

Use Apriori, Genetic Algorithm, and Fuzzy Logic to Foretell the Most Common Amino Acid Sequence

Krupali Patel¹, Pravinbhai Patel² ¹Department of Genetic, ²Department of Computer Science Sardar Patel Unversity, Anand, Gujarat, India

ABSTRACT

Data mining is the practice of discovering connections between seemingly unrelated pieces of biological information. Rapid progress in genomics and proteomics in recent years has resulted in an abundance of biological data. Thus, categorising biological sequences and structures according to essential properties and functions is a pressing issue in the field of biological data processing. Many methods have been used to generate recurrent patterns from published works for use in a wide range of contexts. The frequency with which this algorithm was produced has diminished. Because of this, it's completely pointless. In this case, I want to use two different methods to compare the common pattern and optimise the data. Hence, we find it to be of great value. The contaminated protein sequence is the root cause of several human illnesses, and our method is designed to extract the amino acids that are both hidden and most dominant in the sequence. We deal with this issue by employing a combination of the apriori algorithm, the genetic algorithm, and strong association rules for pattern prediction. Apply fuzzy logic to the optimisation of data and the identification of intriguing common patterns in the protein sequence database. This Recurring Pattern is quite helpful in the Pharmaceutical Industry.

Keywords: Protein structure analysis, Genetic Algorithms, Association methods, Fuzzy Systems for mining biological data

1. Introduction

In order to efficiently and automatically uncover patterns in massive data sets, a set of methods known as "Data Mining" has been developed. Data mining, also known as Knowledge Discovery in Databases [1,2], is discovering helpful information hidden inside massive datasets housed in centralized locations.

When it comes to gaining insight from data, data mining encompasses every step taken to use computational methods, such as cutting-edge research in this area. Despite the abundance of data, a unified explanation of the molecular arrangement of life remains elusive, making Data Mining techniques appear well-suited for Biological Data Mining. The massive amounts of biological data are both problems and possibilities for KDD researchers. Knowledge in medicine, neurology, and allied life sciences can be aided by mining biological data [3].

Putting that information to use is a significant problem for the future of biological science [4]. The current goal of the life sciences is to advance fundamental theories, biotechnology, and medicine by planning large-scale experiments, collecting massive amounts of data, analyzing said data, comparing said data, and ultimately combining said data. Data characterization, data differentiation, association scrutiny

categorization, calculation, grouping, outliers, and association regulation mining are all possible results of employing data mining methods.

1.1 Data Pre-processing

Data mining refers to the process of collecting and analyzing data to conclusions. Many methods exist for preparing data for analysis, including (i) records cleansing, (ii) records processing, (iii) feature extraction, and (iv) subsampling. The plan's implementation included preparing the biological information for Recognizing Patterns.

1.2 Data Reduction

It is a compact version of the original data set that allows for helpful analysis [3]. The following methods are used to accomplish the task of data reduction;

- In order to build the data cube, aggregation procedures must be applied to the raw data.
- When data is unimportant, has little relevance, is redundant, or might have dimensions omitted, attribute subset selection is used.
- The size of a data collection may be decreased by the application of dimensionality reduction.
- Data may be approximated or updated with better precision after undergoing numerosity reduction.

2. Bio-Data mining

Bio-Data mining [3] refers to creating computational tools for use in the biological sciences. Variations in biomedical study, biotechnology, and an explosion of biomedical statistics over the past two decades include, but are not limited to, data gathered from research in the pharmaceutical industry and cancer therapy research; data uncovered in proteomic and genomics studies [5]; and data gleaned from the identification of sequential arrangements, gene roles, and protein-protein connections.

In bioinformatics, there are several awe-inspiring fields of study.

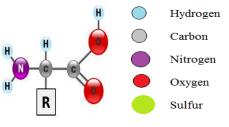
- Analyzing sequences
- Simulating biological processes
- Annotating the Genome
- Membrane protein analysis
- Cancer mutation analysis
- Calculating the 3D structure of proteins
- Interaction between proteins using docking
- Genomic comparisons Etc.

Disease diagnosis, prognosis, therapy optimization, data purification, and sub-cellular location prediction are only a few of the bio-data mining applications of data exploration. Gene discovery, protein function area identification, function theme recognition, protein function inference, and illness diagnosis are just a few of the many bio-data mining systems.

2.1 Amino acids

Amino acids, shown in Fig. 1, are chemical molecules of biological significance. They have the functional assemblies amine (NH2) and carboxylic acid (COOH) and a cross-chain that is unique to

respectively amino acids. Carbon, oxygen, hydrogen, and nitrogen make up the backbone of amino acids, while some amino acids include additional elements in their side chains. About 500 amino acids are recognized and can be categorized in different ways.



(a)Construction of amino acid (b) Basics of amino acid

Fig. 1. Insights into the makeup and composition of amino acids [6]

2.2 Protein

Proteins stand macromolecules made up of linked chains of amino acids. The nucleotide sequence of the gene that encodes the protein sets the priority.

Protein Sequence:- HMAAVVALVSLRRRLPATTLGG HACLQASRGAQAATTTVVV......

The biological function of a protein is determined by the chemical characteristics of its constituent amino acids. Twenty amino acids are used by cells in biosynthesis. The genetic code determines whether an amino acid is necessary for survival, as seen in Table 1.

One letter	Three letter	Amino-acid
code	code	name
А	Ala	Alanine
В	Asx	Aspartic acid
С	Cvs	Cvsteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenvlalanine
G	Glv	Glycine
Н	His	Histidine
Ι	Ile	Isoleucine
K	Lvs	Lysine
L	Leu	Leucine
М	Met	Methionine
N	Asn	Asparagine
0	Pyl	Pyrrolysine
Р	Pro	Proline
0	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
Т	Thr	Threonine
U	Sec	Selenocysteine

V	Val	Valine
W	Trp	Tryptophan
Х	Xaa	Any amino acid
Y	Tvr	Tvrosine
Z	Glx	Glutamic acid

Vol-1 Issue-1 2023 Scientific Research Journal of Science, Engineering and Technology

2.3 Index of Protein Sequences

The term "sequence database" refers to a subcategory of "biological database" that stores, in a digital format, an extensive collection of "sequences" such as "nucleic acid sequences," "protein sequences," or "other polymer sequences," among other types of "polymer sequences." The Protein database is comprised of several different sequences, some of which come from GenBank, RefSeq, and TPA; others come from SwissProt, PIR, and PRF; and yet more come from the Protein Data Bank (PDB). Protein sequences are an essential component in determining how organisms are constructed as well as how they function.

3. Methodology

Performing a literature survey entails investigating the current setup and assessing whether or not it needs to be altered. The current system can only predict how often an item will be used by using an apriori algorithm and an association rule. This recurring pattern is legitimate only if your confidence is at least 90%.

However, my strategy is to broaden this study and identify the commonalities.

The following algorithms, methods, and logic have all been discussed in the literature we surveyed.

- 1. Fuzzy logic
- 2. Genetic Algorithm
- 3. Association Rule Mining

3. 1 Association Rule Mining

Pattern prediction in association rule mining may be accomplished using a variety of different approaches. Nevertheless, specific approaches, such as Apriori and DIC algorithms, have restrictions in areas such as the amount of runtime complexity, the amount of storage required, and the cost. With dataparallel formulation (DPF), the calculations [7] required to determine the occurrence of the different sequences located at each node of the tree may be split up between a large number of processors. The time complexity is decreased, and this equivalent formulation is conceptually similar to the count distribution method used to parallelize the serial Apriori technique. This technique is used to find often recurring groupings of things.

In most cases, the mining process for association rules may be divided into two phases:

- 1. Locate all sets of things used frequently: The specification states that each of these items will happen at least as often as the stated minimum livelihood count.
- 2. Create reliable association rulebooks from the most common item sets: By their very definition, these rules must attain at least a base degree of acceptance and trust.

Allow I to be a collection of things, where Let $I = \langle i1, i2 \dots im \rangle$ etc. Let D, the task-relevant data, be a collection of sequential databases, each of which S is a collection of [8] (amino acid) items such that

that $S \subseteq I$. An identification known as SID is connected to each amino acid (i.e., Sequential Identification). A should represent a collection of things. Only when A S is a sequence S considered to contain A. A rule of association is implied by the formula $\subseteq B$, where $A \subseteq I$, $B \subseteq I$, and $A \cap B = \Phi$ is referred to as the antecedent, while B is consequent; the rule states that A entails B.

The equations (1) and (2) below illustrate the most fundamental metrics for making predictions about association rules: support(s) and confidence.

Support
$$(A => B) = \frac{Sequence of A and B}{Total number of sequence}$$

Equation 1

An exciting association rule $A \Rightarrow B$ can be generated if the percentage is larger than the confidence level. To calculate the confidence(c) of an association rule, we divide the number of sequences containing $A \cup B$ by the number of records containing A. A rule of association's confidence can be used as a measure of its efficacy. If the confidence level of an association rule is 88%, for instance, it means that 87% of all transactions that include A also include B. Users can also establish a threshold for their level of confidence in the rules to guarantee that they are at least somewhat engaging. The confidence that the rule $A \Rightarrow B$ holds in the transaction set D is equal to the fraction of sequences in D that includes both A and B. For a set of rules to be considered vital, they must meet both a minimal threshold of consensus support and minimal confidence criteria.

Confidence
$$(A \Rightarrow B) = \frac{Sequence \text{ of } A \text{ and } B}{Sequence \text{ of } A}$$

Equation 2

The use of the association rule to discover the frequency is beneficial for several different purposes, including the detection of protein function domains, function motifs, protein function inference, disease analysis, disease prediction, disease behavior optimization, reconstruction of the network of protein and gene interactions, data cleaning, and protein sub-cellular location prediction. All of these and more can be accomplished by using the association rule.

3.2 Combining Fuzzy Logic and Genetic Algorithms

Various optimization strategies are based on genetic algorithms. Computational systems have become more appealing for some forms of optimization as their performance has continued to increase. Cells are the fundamental building blocks of every living entity. An identical collection of chromosomes may be found in each individual cell. DNA is strung together in chromosomes, which act as a kind of blueprint for the entire organism. Genes are the building units of DNA that makeup chromosomes.

Inputs:

- 1. Medical record
- 2. The standard range of characteristics (including cholesterol, blood pressure, and other parameters) derived from professional knowledge
- 3. Rules derived from the Apriori algorithm

Output:

Vol-1 Issue-1 2023 Scientific Research Journal of Science, Engineering and Technology

Values of characteristics at their most desirable extremes; for example, the chromosome with the highest fitness potential.

Procedure:

- 1. [Begin] Produce a population of n chromosomes that is chosen at random (suitable solutions for the problem)
- 2. [New people living here] To create a new population, repeat the processes outlined below until the new inhabitants are finished.
 - a) [Selection] Take two sets of parental chromosomes from a populace and evaluate them based on how to fit they are (the better fitness, the bigger chance to be selected)
 - b) [Crossover] A new generation can be produced by crossing the genes of both parents, which has a certain likelihood (children). If there were no genetic crossing in the process, the offspring would be an identical replica of the parents.
 - c) [Mutation] Following the likelihood of mutation, mutate the new offspring at each locus (position in DNA).
 - d) [Receiving] Introduce newly produced children into an existing population
- 3. [Replace] In subsequent iterations of the algorithm, use the newly produced population.
- 4. [Test] If the end condition has been met, come to a halt and return the optimal solution based on the currently available population.
- 5. [Loop] Proceed to the Second Step

The use of GAs may be put to use in the field of bioinformatics for the purpose of solving specific multiobjective problems, which in turn optimizes the amount of computing that is required and produces solutions that are resilient, quick, and approximative. In addition, the error-handling capabilities of GAs allow for the handling of mistakes that are produced during experiments, including data from bioinformatics [9]. These kinds of mistakes might, to a certain extent, be seen as adding to genetic variation, which is a quality that is desired. A new field of study has been opened up by the challenge of merging genetic algorithms with bioinformatics.

The Application of Fuzzy Logic to Bioinformatics

It is not challenging to employ fuzzy logic to create systems of any complexity, from the simplest, smallest, or even embedded ones to the most complex, networked ones. Fuzzy logic is defined by the fact that it is able to accept the reservations that are inherent in the realistic inputs, and it is also able to deal with these uncertainties in such a way that their influence is minimized, which ultimately leads to exact outputs. Since the initial stage of fuzzy logic design is to understand and characterize the system's behavior by utilizing one's knowledge and experience, fuzzy logic can reduce the number of design processes and simplify any complexity that may crop up. Fuzzy logic, often known as FL, is a concept

Vol-1 Issue-1 2023 Scientific Research Journal of Science, Engineering and Technology

that Lotfi Zadeh developed. FL offers a straightforward method for arriving at a definitive conclusion based on information that may be unclear, ambiguous, vague, noisy, or absent entirely. It uses human controller logic as a model.

4. Work Plan and Execution

Work proposal

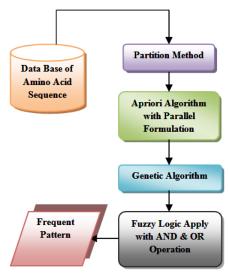


Fig 2. Work Proposal

1. A database was used to gather the standard set of amino acids.

2. The dataset is divided into several parts.

3. The preprocessed data is then sent to the Apriori algorithm, which uses a parallel formulation to generate fascinating correlations and connection rules.

4. The genetic algorithm module receives the results of the apriori module. Finally, do the crossover technique in order to create offspring. Use this offspring to produce the typical pattern through the apriori method.

5. For the fuzzy logic module to provide the required knowledge of recurrent patterns, the results of the genetic algorithm module and the apriori algorithm module are transferred to it. Additionally, improve the recurring pattern that is shown in the results' implementation section.

Execution

Transfer the Database file.

This is the most basic option, providing only the functionality necessary to submit your input file. The file used as input contains the sequence of amino acids.

File:	Select a file	Select File
String Input:		Submit
Output:		

Fig 3. Examine the input file

Method of partitioning

When working with larger input strings, the partitioning technique is what is needed to generate the transaction id. The apriori algorithm will take into consideration this transaction as an input. As a result, it is used for calculations that are not very complex. Because of this, the database of amino acid sequences is straightforward to handle when utilizing the partition approach.

NikitApp - NetBeans IDE 7.0	U 1 1.	
Elle Edit View Navigate Source Refact	stor Bun Debug Brofile Team Iools Window Help	Q = Search (Ctrl+t)
2 🔁 🏙 🧤 🤊 🥙 🖸	rdefaulteendep 💿 🐨 🎲 🕨 🖡 🕫 😗 -	
I Projects	● R StarfAge R (Aglans R McDatages R McDatages R Controllate as R Control Case R Control Case R R McDatages R R R R R R R R R R R R R R R R R R R	
Backgo	10 /* In charge this license baskey, dhose license Baskey in Frederi Properties. * To charge this template file, dhose Total ; Templates * and upper the template in the editor. * y package UL; package UL; pa	
file - Navigator	is L import java.util.Iterator; g H 16 /**	
Members Vew	Conjust - MARADO (rem) Seaded PETIADE: MERICIPEUMINEDIA/PUNEELANA/OPEMPA/VALGLAN/PELPHAADI/VADELETINOTLANDODETINOTINOTLANDODETINOTINOTATICANDODETINOTICANOTANOTINATICANDODETINOTICANOTINATICANDODETINOTICANOTINOTICANOTIC	
	Vert Australia Method Australia Method Australia Method Australia Method Australia Method Australia Method Australia	
Tesks		
	NétrApp (un)	noning 🛛 🖬 🧕 25 5 14

Fig 4. Method of partitioning

Intention Utilizing the Apriori Algorithm

The apriori technique was used to generate a pattern of frequent occurrences. This is the input seed taken from the file you chose as the input. In the approach that has been presented, the emphasis is placed on determining which amino acids are more likely than other amino acids to induce a viral illness by analyzing protein data sets and then feeding this information into a genetic algorithm. Choose two different parent ids based on this recurring pattern. This is the red line pattern that was chosen to be the input for the genetic algorithm.

File:	New Text Document (2).txt	Change File
String Input:		Submit
Output:		
	1538461538463 4 6153846153846 3	
	6153846153846 3	
	2307692307693 21	
	5384615384613 23	
15 M,G,P 18.26	9230769230766 19	
	9230769230766 19	
	6153846153843 30	
18 K,T,C 8.653		
19 K,C,Y 1.923 20 K,C,V 8.653	0769230769231 2	
21 K,C,W 0.961	5384615384616 1 6923076923077 11	

Fig 5. Intention Utilizing the Apriori Algorithm

The Genetic Algorithm Used in the Crossover Process The Parent item sets must also be entered.

In the beginning, a group of common patterns that are formed at random will have some randomness introduced into it. The first red mark is chosen from figure 5. After that, the crossover operation is done on the string, which is followed by the string being fed into the String input field. The apriori algorithm will have the outcome of the crossover operation as one of its inputs. Once again, a consistent pattern will emerge. As seen in figure 6, you will observe. In addition, the apriori algorithm uncovered a second pattern match with the red line.

A Fuzzy Appre	och To Find The Frequent pattern C Using Genetic Algorithm	
File:	New Text Document (2).txt	Change File
String Input: Output:	MGGHGILCGLWG	Submit
Output 1 M 25.0 1 2 Q 25.0 1 3 G 100.0 4		Â
4 C 25.0 1 5 L 50.0 2 6 W 25.0 1		
7 H 25.0 1 8 I 25.0 1 1 M,Q 25.0 1		D
2 M,G 25.0 1		, •

Fig 6. Alternate use with Common item crossed over.

The apriori technique is used to determine the typical pattern seen in this snapshot (fig. 7). The result of the crossover operation often matches the pattern represented by the red marks and the rest of the patterns, as determined by the fuzzy approach.

A fuzzy Approch To Find The Frequent pattern Of Amino Acid Sequed Using Genetic Algorithm File: New Text Document (2).txt Change File: New Text Document (2).txt Change String Input: Submit Output: Submit V 201920769207693 21 Submit File: New Text String String 9 Ho 23:40530769207693 21 Submit 9 Ho 23:40530769207693 24 Submit 10 K,T 25:961398461398463 27 Submit 11 K, C 0:3463398461398463 27 Submit 12 K, G 1:461398461398463 27 Submit 14 K, C 1:461398461398463 27 Submit 15 K, L 2:400398461398463 27 Submit 16 K, C 1:461398461398463 27 Submit 18 K, C 1:460398461398463 27 Submit 18 K, C 1:460398661398463 27 Submit 18 K, C 1:460398661398463 24 Submit 18 K, C 1:461398661398463 24 Submit 18 K, C 1:460398661398463 24 Submit			
File: New Text Document (2).txt Change String Input: Submit Output: Submit 1 % K 0.0120176920769207693 21 Submit 6 % 0.3 2.020769207693 34 T 7 % A 25.96135464154414 1 String Input: 1 % K 0.0120769207693 34 T 9 % J 25.961354641534643 27 String Input: 1 % K 0.0163307693207693 34 T 9 % J 25.961354645138464 27 String Input: 1 % K 0.016334613534645 27 String Input: 1 % K 0.016334613534645 21 String Input: 1 % K 0.0164334645346445 20 String Input: 1 % K 0.01643346453346454 25 String Input: 1 % K 0.016445334645346454 25 String Input:	A Fuzzy Appro	och To Find The Frequent pattern	Of Amino Acid Sequnce
Numeric Cuturget String Input: Submit Output: Submit 13 Kr, 27, 1285613384613857 737 Submit 4 Kr, C 0.9613846613884614 1 Submit 5 Kr, 27, 253507682307683 1 Yestion 1 7 Kr, 23, 263307682307683 1 Yestion 1 9 Kr, D 13, 46138461388462 34 Yestion 1 13 Kr, 27, 278307682307683 21 Yestion 1 9 Kr, D 13, 46138461388463 24 Yestion 1 13 Kr, 27, 278307682307683 21 Yestion 1 13 Kr, 27, 278307682307682 37 Yestion 1 14 Kr, 27, 268307682307682 38 Yestion 1 15 Kr, 13 Yestion 1 Yestion 1 16 Kr, 12 Yestion 1 Yestion 1 16 Kr, 13 Yestion 1 Yestion 1 16 Kr, 14 Yestion 2 Yestion 1 16 Kr, 14 Yestion 2 Yestion 2 16 Kr, 14 Yestion 2 Yestion 2		Using Genetic Algorith	
String Input: Cutury Output: Submit 3 Kr, 7 27,884(15384615384616 Submit 9 Kr, 0 : 961384645384616 Submit 9 Kr, 0 : 961384645384616 Submit 1 Str. 7 : 7884(15384615384616 Submit 1 Kr, 7 : 7884(153846158462 Submit 1 Kr, 7 : 7884(153846158462 Submit 1 Kr, 7 : 788307692307683 Submit 1 Kr, 7 : 788307692307683 Submit 1 Kr, 7 : 788307692307683 Submit 1 Kr, 7 : 788307692307682 Submit <th></th> <th></th> <th></th>			
Instruction New Hold Ducument (2),XX Citizing String Input: Submit Output: Submit 3 K,7 27,854615384615387 29 4 K,6 0.9615384615387 29 4 K,6 0.9615384615384614 1 5 5 K,7 27,8230768207683 1 7 7 Kh, 23,961538461538461 21 7 8 K,8 20,1023768207683 1 7 9 K,9 13,46153846153846 21 1 12 K,6 41,346153846153846 2 1 13 K,7 7,768307682077632 1 1 14 K,7 0.96830768207683 21 1 9 K,9 13,46153846153846 2 1 13 K,8 0.7,683076820777 22 1 14 K,7 7,6930768207682 7 2 15 K,8 0.7,6830768207682 8 1 12 K,6 1,34615384615385 43 1 13 K,8 0.7,6830768207682 8 1 14 K,7 7,6930768207682 8 1 14 K,7 7,6930768207682 8 1 14 K,7 7,6930768207682 8 1 14 K,8 7,76846153864 8 1 15 K,8 7,77 8 1 16 K,8 7,8 40,846538644 8 1 14 K,8 7,7684565386454 28 1 <th></th> <th></th> <th></th>			
String Input: Submit Output: Submit 0 Kp 21 7:0584615584615387 259 5 4 Kp 0 0.9613546436384616 1 5 6 Kp 0 32:052076820768 31 7 7 Kp 22:052076820768 31 7 9 Kp 0 31:461538461538463 47 6 9 Kp 0 31:461538461538463 47 1 10 Kp 0 31:461538461538463 77 1 11 Kp 0 3:461538461538463 41 1 12 Kp 0 41:3461538461538463 42 1 13 Kp 0 7:4623076820769207692 61 1 14 Kp 0 7:4623076820769207692 7 1 15 Kp 0 7:4623076820769207692 7 1 15 Kp 0 7:4684054644 25 1 15 Kp 0 7:468405464 25 1 15 Kp 0 7:468405464 26 1 15 Kp 0 7:468405464 25 1 16 Kp N 42:4645546453054 25 1 16 Kp N 42:46455464536454 25 1	File:	New Text Document (2) txt	Chapae Fi
Output T N,T 27.094615384615387 25 4 N,T 27.094615384615387 25 4 N,T 27.09461538461538461 25 5 N,T 27.09461538461538463 2 5 N,S 20.192076920769 34 7 N,A 25.961539461538463 27 8 N,S 20.192076920769 34 7 N,A 25.961539461538463 2 1 N,C 3.7846153846153846 2 1 N,C 3.784615846153846 2 1 N,C 3.7846158461538461 2 1 N,C 3.784615846153861 2 1 N,C 3.784615846153861 2 1 N,C 3.784615846158461 2 1 N,C 3.78461584615846158461 2 1 N,C 3.784615846158461 2 1 N,C 3.784615846158461 2 1 N,C		not fox boothin (2)	Containgo II
Output 3 N,7 27.5946153946153946153967 25 4 N,6 05.41539461539461539463 2 5 N,7 27.59461539461539463 2 5 N,5 25.626270520768 2 7 N,4 25.961539461539463 2 9 N,2 50.152076920768 2 1 N, C 3.546153946153946 3 1 Z K,6 41.346153946153964 3 1 Z K,6 41.346153946153954 4 2 Z K,6 41.3461539461539461 4 2 Z K,6 41.3461539461539461 4 2 Z K,6 41.3461539461539461 4 2 Z K,6 41.3461539461539641 4 2 Z K,6 41.3461539461 4 2 Z K,6 41.34615454545454545454545454545454545454545	String Input:		
3 Kr 27, 88453544513947 25 4 Kr C 0-961334645384613 5 Kr 20.19230769220769 34 7 Kr 23.29230769220769 34 7 Kr 23.94230769220769 34 1 Kr 23.64235464358463 27 1 Kr C 3.646135464538464 27 1 Kr C 3.646135464538464 27 1 Kr C 3.646135464538463 4 1 2 Kr C 41.346153464538463 4 1 2 Kr C 41.346153464538463 4 1 5 Kr 2 2.0934613546412 5 1 5 Kr 2 3.0934613546412 5 1 6 Kr 24.0034613546412 5 1 6 Kr 24.0034613546412 5 1 6 Kr 24.0034613546412 5 1 7 Kr 25.0034613546412 5 1 7 Kr 25.003461346412 5 1 7 Kr 25.003461346412 5 1 7 Kr 25.003461346134612 5 1 7 Kr 25.00346134612 5 1 7 Kr 25.0034612	oning inport.		Submit
3 Kr 27, 88453544513947 25 4 Kr C 0-961334645384613 5 Kr 20.19230769220769 34 7 Kr 23.29230769220769 34 7 Kr 23.94230769220769 34 1 Kr 23.64235464358463 27 1 Kr C 3.646135464538464 27 1 Kr C 3.646135464538464 27 1 Kr C 3.646135464538463 4 1 2 Kr C 41.346153464538463 4 1 2 Kr C 41.346153464538463 4 1 5 Kr 2 2.0934613546412 5 1 5 Kr 2 3.0934613546412 5 1 6 Kr 24.0034613546412 5 1 6 Kr 24.0034613546412 5 1 6 Kr 24.0034613546412 5 1 7 Kr 25.0034613546412 5 1 7 Kr 25.003461346412 5 1 7 Kr 25.003461346412 5 1 7 Kr 25.003461346134612 5 1 7 Kr 25.00346134612 5 1 7 Kr 25.0034612			
4 UK C 0.961396451394616 1 5 MY 25.012930769230769 31 5 MY 25.012930769230769 32 5 MY 25.012930769230769 34 7 MA 25.96230769230769 32 9 M, 5 13.4615946135446 37 10 K/ 7 35.9613946155461 24 10 K/ 7 35.9613946155461 3 12 K/ 6 41.364515461354 4 13 K/A 30.76923076923077 32 14 K, C 7.6623076923077 32 14 K, C 7.662307692307828 6 15 K/L 24.03846153946154 25 16 K/M 24.03846153946154 25 16 K/M 24.03846153946154 25 17 Ministry 10	Output:		
5 k/ 7 20.122076220769 21 6 k/ 8.1 20.252076220769 34 7 k/ 2.5 4623076220769 34 9 k/ 5 10.12076220764 27 8 k/ 5 10.12076220764 2 10 k/ 7 25.9613964513964 5 12 k/ 6 41.3641538461394 5 12 k/ 6 41.3641538461394 5 15 k/ 20.763207682077 2 15 k/ 20.763207682077 2 15 k/ 20.763207682077 2 16 k/ 21.0264153846134 25 16 k/ 21.0264153846134 25 16 k/ 21.0264153846134 25 17 k/ 21.0264153846134 25 18 k/ 21.026413484514 25 18 k/ 21.026413484514 18 k/ 21.02641348514 18 k/ 21.02641348514 18 k/ 21.02641348514 18 k/ 21.02641348514 18 k/ 21.02641348514 18 k/ 21.026414 18 k/ 21.	3 M.T 27.884615	384615387 29	
6 M, 0 32.6823078230786 34 7 M, 2 5.96330645354645374 9 M, 5 20.192307683207683 21 9 M, 5 20.192307682307683 21 10 K, 7 25.96133641538462 14 10 K, 7 25.96133641538462 24 11 K, C 41.44515846153842 14 12 K, C 41.4451584615384 25 13 K, A 30.78942076923079 32 14 R, C 7.6620706923076928 6 15 H, L 24.00346153846154 25 16 H, N 24.00346153846154 25	4 M,C 0.9615384	615384616 1	
7 % M, 25.961358461358463 27 6 % 8 20.19207062307683 21 9 % 0 13.46135846155462 14 1 % C 3.5646135846155462 14 12 % C 41.346135846135865 43 12 % C 41.34613584613865 43 13 % A 30.7682307662307 32 14 % C 7.462307662307682 5 14 % C 7.462307662307682 5 14 % C 7.462307662307682 5 14 % C 7.462307682307682 5 14 % C 7.462307682307682 5 14 % C 7.462307682376842 5 2 %			I
e M. 5 20.19200769207693 21 9 M.D 13.46153646153642 14 10 K.T 25.96153646153646 27 11 K.C 3.964153646153646 3 12 K.G 41.36415364415396 45 14 K.C 7.6620769207692076920 14 K.C 7.6620769207692076925 15 H.L 24.09346153846154 25 15 H.L 24.09346153846154 25			
9 m, D 13, 461354641354642 14 10 K, T 25, 961538464518463 27 11 K, C 3, 061538464518463 27 12 K, G 41, 3461538461538463 4 13 K, A 30, 76923076922077 32 14 R, C 7, 6523076922076925 8 15 M, L 24, 03846153846154 25 16 M, N 24, 03846153846154 25 17 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7			1
10 kg 2 3 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5			
11 K,C 3.860153640538463 4 12 K,G 41.364538461538 4 13 K,A 30.7692307692077 32 14 R,C 7.692307692076928 8 15 R,L 24.03846153846154 25 16 K,N 24.03846153846154 25			
12 K, 6 41.346135864613585 43 13 K, 8 30.74833076923074925 8 13 K, C .46833076923074925 8 14 K, C .46833076923074925 8 14 K, N 24.0484513946134 25 14 K, N 24.0484513946134 25			1
13 K,A 30.76923076923077 32 14 R,C 7.6523076923076925 6 15 K,L 24.03846153946154 25 16 K,N 24.03846153946154 25			
14 B,C 7.6823076925 8 15 B,L 24.084615384614 25 16 B,N 24.03846153846154 25			1
15 H, L 24.0346153946154 25 16 H, N 24.0366153946154 25			
16 H,N 24.03846153846154 25			
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,			
File Selected New Entergravit	16 H,N 24.03846	3153846154 25	
File Selected MenuEnter equipment Devices of			7.6
File School and Mary Enter any input			
	File Selected Now	Enter any input.	Powered By:- Nikit Pa

Fig 7. Frequency pattern determination using an apriori algorithm.

Fuzzy logic

The outputs of the apriori algorithm module (shown in figure 7) and the genetic algorithm module are fed into the fuzzy logic module, which then uses those results to construct the requisite frequent patterns (fig. 6). As can be seen in Figure 8, the genetic algorithm module and the apriori algorithm module both contain item sets that are identical to one another. Figure 8's red rectangles provide visual confirmation that the aforementioned layout is available in both modules. Protein structure prediction, gene classification, microarray-based cancer classification, gene expression data clustering, statistical modelling of protein-protein interactions, and other similar endeavours, as well as their children, stand to gain a great deal from the application of this pattern.

A Fuzzy Approch To Find The Frequent pattern Of Amino Acid Sequnce Using Genetic Algorithm		
File:	New Text Document (2).txt	Change File
String Input:	MGQHGI LCGLGW	Submit
Output:		
*****Fuzzy Log		
м, с		
L,W C,W G,I		
**** Match No	tFound Or Offspring Generated ***	
M, Q Q, G		
Output		Powered By:- Nikit Patel

Fig 8. The final product of the apriori method compared to that of the genetic algorithm.

5. CONCLUSION

The process of doing research on the existing system in order to ascertain whether or not it is essential for the system to be re-engineered is known as a literature survey. The current state of affairs makes it such that using an apriori algorithm and an association rule is the only method to foresee a popular item. Ninety percent is the minimum required to consider this prevalent pattern to be an authentic one.

The partition method, the apriori algorithm, the genetic algorithm, and fuzzy logic are the four different algorithms I will use to predict the familiar pattern of amino acids. In order to produce the required knowledge in the form of simple patterns, the fuzzy logic module takes input from both the genetic algorithm module and the apriori algorithm module. Additionally, it is essential to make the most of the regimen that is followed. It is to everyone's advantage for companies in the pharmaceutical sector to use this Common Pattern. The major objective is to make an educated guess as to which combinations of amino acids would be the most beneficial when developing treatments for ill health.

In the future, this work will possibly be developed to employ the Eclat algorithm rather than the apriori method to improve the accuracy with which it forecasts the typical pattern. In addition, the protein sequences of some diseases, such as HIV/AIDS, influenza, dengue fever, viral fever, swine flu, and others, are compared to find similarities among them. If medicines treating viral infections adhere to this more constant pattern, they will be more successful in curing the conditions they are intended to treat.

Bibliography

- [1] H. C. D. X. a. X. Y. Jiawei Han, "Frequent pattern mining: current ststus and future directions," *Data Mining Knowledge Discovery*, vol. 1, no. 5, pp. 55 -86, 2007.
- [2] S. H. Lakshmi Priya. G., "A Study On Predicting Patterns Over The Protein Sequence Datasets Using Association Rule MINING," *Journal Of Engineering Science And Technology*, vol. 7, no. 5, p. 563 – 573, 2012.
- [3] D. U. A. J. H. Moore, "Data Mining And Thev Evolution Of Biological Complexity," *Biodata Mining*, pp. 4-7, 2011.
- [4] K. Raza, "Application Of Data Mining In Bioinformatics," *Indian Journal Of Computer Science And Engineering*, vol. 1, no. 2, pp. 114-118, 2010.
- [5] wiki, "Amino Acid".
- [6] Valerie Guralnik And George Karypis Parallel Formulations Of Tree-Projection-Based Sequence Mining Algorithm.
- [7] A. R. A. S. R., "Fast Algorithms For Mining Association Rules," in *Proceedings Of The 20th International Conference On Very Large Data Bases*, Santiago De Chile, 1994.
- [8] R. S. I. Dr. Tryambak A. Hiwarkar, "New Applications Of Soft Computing, Artificial Intelligence, Fuzzy Logic & Genetic Algorithm In Bioinformatics," *IJCSMC*, vol. 2, no. 5, p. 202 – 207, May 2013.
- [9] S. H. Lakshmipriya, "An Efficient Approach For Generating Frequent Patterns Without Candidate Generation," *ACM*, pp. 3-5, 8 2012.