Vol-1 Issue-1 2023
Scientific Research Journal of Science, Engineering and Technology
ISSN: 2584-0584, Peer Reviewed Journal

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

Krupali Patel[1], Pravinbhai Patel[2]
[1]Department of Genetic, [2]Department 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.

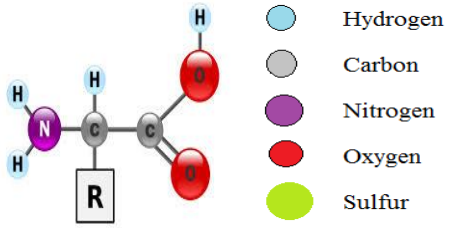
- Analyzing sequences
- Simulating biological processes
- Annotating the Genome
- Membrane protein analysis
- Cancer mutation analysis
- Calculating the 3D structure of proteins
- Interaction between proteins using docking
- Genomic comparisons Etc.

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.

Table 1. One- and three-letter codes for representing amino acids

One letter code	Three letter code	Amino-acid name
A	Ala	Alanine
B	Asx	Aspartic acid
C	Cys	Cysteine
D	Asp	Aspartic acid
E	Glu	Glutamic acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
O	Pyl	Pyrrolysine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
U	Sec	Selenocysteine

V	Val	Valine
W	Trp	Tryptophan
X	Xaa	Any amino acid
Y	Tyr	Tyrosine
Z	Glx	Glutamic acid

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 data-parallel 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 item**s 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 $I = < i1, i2 \ldots im >$ 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 ⊆ 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 ⊆ B, where A ⊆ I, B ⊆ I, and A ∩ B = Φ 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.

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

Equation 1

An exciting association rule A => **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 ∪ 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 => **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.

$$\boldsymbol{Confidence}\ (\boldsymbol{A} => \boldsymbol{B}) = \frac{\boldsymbol{Sequence\ of\ A\ and\ B}}{\boldsymbol{Sequence\ of\ A}}$$

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:

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 multi-objective 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

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

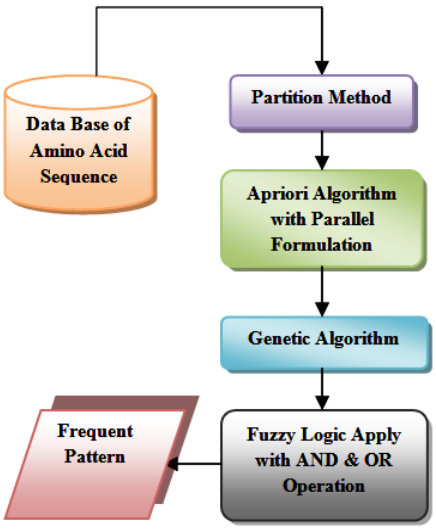


Fig 2. ***Work Proposal***

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.

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.

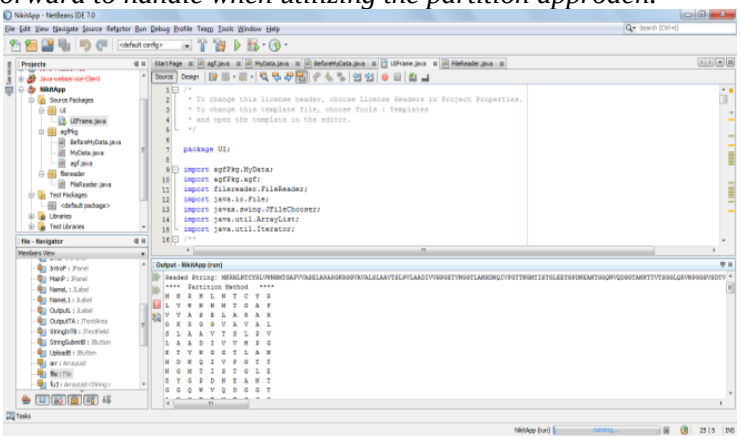


Fig 4. *Method of partitioning*

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.

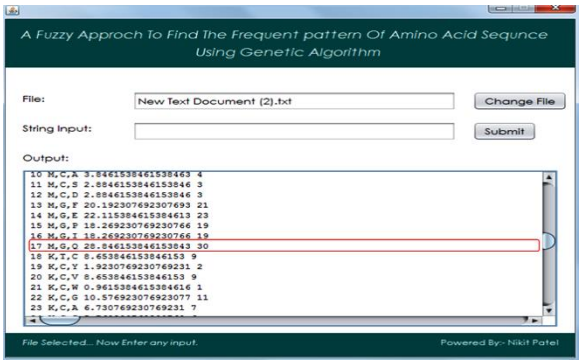


Fig 5. ***Intention Utilizing the Apriori Algorithm***

.

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.

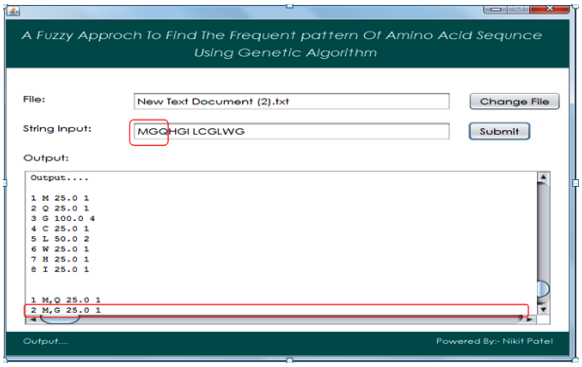


Fig 6. Alternate use with Common item crossed over.

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.

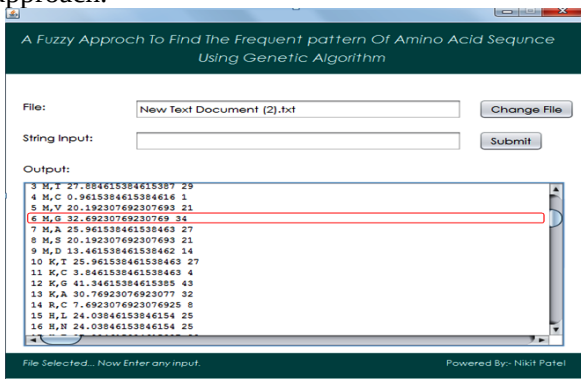


Fig 7. Frequency pattern determination using an apriori algorithm.

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.

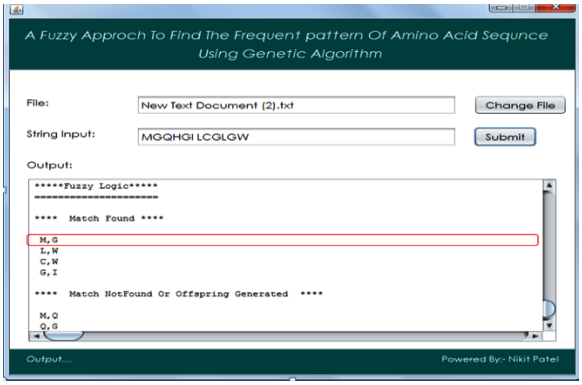


Fig 8. The final product of the apriori method compared to that of the genetic algorithm.

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.