5

Ser	Itemsets	Suport Count
1590	{V, I, P, G}	5
1591	{V, I, P, T}	5
1592	{V, I, T, G}	5
1593	{V, K, P, G}	7
1594	{V, K, P, T}	6
1595	{V, N, I, G}	6
1596	{V, N, I, K}	6
1597	{V, N, K, G}	5
1598	{V, N, P, G}	6
1599	{V, N, T, G}	5
1600	{V, P, T, G}	6
1601	{V, Q, I, K}	5
1602	{V, Q, K, G}	6
1603	{V, Q, K, P}	10
1604	{V, Q, K, T}	5
1605	{V, Q, N, G}	7
1606	{V, Q, N, I}	5
1607	{V, Q, N, K}	5
1608	{V, Q, P, G}	7
1609	{V, Q, P, T}	6

Ser	Itemsets	Suport Count
1610	{A, E, N, T, G}	5
1611	{A, E, P, T, G}	5
1612	{A, N, P, T, G}	6
1613	{D, A, E, N, G}	5
1614	{D, A, E, N, T}	7
1615	{D, E, N, K, T}	5
1616	{D, E, N, T, G}	5
1617	{D, L, A, E, N}	5
1618	{D, L, E, N, P}	6
1619	{D, L, E, N, T}	5
1620	{D, L, E, P, H}	5
1621	{D, L, E, P, T}	7
1622	{D, L, E, Q, P}	5
1623	{D, L, E, Q, T}	5
1624	{D, L, I, P, T}	5
1625	{D, L, N, P, T}	6
1626	{D, L, Q, P, T}	7
1627	{D, L, R, E, P}	5
1628	{D, L, R, E, T}	5
1629	{D, L, S, A, E}	7

Ser	Itemsets	Suport Count
1787	{S, V, E, N, I}	5
1788	{S, V, E, N, K}	6
1789	{S, V, E, N, P}	5
1790	{S, V, E, P, G}	6
1791	{S, V, E, P, T}	7
1792	{S, V, E, Q, G}	6
1793	{S, V, E, Q, K}	5
1794	{S, V, E, T, G}	5
1795	{S, V, I, K, P}	5
1796	{S, V, K, P, G}	5
1797	{S, V, K, P, T}	5
1798	{S, V, N, I, K}	5
1799	{S, V, Q, K, P}	6
1800	{V, Q, K, P, G}	5
1801	{D, L, S, E, N, P}	5
1802	{L, S, E, N, P, T}	5
1803	{L, S, E, Q, N, P}	5
1804	{L, S, Q, N, P, T}	5
1805	{L, S, V, E, Q, G}	5
1806	{L, S, V, E, Q, K}	5

Appendix-C

Valid Itemsets Generation

Disease-3: Cystic Fibrosis
(Protein: Cystic Fibrosis Transmembrane Conductance Regulator)

Minimum Support Count Considered: 5

In the following tables, some itemsets of total 1464 are shown.

Ser	Itemsets	Suport Count
1	{A}	83
2	{C}	18
3	{D}	58
4	{E}	93
5	{F}	85
6	{G}	84
7	{H}	25
8	{I}	119
9	{K}	92
10	{L}	183
11	{M}	37
12	{N}	54
13	{P}	45
14	{Q}	67
15	{R}	78
16	{S}	123
17	{T}	83
18	{V}	90
19	{W}	23
20	{Y}	40

Ser	Itemsets	Suport Count
21	{A, C}	9
22	{A, D}	19
23	{A, F}	22
24	{A, G}	20
25	{A, H}	8
26	{A, I}	31
27	{A, N}	20
28	{A, T}	26
29	{A, V}	35
30	{A, W}	12
31	{A, Y}	11
32	{D, C}	5
33	{D, H}	5
34	{D, N}	15
35	{E, A}	20
36	{E, C}	5
37	{E, D}	22
38	{E, F}	19
39	{E, G}	20
40	{E, H}	7

Ser	Itemsets	Suport Count
179	{T, N}	19
180	{T, Y}	13
181	{V, C}	7
182	{V, D}	20
183	{V, F}	29
184	{V, G}	26
185	{V, H}	11
186	{V, I}	36
187	{V, N}	15
188	{V, T}	33
189	{V, W}	11
190	{V, Y}	14
191	{W, D}	7
192	{W, G}	10
193	{W, I}	9
194	{W, N}	7
195	{W, T}	10
196	{W, Y}	5
197	{Y, D}	12
198	{Y, N}	8

Ser	Itemsets	Suport Count
199	{A, D, N}	5
200	{A, F, G}	6
201	{A, F, I}	10
202	{A, F, N}	7
203	{A, F, T}	9

Ser	Itemsets	Suport Count
501	{Q, A, F}	8
502	{Q, A, G}	6
503	{Q, A, I}	11
504	{Q, A, N}	5
505	{Q, A, T}	8

Ser	Itemsets	Suport Count
786	{V, F, G}	10
787	{V, F, I}	20
788	{V, F, N}	6
789	{V, F, T}	13
790	{V, G, D}	7

204	{A, F, W}	5
205	{A, G, D}	6
206	{A, I, D}	11
207	{A, I, G}	11
208	{A, I, N}	8
209	{A, I, Y}	5
210	{A, T, D}	6
211	{A, T, G}	7
212	{A, T, I}	7
213	{A, T, N}	9
214	{A, T, Y}	5
215	{A, V, C}	5
216	{A, V, D}	9
217	{A, V, F}	11
218	{A, V, G}	10

506	{Q, A, V}	13
507	{Q, A, W}	5
508	{Q, E, A}	8
509	{Q, E, D}	10
510	{Q, E, F}	7
511	{Q, E, G}	7
512	{Q, E, I}	13
513	{Q, E, K}	12
514	{Q, E, N}	8
515	{Q, E, T}	7
516	{Q, E, V}	5
517	{Q, F, G}	5
518	{Q, F, I}	11
519	{Q, F, T}	8
520	{Q, F, Y}	5

791	{V, I, D}	9
792	{V, I, G}	12
793	{V, I, H}	6
794	{V, I, N}	5
795	{V, I, Y}	6
796	{V, T, D}	7
797	{V, T, G}	13
798	{V, T, H}	8
799	{V, T, I}	14
800	{V, T, N}	8
801	{V, T, Y}	5
802	{V, W, I}	5
803	{V, W, T}	6
804	{V, Y, D}	5
805	{W, I, G}	6

Ser	Itemsets	Suport Count
521	{Q, G, D}	6
522	{Q, I, D}	11
523	{Q, I, G}	15
524	{Q, I, N}	5
525	{Q, I, Y}	6
526	{Q, K, A}	11
527	{Q, K, D}	7
528	{Q, K, F}	7
529	{Q, K, G}	8
530	{Q, K, I}	14
531	{Q, K, N}	6
532	{Q, K, T}	8
533	{Q, K, V}	12
534	{Q, K, Y}	8
535	{Q, L, A}	18
536	{Q, L, D}	12
537	{Q, L, E}	19
538	{Q, L, F}	12
539	{Q, L, G}	16
540	{Q, L, I}	23

Ser	Itemsets	Suport Count
786	{V, F, G}	10
787	{V, F, I}	20
788	{V, F, N}	6
789	{V, F, T}	13
790	{V, G, D}	7
791	{V, I, D}	9
792	{V, I, G}	12
793	{V, I, H}	6
794	{V, I, N}	5
795	{V, I, Y}	6
796	{V, T, D}	7
797	{V, T, G}	13
798	{V, T, H}	8
799	{V, T, I}	14
800	{V, T, N}	8
801	{V, T, Y}	5
802	{V, W, I}	5
803	{V, W, T}	6
804	{V, Y, D}	5
805	{W, I, G}	6

Ser	Itemsets	Suport Count
806	{A, V, F, I}	8
807	{A, V, F, T}	5
808	{A, V, I, G}	5
809	{A, V, T, D}	5
810	{A, V, T, G}	5
811	{A, V, T, N}	5
812	{E, A, I, D}	5
813	{E, I, D, N}	7
814	{E, K, A, I}	6
815	{E, K, A, V}	5
816	{E, K, D, N}	5
817	{E, K, F, I}	6
818	{E, K, F, N}	5
819	{E, K, I, D}	7
820	{E, K, I, N}	8
821	{E, K, V, I}	7
822	{K, A, I, D}	7
823	{K, A, I, N}	5
824	{K, A, V, G}	5
825	{K, A, V, I}	5

Ser	Itemsets	Suport Count
1196	{R, S, L, I}	16
1197	{R, S, L, K}	12
1198	{R, S, L, T}	9
1199	{R, S, L, V}	11
1200	{R, S, L, Y}	7
1201	{R, S, P, I}	6
1202	{R, S, P, K}	8
1203	{R, S, P, L}	6
1204	{R, S, P, V}	6
1205	{R, S, T, I}	6
1206	{R, S, V, D}	5
1207	{R, S, V, F}	5
1208	{R, S, V, I}	10
1209	{R, S, Y, D}	5
1210	{R, V, F, I}	8
1211	{R, V, I, G}	5
1212	{S, A, D, N}	5
1213	{S, A, I, D}	7
1214	{S, A, I, G}	6
1215	{S, A, I, N}	6

Ser	Itemsets	Suport Count
1445	{S, L, K, V, H}	5
1446	{S, L, K, V, I}	10
1447	{S, L, K, V, T}	8
1448	{S, L, T, I, H}	5
1449	{S, L, V, F, I}	5
1450	{S, L, V, F, T}	5
1451	{S, L, V, G, D}	5
1452	{S, L, V, I, D}	5
1453	{S, L, V, I, G}	5
1454	{S, L, V, T, G}	6
1455	{S, L, V, T, H}	6
1456	{S, L, V, T, I}	7
1457	{S, P, L, A, V}	6
1458	{S, P, L, K, A}	5
1459	{S, P, L, K, I}	5
1460	{S, P, L, K, V}	6
1461	{S, P, L, V, D}	5
1462	{S, P, L, V, G}	5
1463	{S, P, L, V, T}	5
1464	{S, L, K, V, T, I}	5

Appendix-D

Valid Itemsets Generation

Disease-4: Nephrogenic Diabetes Insipidus
(Protein: Vasopressin V2 Receptor)

Minimum Support Count Considered: 4

In the following tables, some itemsets of total 234 are shown.

Itemsets	Suport Count
1. {A}	47
2. {C}	11
3. {D}	10
4. {E}	11
5. {F}	14
6. {G}	24
7. {H}	9
8. {I}	12
9. {K}	4
10. {L}	49
11. {M}	10
12. {N}	6
13. {P}	26
14. {Q}	9
15. {R}	29
16. {S}	35
17. {T}	17
18. {V}	30
19. {W}	11
20. {Y}	7

Itemsets	Suport Count
21. {A, C}	8
22. {A, D}	7
23. {A, E}	6
24. {A, F}	12
25. {A, G}	12
26. {A, H}	7
27. {A, I}	9
28. {A, K}	4
29. {A, N}	4
30. {A, P}	16
31. {A, Q}	5
32. {A, R}	11
33. {A, S}	14
34. {A, T}	11
35. {A, V}	17
36. {A, W}	9
37. {A, Y}	5
38. {E, D}	5
39. {E, R}	6
40. {F, W}	6

Itemsets	Suport Count
90. {S, P}	10
91. {S, R}	7
92. {S, T}	8
93. {S, V}	10
94. {T, D}	5
95. {T, E}	4
96. {T, G}	7
97. {T, P}	6
98. {T, R}	8
99. {T, V}	6
100. {V, C}	4
101. {V, E}	4
102. {V, F}	8
103. {V, G}	9
104. {V, H}	4
105. {V, I}	7
106. {V, P}	7
107. {V, R}	7
108. {V, W}	6
109. {V, Y}	4

Itemsets	Suport Count
110. {A, F, W}	6
111. {A, G, F}	4
112. {A, G, H}	4
113. {A, G, I}	4
114. {A, G, R}	6
115. {A, H, R}	4
116. {A, I, F}	5
117. {A, I, Y}	4
118. {A, P, C}	4
119. {A, P, D}	4
120. {A, P, E}	6
121. {A, P, F}	9
122. {A, P, G}	8
123. {A, P, H}	4
124. {A, P, I}	5
125. {A, P, R}	6
126. {A, P, W}	7
127. {A, R, C}	4
128. {A, S, F}	4
129. {A, S, G}	6

Itemsets	Suport Count
189. {P, G, E}	5
190. {P, G, F}	4
191. {P, G, H}	4
192. {P, G, R}	5
193. {S, E, R}	5

Itemsets	Suport Count
209. {A, P, F, W}	5
210. {A, P, G, F}	4
211. {A, P, G, R}	4
212. {A, S, P, G}	4
213. {A, S, V, G}	4

Itemsets	Suport Count
229. {L, P, E, D}	4
230. {L, V, P, F}	4
231. {M, L, A, T}	4
232. {S, P, G, E}	4
233. {S, P, G, R}	4

Itemsets	Suport Count
194. {S, G, E}	5
195. {S, G, R}	5
196. {S, P, E}	5
197. {S, P, G}	7
198. {S, P, R}	5
199. {S, T, G}	4
200. {S, V, G}	5
201. {T, G, R}	5
202. {T, P, G}	5
203. {T, P, R}	4
204. {V, F, W}	4
205. {V, G, I}	4
206. {V, P, F}	5
207. {V, P, G}	5
208. {V, P, W}	4

Itemsets	Suport Count
214. {A, T, G, R}	4
215. {A, T, P, G}	4
216. {A, T, P, R}	4
217. {A, V, F, W}	4
218. {A, V, G, I}	4
219. {A, V, P, F}	5
220. {A, V, P, G}	5
221. {A, V, P, W}	4
222. {L, A, P, E}	4
223. {L, A, P, F}	5
224. {L, A, S, V}	5
225. {L, A, V, C}	4
226. {L, A, V, F}	7
227. {L, A, V, G}	4
228. {L, A, V, P}	5

Itemsets	Suport Count
234. {L, A, V, P, F}	4

Appendix-E

Valid Itemsets Generation

Disease-5: Retinitis Pigmentosa 4
(Protein: Rhodopsin)

Minimum Support Count Considered: 4

In the following tables, some itemsets of total 268 are shown.

Itemsets	Suport Count
1. {A}	32
2. {C}	10
3. {D}	4
4. {E}	16
5. {F}	30
6. {G}	22
7. {H}	5
8. {I}	24
9. {K}	11
10. {L}	29
11. {M}	15
12. {N}	16
13. {P}	20
14. {Q}	12
15. {R}	7
16. {S}	17
17. {T}	24
18. {V}	30
19. {W}	5
20. {Y}	19

Itemsets	Suport Count
21. {A, I}	10
22. {A, K}	4
23. {A, L}	10
24. {A, Q}	5
25. {A, W}	5
26. {E, A}	8
27. {E, F}	6
28. {E, I}	4
29. {E, K}	5
30. {E, L}	6
31. {E, P}	6
32. {E, Q}	6
33. {E, S}	9
34. {E, V}	7
35. {E, Y}	5
36. {F, A}	9
37. {F, C}	4
38. {F, H}	4
39. {F, I}	9
40. {F, L}	9

Itemsets	Suport Count
108. {T, K}	5
109. {T, L}	7
110. {T, P}	8
111. {T, Q}	6
112. {T, S}	5
113. {T, V}	11
114. {T, Y}	7
115. {V, A}	13
116. {V, C}	5
117. {V, I}	9
118. {V, K}	4
119. {V, L}	9
120. {V, Q}	5
121. {V, S}	6
122. {Y, A}	6
123. {Y, C}	4
124. {Y, I}	9
125. {Y, L}	9
126. {Y, S}	5
127. {Y, V}	9

Itemsets	Suport Count
128. {A, L, I}	6
129. {A, L, W}	5
130. {E, A, L}	4
131. {E, A, Q}	4
132. {E, F, S}	4
133. {E, P, F}	4
134. {E, P, S}	4
135. {E, S, A}	5
136. {E, S, Q}	5
137. {E, V, A}	5
138. {E, V, S}	4
139. {F, A, I}	5
140. {F, A, L}	5
141. {F, S, A}	4
142. {F, S, I}	4
143. {F, V, A}	6
144. {F, V, I}	5
145. {F, V, L}	4
146. {F, Y, I}	5
147. {F, Y, L}	5

Itemsets	Suport Count
221. {T, F, A}	4
222. {T, F, H}	4
223. {T, F, I}	5
224. {T, F, L}	5
225. {T, F, V}	5

Itemsets	Suport Count
241. {Y, A, L}	5
242. {Y, L, I}	6
243. {Y, V, A}	4
244. {Y, V, C}	4
245. {Y, V, I}	6

Itemsets	Suport Count
261. {P, A, L, W}	4
262. {P, F, S, A}	4
263. {P, Y, L, I}	4
264. {S, A, L, W}	4
265. {V, A, L, I}	4

Itemsets	Suport Count
226. {T, F, Y}	6
227. {T, P, F}	6
228. {T, P, V}	4
229. {T, P, Y}	4
230. {T, V, A}	7
231. {T, V, K}	4
232. {T, V, Q}	5
233. {T, Y, L}	4
234. {T, Y, V}	4
235. {V, A, I}	7
236. {V, A, L}	6
237. {V, L, C}	4
238. {V, L, I}	4
239. {V, S, A}	5
240. {Y, A, I}	6

Itemsets	Suport Count
246. {Y, V, L}	5
247. {F, V, A, I}	4
248. {F, Y, V, I}	4
249. {G, F, V, A}	4
250. {G, T, F, A}	4
251. {G, T, F, I}	4
252. {G, T, F, V}	5
253. {G, T, F, Y}	5
254. {G, T, P, F}	5
255. {M, N, G, F}	4
256. {M, N, T, F}	4
257. {N, G, P, F}	5
258. {N, G, T, F}	6
259. {N, G, T, P}	4
260. {N, T, P, F}	4

Itemsets	Suport Count
266. {Y, A, L, I}	5
267. {Y, V, A, I}	4
268. {N, G, T, P, F}	4

Appendix-F

Generation of Strong Association Rules

Disease-2: Breast Cancer
(Protein: Breast Cancer Type 1 Susceptibility Protein)

Minimum Support Count Considered: 5

The list of accepted strong association rules (having minimum confidence 90%) generated from 1806 valid itemsets are shown in the following table:

Ser	Association Rule	Confidence		Ser	Association Rule	Confidence
1	AD -> E	100.00%		30	QRT -> L	100.00%
2	DH -> E	90.00%		31	LMS -> E	90.00%
3	MS -> E	93.30%		32	KMR -> E	100.00%
4	CV -> E	90.00%		33	GMS -> E	100.00%
5	ADG -> E	100.00%		34	MNS -> E	100.00%
6	ADK -> E	100.00%		35	MQS -> E	100.00%
7	ADN -> E	100.00%		36	MSV -> E	100.00%
8	ADP -> E	100.00%		37	ACK -> S	100.00%
9	ADQ -> E	100.00%		38	EPY -> S	100.00%
10	ADT -> E	100.00%		39	EQR -> S	90.00%
11	ADR -> E	100.00%		40	IKR -> S	100.00%
12	ADV -> E	100.00%		41	FKV -> S	90.00%
13	DGH -> E	100.00%		42	ANPT -> G	100.00%
14	ADL -> E	100.00%		43	ADGN -> E	100.00%
15	DHP -> L	100.00%		44	ADNT -> E	100.00%
16	DPY -> L	100.00%		45	ADLN -> E	100.00%
17	DFR -> E	100.00%		46	DEHP -> L	100.00%
18	DGR -> E	100.00%		47	ADLS -> E	100.00%
19	DNR -> E	100.00%		48	DHLS -> E	100.00%
20	DRT -> E	90.00%		49	DRTV -> E	100.00%
21	DRV -> E	100.00%		50	ADKS -> E	100.00%
22	ADS -> E	100.00%		51	ADNS -> E	100.00%
23	DHS -> E	100.00%		52	DGKS -> E	100.00%
24	DKS -> E	92.30%		53	DRST -> E	100.00%
25	DRS -> E	90.90%		54	DRSV -> E	100.00%
26	DSV -> E	100.00%		55	DKSV -> E	100.00%
27	DNV -> E	100.00%		56	GKLN -> P	100.00%
28	FLN -> P	100.00%		57	GLNT -> P	100.00%
29	NQR -> L	90.00%		58	ELPY -> S	100.00%

Ser	Association Rule	Confidence
59	FPST -> L	100.00%
60	GPST -> L	100.00%
61	ILQS -> N	100.00%
62	GQRS -> L	100.00%
63	NQRS -> L	100.00%
64	LRSV -> E	100.00%
65	EKQV -> L	100.00%
66	EGLM -> S	100.00%
67	GLMS -> E	100.00%
68	LMNS -> E	100.00%
69	LMQS -> E	100.00%
70	AEFN -> S	100.00%
71	NQST -> P	100.00%
72	PRSV -> E	100.00%
73	EGKV -> S	100.00%
74	EINV -> S	100.00%
75	EGQV -> S	100.00%
76	IKPV -> S	100.00%
77	IPSV -> K	100.00%
78	LNQST -> P	100.00%
79	EGLQV -> S	100.00%
80	EKQSV -> L	100.00%

Appendix-G

Generation of Strong Association Rules

Disease-3: Cystic Fibrosis
(Protein: Cystic Fibrosis Transmembrane Conductance Regulator)

Minimum Support Count Considered: 5

The list of accepted strong association rules (having minimum confidence 90%) generated from 1464 valid itemsets are shown in the following table:

Ser	Association Rule	Confidence		Ser	Association Rule	Confidence
1	AG -> L	90.00%		30	DIM -> S	100.00%
2	DT -> L	91.70%		31	AMS -> L	100.00%
3	HV -> L	90.90%		32	APW -> V	100.00%
4	NW -> L	100.00%		33	PVW -> A	100.00%
5	TW -> L	90.00%		34	KPY -> L	100.00%
6	AM -> L	92.90%		35	PVY -> L	100.00%
7	PY -> L	100.00%		36	AQW -> V	100.00%
8	QY -> L	91.70%		37	FGQ -> I	100.00%
9	FIN -> K	100.00%		38	EGQ -> L	100.00%
10	ADG -> L	100.00%		39	FQY -> L	100.00%
11	ADT -> L	100.00%		40	IQY -> L	100.00%
12	AGT -> L	100.00%		41	QTY -> L	100.00%
13	ANW -> L	100.00%		42	EPQ -> L	100.00%
14	AEG -> L	100.00%		43	KPQ -> L	91.70%
15	DET -> L	100.00%		44	PQY -> L	100.00%
16	DFG -> L	100.00%		45	EQR -> L	90.00%
17	FHT -> L	100.00%		46	QSY -> L	100.00%
18	GIY -> L	100.00%		47	ACR -> L	100.00%
19	DIY -> L	100.00%		48	CLR -> A	100.00%
20	IKY -> L	100.00%		49	IRY -> L	100.00%
21	KTY -> L	100.00%		50	RTY -> L	100.00%
22	DIT -> L	100.00%		51	FPR -> V	100.00%
23	DGV -> L	100.00%		52	PRT -> V	100.00%
24	DTV -> L	100.00%		53	DIR -> S	90.00%
25	HTV -> L	100.00%		54	HKR -> S	100.00%
26	AIM -> L	100.00%		55	ADN -> S	100.00%
27	AKM -> L	100.00%		56	HKV -> S	100.00%
28	FMR -> I	100.00%		57	AGS -> L	100.00%
29	AMR -> L	100.00%		58	STW -> L	100.00%

Ser	Association Rule	Confidence
59	APS -> L	90.00%
60	AFLV -> I	100.00%
61	ADTV -> L	100.00%
62	AGTV -> L	100.00%
63	AKTV -> L	100.00%
64	HILV -> T	100.00%
65	HITV -> L	100.00%
66	AKMS -> L	100.00%
67	FILP -> V	100.00%
68	EGIQ -> L	100.00%
69	EKLP -> Q	100.00%
70	EKPQ -> L	100.00%
71	LPQR -> K	100.00%
72	ALQR -> S	100.00%
73	KPQS -> L	100.00%
74	FIPR -> V	100.00%
75	DLRS -> I	100.00%
76	IRSY -> L	100.00%
77	ADKS -> I	100.00%
78	AIKN -> S	100.00%
79	HIKT -> S	100.00%
80	DIKV -> S	100.00%
81	DKSV -> I	100.00%
82	HIKV -> S	100.00%
83	HISV -> K	100.00%
84	AGIS -> L	100.00%
85	AGSV -> L	100.00%
86	DISY -> L	100.00%
87	AGKS -> L	100.00%
88	HKLS -> V	100.00%
89	HKLV -> S	100.00%
90	IKLV -> S	90.90%
91	HILS -> T	100.00%
92	DGSV -> L	100.00%
93	HSTV -> L	100.00%
94	APSV -> L	100.00%
95	LPST -> V	100.00%
96	IKLTV -> S	100.00%
59	APS -> L	90.00%
60	AFLV -> I	100.00%

Appendix-H

Generation of Strong Association Rules

Disease-4: Nephrogenic Diabetes Insipidus
(Protein: Vasopressin V2 Receptor)

Minimum Support Count Considered: 4

The list of accepted strong association rules (having minimum confidence 90%) generated from 234 valid itemsets are shown in the following table:

Ser	Association Rule	Confidence	Ser	Association Rule	Confidence
1	K -> A	100.00%	30	GI -> V	100.00%
2	N -> S	100.00%	31	FPW -> A	100.00%
3	FW -> A	100.00%	32	AFG -> P	100.00%
4	FG -> A	100.00%	33	FG -> AP	100.00%
5	GI -> A	100.00%	34	FGP -> A	100.00%
6	HR -> A	100.00%	35	PRT -> A	100.00%
7	FI -> A	100.00%	36	FVW -> A	100.00%
8	AE -> P	100.00%	37	AGI -> V	100.00%
9	FP -> A	100.00%	38	GI -> AV	100.00%
10	IP -> A	100.00%	39	GIV -> A	100.00%
11	PW -> A	100.00%	40	FPV -> A	100.00%
12	CR -> A	100.00%	41	GPV -> A	100.00%
13	FS -> A	100.00%	42	PVW -> A	100.00%
14	IS -> A	100.00%	43	AEL -> P	100.00%
15	AN -> S	100.00%	44	FLP -> A	100.00%
16	CV -> A	100.00%	45	ACV -> L	100.00%
17	FV -> A	100.00%	46	CLV -> A	100.00%
18	HV -> A	100.00%	47	CV -> AL	100.00%
19	PV -> A	100.00%	48	FLV -> A	100.00%
20	VW -> A	100.00%	49	GLV -> A	100.00%
21	FL -> A	100.00%	50	LPV -> A	100.00%
22	AQ -> L	100.00%	51	DEL -> P	100.00%
23	LW -> A	100.00%	52	DLP -> E	100.00%
24	PQ -> L	100.00%	53	AMT -> L	100.00%
25	CV -> L	100.00%	54	FLPV -> A	100.00%
26	MR -> L	100.00%			
27	MT -> L	100.00%			
28	DE -> P	100.00%			
29	FG -> P	100.00%			

Appendix-I

Generation of Strong Association Rules

Disease-5: Retinitis Pigmentosa 4
(Protein: Rhodopsin)

Minimum Support Count Considered: 4

The list of accepted strong association rules (having minimum confidence 90%) generated from 268 valid itemsets are shown in the following table:

Ser	Association Rule	Confidence
1	W -> A	100.00%
2	W -> L	100.00%
3	H -> T	100.00%
4	AW -> L	100.00%
5	LW -> A	100.00%
6	W -> AL	100.00%
7	QS -> E	100.00%
8	GT -> F	100.00%
9	CI -> L	100.00%
10	EM -> F	100.00%
11	MS -> F	100.00%
12	GM -> F	100.00%
13	NY -> P	100.00%
14	PW -> A	100.00%
15	PW -> L	100.00%
16	SW -> A	100.00%
17	SW -> L	100.00%
18	FH -> T	100.00%
19	KV -> T	100.00%
20	QV -> T	100.00%
21	AY -> I	100.00%
22	CY -> V	100.00%
23	AFT -> G	100.00%
24	AGT -> F	100.00%
25	FGI -> T	100.00%
26	GIT -> F	100.00%
27	FTV -> G	100.00%
28	GTV -> F	100.00%
29	FGY -> T	100.00%
30	GTY -> F	100.00%
31	GPT -> F	100.00%
32	GMN -> F	100.00%
33	MNT -> F	100.00%
34	FNP -> G	100.00%
35	GNT -> F	100.00%
36	APW -> L	100.00%
37	LPW -> A	100.00%
38	PW -> AL	100.00%
39	AFP -> S	100.00%
40	AFS -> P	100.00%
41	ALS -> W	100.00%
42	ASW -> L	100.00%
43	LSW -> A	100.00%
44	SW -> AL	100.00%
45	ILV -> A	100.00%
46	ALY -> I	100.00%
47	AVY -> I	100.00%
48	FNPT -> G	100.00%
49	GNPT -> F	100.00%

Appendix-J

Generation of Useful Strong Association Rules

Disease-1: Sickle Cell Anemia
(Protein: Hemoglobin Subunit Beta)

Minimum Support Count Considered: 3

The list of useful strong association rules generated from 135 valid itemsets are shown in the following table:

Ser	Association Rules	*Lift*	*Bi-lift*	*Bi-improve*	*Bi-confidence*
1	GT -> AN	3.75	12	0.183	0.917
2	GT -> KN	3.75	12	0.183	0.917
3	AGT -> KN	3.75	12	0.183	0.917
4	GKT -> AN	3.75	12	0.183	0.917
5	GT -> AKN	3.75	12	0.183	0.917
6	AN -> GK	3	11	0.242	0.909
7	GS -> FL	3	6	0.167	0.833
8	NT -> GK	3	6	0.167	0.833
9	KP -> TV	3	6	0.167	0.833
10	ANT -> GK	3	6	0.167	0.833
11	NT -> AGK	3	6	0.167	0.833
12	ANV -> GK	3	6	0.167	0.833
13	GT -> N	2.5	4	0.15	0.75
14	AGT -> N	2.5	4	0.15	0.75
15	GKT -> N	2.5	4	0.15	0.75
16	AGKT -> N	2.5	4	0.15	0.75
17	KP -> T	2.143	3	0.133	0.667
18	GH -> AL	2.143	3	0.133	0.667
19	GT -> AK	2.143	3	0.133	0.667
20	NT -> AK	2.143	3	0.133	0.667
21	KPV -> T	2.143	3	0.133	0.667
22	GNT -> AK	2.143	3	0.133	0.667
23	KN -> AG	1.875	2.75	0.17	0.636
24	GS -> F	1.875	2.4	0.117	0.583
25	FS -> GL	1.875	2.4	0.117	0.583
26	GLS -> F	1.875	2.4	0.117	0.583
27	NT -> AG	1.875	2.4	0.117	0.583
28	KNT -> AG	1.875	2.4	0.117	0.583
29	KNV -> AG	1.875	2.4	0.117	0.583
30	AN -> K	1.364	1.571	0.097	0.364
31	AT -> K	1.364	1.571	0.097	0.364
32	AGN -> K	1.364	1.571	0.097	0.364

Useful Strong Association Rules

Ser	Association Rules	*Lift*	*Bi-lift*	*Bi-improve*	*Bi-confidence*	
33	GT -> K	1.364	1.5	0.067	0.333	
34	NT -> K	1.364	1.5	0.067	0.333	
35	AGT -> K	1.364	1.5	0.067	0.333	
36	ANT -> K	1.364	1.5	0.067	0.333	
37	GNT -> K	1.364	1.5	0.067	0.333	
38	ANV -> K	1.364	1.5	0.067	0.333	
39	ATV -> K	1.364	1.5	0.067	0.333	
40	AGNT -> K	1.364	1.5	0.067	0.333	
41	AGNV -> K	1.364	1.5	0.067	0.333	
42	FL -> G	1.154	1.25	0.067	0.2	
43	AN -> G	1.154	1.222	0.048	0.182	
44	KN -> G	1.154	1.222	0.048	0.182	
45	NV -> G	1.154	1.222	0.048	0.182	
46	AKN -> G	1.154	1.222	0.048	0.182	
47	ALV -> G	1.154	1.222	0.048	0.182	
48	AD -> G	1.154	1.2	0.033	0.167	
49	LN -> G	1.154	1.2	0.033	0.167	
50	FS -> G	1.154	1.2	0.033	0.167	
51	NT -> G	1.154	1.2	0.033	0.167	
52	AFL -> G	1.154	1.2	0.033	0.167	
53	FLS -> G	1.154	1.2	0.033	0.167	
54	ANT -> G	1.154	1.2	0.033	0.167	
55	KNT -> G	1.154	1.2	0.033	0.167	
56	ANV -> G	1.154	1.2	0.033	0.167	
57	KNV -> G	1.154	1.2	0.033	0.167	
58	AKNT -> G	1.154	1.2	0.033	0.167	
59	AKNV -> G	1.154	1.2	0.033	0.167	
60	GH -> A	1	1	0	0	Redundant Rules
61	GK -> A	1	1	0	0	
62	KN -> A	1	1	0	0	
63	GT -> A	1	1	0	0	
64	NT -> A	1	1	0	0	
65	GHL -> A	1	1	0	0	
66	GKN -> A	1	1	0	0	
67	GKL -> A	1	1	0	0	
68	GNT -> A	1	1	0	0	
69	GKT -> A	1	1	0	0	
70	KNT -> A	1	1	0	0	
71	GKV -> A	1	1	0	0	
72	KNV -> A	1	1	0	0	
73	GKNT -> A	1	1	0	0	
74	GKNV -> A	1	1	0	0	
75	R -> L	0.833	0.8	-0.05	-0.25	
76	Q -> V	0.833	0.8	-0.05	-0.25	
77	GH -> L	0.833	0.8	-0.05	-0.25	

Ser	Association Rules	*Lift*	*Bi-lift*	*Bi-improve*	*Bi-confidence*	
78	DE -> L	0.833	0.8	-0.05	-0.25	
79	EG -> L	0.833	0.8	-0.05	-0.25	
80	FS -> L	0.833	0.8	-0.05	-0.25	
81	GS -> L	0.833	0.8	-0.05	-0.25	
82	EK -> V	0.833	0.8	-0.05	-0.25	
83	LP -> V	0.833	0.8	-0.05	-0.25	
84	EP -> V	0.833	0.8	-0.05	-0.25	
85	KP -> V	0.833	0.8	-0.05	-0.25	
86	AGH -> L	0.833	0.8	-0.05	-0.25	
87	AFG -> L	0.833	0.8	-0.05	-0.25	
88	FGS -> L	0.833	0.8	-0.05	-0.25	
89	HKV -> L	0.833	0.8	-0.05	-0.25	
90	KPT -> V	0.833	0.8	-0.05	-0.25	
91	AH -> L	0.833	0.786	-0.073	-0.273	
92	HK -> L	0.833	0.786	-0.073	-0.273	
93	HV -> L	0.833	0.786	-0.073	-0.273	
94	PT -> V	0.833	0.786	-0.073	-0.273	
95	FG -> L	0.833	0.769	-0.1	-0.3	

Appendix-K

Generation of Useful Strong Association Rules

Disease-2: Breast Cancer
(Protein: Breast Cancer Type 1 Susceptibility Protein)

Minimum Support Count Considered: 5

The list of useful strong association rules generated from 1806 valid itemsets is shown in the following table:

Ser	Association Rules	*Lift*	*Bi-lift*	*Bi-improve*	*Bi-confidence*	
1	ANPT -> G	2.149	2.235	0.018	0.552	Useful Strong Association Rules
2	NQST -> P	1.948	2.011	0.016	0.503	
3	FLN -> P	1.948	2.0	0.013	0.5	
4	GKLN -> P	1.948	2.0	0.013	0.5	
5	GLNT -> P	1.948	2.0	0.013	0.5	
6	LNQST -> P	1.948	2.0	0.013	0.5	
7	ILQS -> N	1.545	1.569	0.01	0.363	
8	IPSV -> K	1.365	1.379	0.007	0.275	
9	EKQV -> L	1.199	1.208	0.006	0.172	
10	DHP -> L	1.199	1.207	0.005	0.171	
11	QRT -> L	1.199	1.207	0.005	0.171	
12	GPST -> L	1.199	1.207	0.005	0.171	
13	GQRS -> L	1.199	1.207	0.005	0.171	
14	NQRS -> L	1.199	1.207	0.005	0.171	
15	DPY -> L	1.199	1.205	0.005	0.17	
16	DEHP -> L	1.199	1.205	0.005	0.17	
17	FPST -> L	1.199	1.205	0.005	0.17	
18	EKQSV -> L	1.199	1.205	0.005	0.17	
19	NQR -> L	1.079	1.084	0.004	0.069	
20	ADR -> E	0.944	0.943	-0.002	-0.06	Redundant Rules
21	ADV -> E	0.944	0.943	-0.002	-0.06	
22	DGH -> E	0.944	0.943	-0.002	-0.06	
23	DFR -> E	0.944	0.943	-0.002	-0.06	
24	DGR -> E	0.944	0.943	-0.002	-0.06	
25	KMR -> E	0.944	0.943	-0.002	-0.06	
26	ADGN -> E	0.944	0.943	-0.002	-0.06	
27	ADLN -> E	0.944	0.943	-0.002	-0.06	
28	DHLS -> E	0.944	0.943	-0.002	-0.06	
29	DRTV -> E	0.944	0.943	-0.002	-0.06	

Ser	Association Rules	*Lift*	*Bi-lift*	*Bi-improve*	*Bi-confidence*
30	ADKS -> E	0.944	0.943	-0.002	-0.06
31	DGKS -> E	0.944	0.943	-0.002	-0.06
32	DRST -> E	0.944	0.943	-0.002	-0.06
33	DRSV -> E	0.944	0.943	-0.002	-0.06
34	DKSV -> E	0.944	0.943	-0.002	-0.06
35	GLMS -> E	0.944	0.943	-0.002	-0.06
36	LMNS -> E	0.944	0.943	-0.002	-0.06
37	LMQS -> E	0.944	0.943	-0.002	-0.06
38	ADG -> E	0.944	0.942	-0.002	-0.061
39	ADK -> E	0.944	0.943	-0.002	-0.061
40	ADP -> E	0.944	0.943	-0.002	-0.061
41	ADQ -> E	0.944	0.943	-0.002	-0.061
42	DNR -> E	0.944	0.943	-0.002	-0.061
43	DHS -> E	0.944	0.943	-0.002	-0.061
44	DNV -> E	0.944	0.943	-0.002	-0.061
45	GMS -> E	0.944	0.943	-0.002	-0.061
46	MNS -> E	0.944	0.942	-0.002	-0.061
47	MQS -> E	0.944	0.942	-0.002	-0.061
48	MSV -> E	0.944	0.943	-0.002	-0.061
49	ADNT -> E	0.944	0.942	-0.002	-0.061
50	ADLS -> E	0.944	0.942	-0.002	-0.061
51	ADNS -> E	0.944	0.942	-0.002	-0.061
52	LRSV -> E	0.944	0.943	-0.002	-0.061
53	PRSV -> E	0.944	0.942	-0.002	-0.061
54	ADT -> E	0.944	0.941	-0.003	-0.062
55	DRV -> E	0.944	0.942	-0.003	-0.062
56	DSV -> E	0.944	0.942	-0.003	-0.062
57	ADN -> E	0.944	0.941	-0.004	-0.063
58	ADL -> E	0.944	0.941	-0.004	-0.063
59	ADS -> E	0.944	0.941	-0.004	-0.063
60	AD -> E	0.944	0.938	-0.007	-0.066
61	MS -> E	0.881	0.872	-0.011	-0.136
62	DKS -> E	0.872	0.864	-0.01	-0.146
63	DRS -> E	0.859	0.851	-0.009	-0.159
64	DH -> E	0.85	0.843	-0.009	-0.168
65	CV -> E	0.85	0.843	-0.009	-0.168
66	DRT -> E	0.85	0.843	-0.009	-0.168
67	LMS -> E	0.85	0.843	-0.009	-0.168
68	ACK -> S	0.835	0.831	-0.005	-0.203
69	IKR -> S	0.835	0.831	-0.005	-0.203
70	ELPY -> S	0.835	0.831	-0.005	-0.203

Ser	Association Rules	*Lift*	*Bi-lift*	*Bi-improve*	*Bi-confidence*	
71	EGLM -> S	0.835	0.831	-0.005	-0.203	
72	AEFN -> S	0.835	0.831	-0.005	-0.203	
73	EINV -> S	0.835	0.831	-0.005	-0.203	
74	IKPV -> S	0.835	0.831	-0.005	-0.203	
75	EGLQV -> S	0.835	0.831	-0.005	-0.203	
76	EPY -> S	0.835	0.83	-0.007	-0.204	
77	EGQV -> S	0.835	0.83	-0.007	-0.204	
78	EGKV -> S	0.835	0.829	-0.008	-0.206	
79	EQR -> S	0.751	0.741	-0.017	-0.315	
80	FKV -> S	0.751	0.741	-0.017	-0.315	

Appendix-L

Generation of Useful Strong Association Rules

Disease-3: Cystic Fibrosis
(Cystic Fibrosis Transmembrane Conductance Regulator)

Minimum Support Count Considered: 5

The list of useful strong association rules generated from 1464 valid itemsets is shown in the following table:

Ser	Association Rules	*Lift*	*Bi-lift*	*Bi-improve*	*Bi-confidence*
1	EKLP -> Q	2.209	2.328	0.023	0.57
2	PVW -> A	1.783	1.833	0.015	0.455
3	CLR -> A	1.783	1.833	0.015	0.455
4	HILV -> T	1.783	1.833	0.015	0.455
5	HILS -> T	1.783	1.833	0.015	0.455
6	FPR -> V	1.644	1.707	0.022	0.414
7	FIPR -> V	1.644	1.69	0.017	0.408
8	APW -> V	1.644	1.682	0.014	0.406
9	AQW -> V	1.644	1.682	0.014	0.406
10	PRT -> V	1.644	1.682	0.014	0.406
11	FILP -> V	1.644	1.682	0.014	0.406
12	HKLS -> V	1.644	1.682	0.014	0.406
13	LPST -> V	1.644	1.682	0.014	0.406
14	FIN -> K	1.609	1.644	0.013	0.392
15	LPQR -> K	1.609	1.644	0.013	0.392
16	HISV -> K	1.609	1.644	0.013	0.392
17	DLRS -> I	1.244	1.259	0.01	0.206
18	AFLV -> I	1.244	1.257	0.008	0.204
19	DKSV -> I	1.244	1.257	0.008	0.204
20	FMR -> I	1.244	1.254	0.007	0.203
21	FGQ -> I	1.244	1.254	0.007	0.203
22	ADKS -> I	1.244	1.254	0.007	0.203
23	ALQR -> S	1.203	1.216	0.008	0.177
24	HKV -> S	1.203	1.214	0.007	0.176
25	DIKV -> S	1.203	1.214	0.007	0.176
26	DIM -> S	1.203	1.212	0.006	0.175
27	HKR -> S	1.203	1.212	0.006	0.175
28	ADN -> S	1.203	1.212	0.006	0.175
29	AIKN -> S	1.203	1.212	0.006	0.175

Useful Strong Association Rules

30	HIKT -> S	1.203	1.212	0.006	0.175	
31	HIKV -> S	1.203	1.212	0.006	0.175	
32	HKLV -> S	1.203	1.212	0.006	0.175	
33	IKLTV -> S	1.203	1.212	0.006	0.175	
34	IKLV -> S	1.094	1.102	0.006	0.084	
35	DIR -> S	1.083	1.089	0.005	0.074	
36	ANW -> L	0.809	0.803	-0.008	-0.245	Redundant Rules
37	DET -> L	0.809	0.803	-0.008	-0.245	
38	DIT -> L	0.809	0.803	-0.008	-0.245	
39	KPY -> L	0.809	0.803	-0.008	-0.245	
40	PVY -> L	0.809	0.803	-0.008	-0.245	
41	FQY -> L	0.809	0.803	-0.008	-0.245	
42	QTY -> L	0.809	0.803	-0.008	-0.245	
43	PQY -> L	0.809	0.803	-0.008	-0.245	
44	QSY -> L	0.809	0.803	-0.008	-0.245	
45	ACR -> L	0.809	0.803	-0.008	-0.245	
46	RTY -> L	0.809	0.803	-0.008	-0.245	
47	STW -> L	0.809	0.803	-0.008	-0.245	
48	ADTV -> L	0.809	0.803	-0.008	-0.245	
49	AGTV -> L	0.809	0.803	-0.008	-0.245	
50	AKTV -> L	0.809	0.803	-0.008	-0.245	
51	HITV -> L	0.809	0.803	-0.008	-0.245	
52	AKMS -> L	0.809	0.803	-0.008	-0.245	
53	EGIQ -> L	0.809	0.803	-0.008	-0.245	
54	IRSY -> L	0.809	0.803	-0.008	-0.245	
55	AGSV -> L	0.809	0.803	-0.008	-0.245	
56	DISY -> L	0.809	0.803	-0.008	-0.245	
57	AGKS -> L	0.809	0.803	-0.008	-0.245	
58	DGSV -> L	0.809	0.803	-0.008	-0.245	
59	ADG -> L	0.809	0.802	-0.01	-0.246	
60	ADT -> L	0.809	0.802	-0.01	-0.246	
61	AEG -> L	0.809	0.802	-0.01	-0.246	
62	DFG -> L	0.809	0.802	-0.01	-0.246	
63	FHT -> L	0.809	0.802	-0.01	-0.246	
64	GIY -> L	0.809	0.802	-0.01	-0.246	
65	DIY -> L	0.809	0.802	-0.01	-0.246	
66	KTY -> L	0.809	0.802	-0.01	-0.246	
67	AIM -> L	0.809	0.802	-0.01	-0.246	
68	AKM -> L	0.809	0.802	-0.01	-0.246	
69	IQY -> L	0.809	0.802	-0.01	-0.246	
70	EKPQ -> L	0.809	0.802	-0.01	-0.246	

71	KPQS -> L	0.809	0.802	-0.01	-0.246	
72	AGIS -> L	0.809	0.802	-0.01	-0.246	
73	HSTV -> L	0.809	0.802	-0.01	-0.246	
74	APSV -> L	0.809	0.802	-0.01	-0.246	
75	NW -> L	0.809	0.801	-0.012	-0.248	
76	AGT -> L	0.809	0.801	-0.012	-0.248	
77	IKY -> L	0.809	0.801	-0.012	-0.248	
78	DGV -> L	0.809	0.801	-0.012	-0.248	
79	DTV -> L	0.809	0.801	-0.012	-0.248	
80	AMR -> L	0.809	0.801	-0.012	-0.248	
81	AMS -> L	0.809	0.801	-0.012	-0.248	
82	EGQ -> L	0.809	0.801	-0.012	-0.248	
83	IRY -> L	0.809	0.801	-0.012	-0.248	
84	PY -> L	0.809	0.8	-0.014	-0.25	
85	HTV -> L	0.809	0.8	-0.014	-0.25	
86	EPQ -> L	0.809	0.8	-0.014	-0.25	
87	AGS -> L	0.809	0.797	-0.019	-0.255	
88	AM -> L	0.751	0.732	-0.032	-0.34	
89	DT -> L	0.741	0.725	-0.028	-0.348	
90	QY -> L	0.741	0.725	-0.028	-0.348	
91	KPQ -> L	0.741	0.725	-0.028	-0.348	
92	HV -> L	0.735	0.72	-0.026	-0.354	
93	TW -> L	0.728	0.714	-0.024	-0.361	
94	EQR -> L	0.728	0.714	-0.024	-0.361	
95	APS -> L	0.728	0.714	-0.024	-0.361	
96	AG -> L	0.728	0.698	-0.053	-0.389	

Appendix-M

Generation of Useful Strong Association Rules

Disease-4: Nephrogenic Diabetes Insipidus
(Protein: Vasopressin V2 Receptor)

Minimum Support Count Considered: 4

The list of useful strong association rules generated from 234 valid itemsets is shown in the following table:

Ser	Association Rules	*Lift*	*Bi-lift*	*Bi-improve*	*Bi-confidence*	
1	DLP -> E	3.455	4.857	0.084	0.794	Useful Strong Association Rules
2	FG -> AP	2.375	2.833	0.068	0.647	
3	GI -> AV	2.235	2.615	0.065	0.618	
4	CV -> AL	1.9	2.125	0.056	0.529	
5	AE -> P	1.462	1.6	0.059	0.375	
6	DE -> P	1.462	1.571	0.048	0.364	
7	FG -> P	1.462	1.545	0.037	0.353	
8	AFG -> P	1.462	1.545	0.037	0.353	
9	AEL -> P	1.462	1.545	0.037	0.353	
10	DEL -> P	1.462	1.545	0.037	0.353	
11	GI -> V	1.267	1.308	0.025	0.235	
12	AGI -> V	1.267	1.308	0.025	0.235	
13	N -> S	1.086	1.103	0.015	0.094	
14	AN -> S	1.086	1.097	0.009	0.088	
15	K -> A	0.809	0.791	-0.028	-0.265	Redundant Rules
16	FG -> A	0.809	0.791	-0.028	-0.265	
17	GI -> A	0.809	0.791	-0.028	-0.265	
18	HR -> A	0.809	0.791	-0.028	-0.265	
19	CR -> A	0.809	0.791	-0.028	-0.265	
20	FS -> A	0.809	0.791	-0.028	-0.265	
21	CV -> A	0.809	0.791	-0.028	-0.265	
22	HV -> A	0.809	0.791	-0.028	-0.265	
23	LW -> A	0.809	0.791	-0.028	-0.265	
24	FGP -> A	0.809	0.791	-0.028	-0.265	
25	PRT -> A	0.809	0.791	-0.028	-0.265	
26	FVW -> A	0.809	0.791	-0.028	-0.265	
27	GIV -> A	0.809	0.791	-0.028	-0.265	
28	PVW -> A	0.809	0.791	-0.028	-0.265	
29	CLV -> A	0.809	0.791	-0.028	-0.265	

30	GLV -> A	0.809	0.791	-0.028	-0.265	
31	FLPV -> A	0.809	0.791	-0.028	-0.265	
32	FI -> A	0.809	0.786	-0.036	-0.273	
33	IP -> A	0.809	0.786	-0.036	-0.273	
34	IS -> A	0.809	0.786	-0.036	-0.273	
35	FPW -> A	0.809	0.786	-0.036	-0.273	
36	FPV -> A	0.809	0.786	-0.036	-0.273	
37	GPV -> A	0.809	0.786	-0.036	-0.273	
38	FLP -> A	0.809	0.786	-0.036	-0.273	
39	LPV -> A	0.809	0.786	-0.036	-0.273	
40	FW -> A	0.809	0.78	-0.044	-0.281	
41	VW -> A	0.809	0.78	-0.044	-0.281	
42	PW -> A	0.809	0.775	-0.053	-0.29	
43	PV -> A	0.809	0.775	-0.053	-0.29	
44	FLV -> A	0.809	0.775	-0.053	-0.29	
45	FV -> A	0.809	0.769	-0.063	-0.3	
46	FL -> A	0.809	0.769	-0.063	-0.3	
47	FP -> A	0.809	0.763	-0.074	-0.31	
48	CV -> L	0.776	0.756	-0.034	-0.324	
49	MR -> L	0.776	0.756	-0.034	-0.324	
50	ACV -> L	0.776	0.756	-0.034	-0.324	
51	AMT -> L	0.776	0.756	-0.034	-0.324	
52	AQ -> L	0.776	0.75	-0.044	-0.333	
53	PQ -> L	0.776	0.75	-0.044	-0.333	
54	MT -> L	0.776	0.75	-0.044	-0.333	

Appendix-N

Generation of Useful Strong Association Rules

Disease-5: Retinitis Pigmentosa 4
(Rhodopsin)

Minimum Support Count Considered: 4

The list of useful strong association rules generated from 268 valid itemsets is shown in the following table:

Ser	Association Rules	*Lift*	*Bi-lift*	*Bi-improve*	*Bi-confidence*	
1	ALS -> W	7	31	0.111	0.968	Useful Strong Association Rules
2	W -> AL	3.5	6	0.119	0.833	
3	PW -> AL	3.5	5.167	0.092	0.806	
4	SW -> AL	3.5	5.167	0.092	0.806	
5	QS -> E	2.188	2.727	0.09	0.633	
6	AFP -> S	2.059	2.385	0.066	0.581	
7	NY -> P	1.75	2	0.071	0.500	
8	AFS -> P	1.75	1.938	0.055	0.484	
9	AFT -> G	1.591	1.722	0.048	0.419	
10	FTV -> G	1.591	1.765	0.062	0.433	
11	FNP -> G	1.591	1.765	0.062	0.433	
12	FNPT -> G	1.591	1.722	0.048	0.419	
13	H -> T	1.458	1.579	0.052	0.367	
14	FH -> T	1.458	1.55	0.041	0.355	
15	KV -> T	1.458	1.55	0.041	0.355	
16	QV -> T	1.458	1.579	0.052	0.367	
17	AY -> I	1.458	1.611	0.065	0.379	
18	FGI -> T	1.458	1.55	0.041	0.355	
19	FGY -> T	1.458	1.579	0.052	0.367	
20	ALY -> I	1.458	1.579	0.052	0.367	
21	AVY -> I	1.458	1.55	0.041	0.355	
22	W -> L	1.207	1.25	0.029	0.200	
23	AW -> L	1.207	1.25	0.029	0.200	
24	CI -> L	1.207	1.24	0.022	0.194	
25	PW -> L	1.207	1.24	0.022	0.194	
26	SW -> L	1.207	1.24	0.022	0.194	
27	APW -> L	1.207	1.24	0.022	0.194	
28	ASW -> L	1.207	1.24	0.022	0.194	
29	GT -> F	1.167	1.238	0.049	0.192	

30	EM -> F	1.167	1.192	0.018	0.161	
31	MS -> F	1.167	1.192	0.018	0.161	
32	GM -> F	1.167	1.2	0.024	0.167	
33	CY -> V	1.167	1.192	0.018	0.161	
34	AGT -> F	1.167	1.192	0.018	0.161	
35	GIT -> F	1.167	1.192	0.018	0.161	
36	GTV -> F	1.167	1.2	0.024	0.167	
37	GTY -> F	1.167	1.2	0.024	0.167	
38	GPT -> F	1.167	1.2	0.024	0.167	
39	GMN -> F	1.167	1.192	0.018	0.161	
40	MNT -> F	1.167	1.192	0.018	0.161	
41	GNT -> F	1.167	1.208	0.03	0.172	
42	GNPT -> F	1.167	1.192	0.018	0.161	
43	W -> A	1.094	1.111	0.014	0.100	
44	LW -> A	1.094	1.111	0.014	0.100	
45	PW -> A	1.094	1.107	0.011	0.097	
46	SW -> A	1.094	1.107	0.011	0.097	
47	LPW -> A	1.094	1.107	0.011	0.097	
48	LSW -> A	1.094	1.107	0.011	0.097	
49	ILV -> A	1.094	1.107	0.011	0.097	