Predict the Frequent Pattern of Amino Acids Using Apriori Algorithm, Genetic Algorithm and Fuzzy Logic

Nikit P. Patel[1] Pratik Kumar[2]
[1]Department of Information Technology, [2]Department of Computer Science & Engineering
[1, 2] Parul Institutes of Engineering & Technology, Vadodara, Gujarat, India

Abstract--- Data Mining is the process of extracting or mining the patterns from very large amount of biological datasets. In recent years, rapid developments in genomics and proteomics have generated a large amount of biological data. So a critical problem in biological data analysis is to classify the biological sequences and structures based on their critical features and functions. From the literature, various algorithms have been employed in generating frequent patterns for distinct application. This algorithm has been lost of frequent produce. So it's meaningless. Here my approach is to compare the frequent pattern using two algorithms and optimise the data. So it's very useful for us. Our approach aims at extracting the hidden and the most dominating amino acids among the infected protein sequence which causes some infections in human. We handle this problem by predicting patterns apply strong association rules along with apriori algorithm and genetic algorithm. Also apply the fuzzy logic to optimise data and interesting frequent pattern get form the protein sequence database. This Frequent Pattern is very useful to drug design, drug discovery etc

Keywords: - Bio-Data Mining, Protein sequence, Association Rules, Genetic Algorithm & Fuzzy logic

I. INTRODUCTION

Data Mining is a collection of techniques for efficient automated discovery of patterns from large data sets. 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" [1] and it may be called Knowledge Discovery in Databases (KDD).

Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery from data. Data Mining approaches seem ideally suited for Biological Data Mining, since it is data-rich, but lacks a comp- rehensive theory of life's organization at the molecular level. The extensive databases of biological information create both challenges and opportunities for development of novel KDD methods. 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 [2].

A crucial challenge in the future of biological science involves putting that data to work. Now life scientists hope to plan large experiments [3] collect lots of data, analyze it, compare data between experiments, and eventually combine all of that information to improve basic theories, biotechnology, and medicine. The functionalities of the Data mining techniques are as follows; data characterization, data discrimination, association analysis classification, prediction, clustering, outliers and association rule mining.

A. Data Preprocessing

"The collection and manipulation of items of data to produce meaningful information." There are various data preprocessing techniques, such as, (i) data cleaning, (ii) data integration, (iii) data transformation and, (iv) data reduction. The proposed work was carried out in section 4 by preprocessing the biological dataset for pattern prediction.

1) Data reduction

It is a reduced representation of any data set in smaller volume and produces almost same analytical results [4]. The data reduction is handled by applying the following strategies;

- Data aggregation operations are performed on the data for constructing the data cube.
- Attribute subset selection is applied on the data when it's irrelevant, weakly relevant, redundant or dimensions may be deleted or removed.
- Dimensionality reduction is used to reduce the data set size.
- Numerosity reduction helps to the data to be replaced or estimated by smaller data representation.

II. BIO-DATA MINING

Bio-Data mining [2] is the design and development of computer based technology that supports life science. In the past two decades the changes in biomedical research, biotechnology and an explosive growth of biomedical data, ranging from those collected in pharmaceutical studies and cancer therapy investigations [5] to those identified in genomics and proteomic research by discovering sequential patterns, gene functions, and proteinprotein interactions.

Some of the grand area of research in bioinformatics

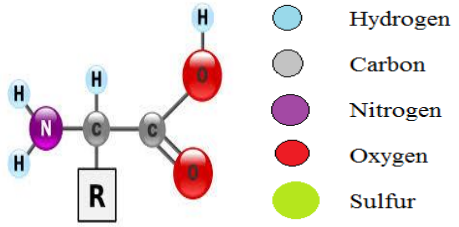
- Sequence analysis
- Modeling biological systems
- Genome annotation
- Analysis of protein expression
- Analysis of mutations in cancer
- Protein structure prediction
- Protein-protein docking
- Comparative genomics Etc…

Applications of data mining to Bio Data Mining includes gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

A. Amino acids

Amino acids are Fig1. biologically important organic compounds composed of amine (NH_2)and carboxylic acid (COOH) functional groups, along with a side-chain specific to each amino acid. The key elements of an amino acid are carbon, hydrogen, oxygen, and nitrogen, though other

elements are found in the side-chains of certain amino acids. About 500 amino acids are known and can be classified in many ways.



(a)Structure of amino acid (b) Elements of amino acid

Fig. 1: The structure and elements of amino acid [6]

B. Protein

Proteins are large molecules composed of one or more chains of amino acids in a specific order. The order is determined by the base sequence of nucleotides in the gene that codes the protein.

C. Protein Sequence: - Hmaavvalslrrrlpattlgg Haclqasrgaq

The chemical properties of the amino acids in proteins determine the biological activity of the protein. Cells use the combination of twenty amino acids during biosynthesis. It is specified by genetic code and they can either be essential or non- essential amino acid as illustrated in Table 1.

Table 1: Representation of amino acid in one and three letter code

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

D. Protein Sequence Database

A sequence database is a type of biological database that is composed of a large collection of computerized nucleic acid sequences, protein seque- nces, or other polymer sequences stored on a computer. The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

III. Related Work

Literature survey is the act of studying the existing system and analyzing the need for the system to be reengineered. In existing system only using association rule along with apriori algorithm using to predict the frequent item. And consider only confidence above 90% this frequent pattern is valid.

But my approach is the extend this work and find out the frequent.

Literature survey has covered the following Algorithm, method and logic.

- Association Rule Mining
- Genetic Algorithm
- Fuzzy logic

A. Association Rule Mining

In association rule mining, several algorithms are available for predicting frequent patterns. But few algorithms (Apriori, DIC algorithms) have certain drawbacks such as time complexity, space complexity and cost. Usinf data parallel formulation (DPF) is to divide between the different processors the computations[7] required to determine the frequency of the various sequences at each node of the tree. And decrease the time complexity the To a large extent, this parallel formulation is similar in nature to the count distribution method developed for parallelizing the serial Apriori algorithm for finding frequent itemsets.

In general, association rule mining can be viewed as a two-step process:

- Find all frequent itemsets: By definition, each of these itemsets will occur at least as frequently as a predetermined minimum support count.
- Generate strong association rules from the frequent itemsets**:** By definition, these rules must satisfy minimum support and minimum confidence.

Let $I = \{i_1, i_2 \ldots i_m\}$ be a set of items. Let D, the task relevant data, be a set of sequential database where each sequence S is a set of items[8] (amino acid) such that $S \subseteq I$. Each amino acid is associated with an identifier, called SID (i.e. Sequential Identification). Let A be a set of items. A sequence S is said to contain A if and only if $A \subseteq S$. An association rule is an implication of the form $A \subseteq B$, where $A \subseteq I$, $B \subseteq I$ and $A \cap B = \Phi$. A is called antecedent, while B is called consequent; the rule specifies A implies B.

Basic measures for predicting association rules are support(s) and confidence which is illustrated in the following equations (1) and (2);

$$Support\ (A \Rightarrow B) = \frac{Sequence\ of\ A\ and\ B}{Total\ number\ of\ Sequences} \quad \text{--- (1)}$$

Confidence(c) of an association rule is defined as the percentage/fraction of the number of sequence that contain A U B to the total number of records that contain A, where if the percentage exceeds the threshold of confidence an interesting association rule A=>B can be generated. Confidence is a measure of strength of the association rules, suppose the confidence of the association rule A => B is

80%, it means that 80% of the transactions that contain A also contain B together, similarly to ensure the interestingness of the rules specified minimum confidence is also pre-defined by users. The rule A=>B has confidence c, in the transaction set D if c is the percentage of sequence in D containing A which also contain B. Rules that satisfy both a minimum support threshold (min sup) and a minimum confidence threshold (min_conf) are called strong.

$$Confidence\ (A => B) = \frac{Sequence\ of\ A\ and\ B}{Sequence\ of\ A} \quad \text{--- (2)}$$

Using the association rule to find the frequent its very useful for the gene finding, protein function domain detection, function motif detection, protein function inference, disease diagnosis, disease prognosis, disease treatment optimization, protein and gene interaction network reconstruction, data cleansing, and protein sub-cellular location prediction.

B. Genetic Algorithm & Fuzzy logic

Genetic algorithm methods for optimization. The continuing performance improvements of computa- tional systems have made them attractive for some types of optimization. All living organisms consist of cells. In each cell there is the same set of chromosomes. Chromosomes are strings of DNA and serves as a model for the whole organism. A chromosome consists of genes, blocks of DNA.

1) Inputs:

1. Medical database
2. Standard range of attributes(e.g.cholesterol, BP, etc.) from expertise
3. Rules generated from Apriori

2) Output:

Optimal values of attributes i.e. the best chromosome with highest fitness value

C. Algorithm:

1. [Start] Generate random population of n chromosomes (suitable solutions for the problem)
2. [New population] Create a new population by repeating following steps until the new population is complete
   a. [Selection] Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
   b. [Crossover] With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, off-spring is an exact copy of parents.
   c. [Mutation] With a mutation probability mutate new offspring at each locus (position in chromo-some).
   d. [Accepting] Place new offspring in a new population
3. [Replace] Use new generated population for a further run of algorithm
4. [Test] If the end condition is satisfied stop, and return the best solution in current population
5. [Loop] Go to step 2

The applications of GAs are for solving certain multi objective problems of bioinformatics, which yields optimization of computation requirements, and robust, fast and close approximate solutions. Moreover, the errors generated in experiments with bioinformatics data can be handled with the robust characteristics of GAs[9]. To some extent, such errors may be regarded as contributing to genetic diversity, a desirable property. The problem of integrating GAs and bioinformatics constitutes a new research area.

D. Fuzzy Logic in Bioinformatics

Fuzzy logic can be easily used to implement systems ranging from simple, small or even embedded Up to large networked ones. Fuzzy logic is that it accepts the uncertainties that are inherited in the realistic inputs and it deals with these uncertainties in their affect is negligible and thus resulting in a precise outputs. Fuzzy Logic reduces the design steps and simplifies complexity that might arise since the first step is to Understand and characterize the system behavior by using knowledge and experience. The concept of Fuzzy Logic (FL) was conceived by Lotfi Zadeh, FL provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information. It mimics human control logic.

IV. PROPOSED WORK AND IMPLEMENTATION

A. Proposed work

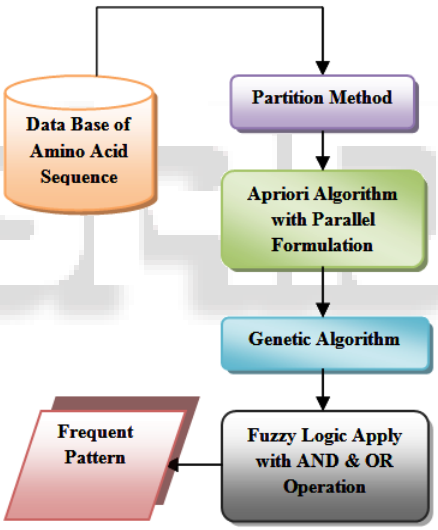


Fig. 2: Proposed work

1. Originally, the standard dataset of amino acid is taken form database.
2. The dataset is divided into partition.
3. The pre-processed data is given to the Apriori algorithm with parallel formulation for generation of associations rules and interesting correlations frequent pattern generated.
4. The results of apriori module are given to Genetic algorithm module. And generate the offspring using crossover operation. Using this offspring to produce the frequent pattern by apriori algorithm.
5. The results from genetic algorithm module and results from apriori algorithm module are given to the fuzzy logic module for generation of required knowledge of frequent patterns. And also optimize the frequent pattern these results see in implementation part.

Implementation

B. Upload Database file

Here is the basic one that simply gives functionality for uploading your input file. The input file which contain amino acid sequence.

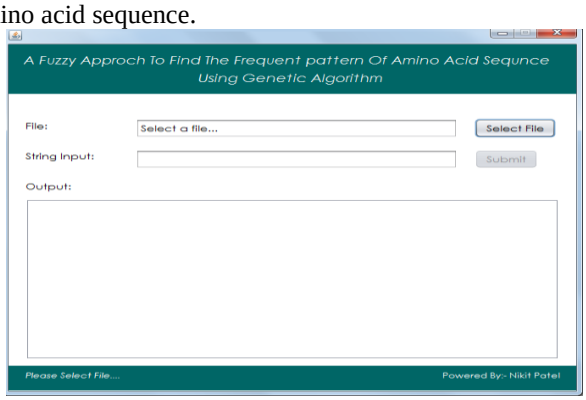


Fig 3: Browse input file.

C. Partition method

The partition method is used for generating transaction id for longer input string. This transaction id consider as input of apriori algorithm. Hence it is for uncomplicated calculation. So, here using partition method easy to manage the database of amino acid sequence.

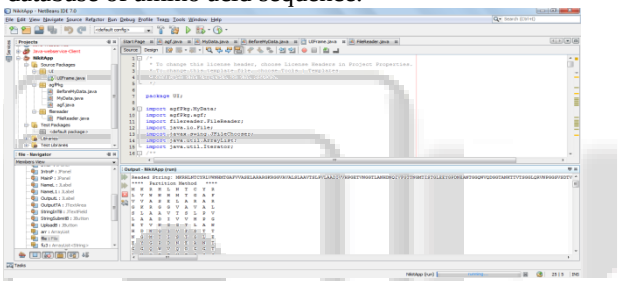


Fig 4: Partitioning Of Amino Acid

D. Calculation of Apriori Algorithm

Generation of frequent pattern using apriori algorithm. Here is the input seed from your selected input file. In this proposed system focus on predicting the most dominating amino acids than the other amino acids to cause the viral disease from the protein data sets and this input give to genetic algorithm. Select two parent id form this frequent pattern. So here is red line pattern selected as input of genetic algorithm.

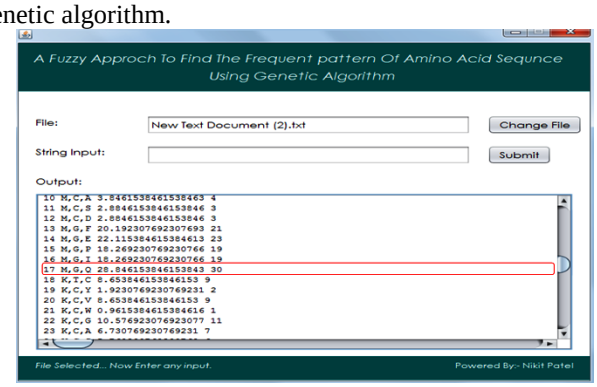


Fig 5: Calculation of frequent pattern by apriori algo.

E. Genetic Algorithm of Crossover operation And Enter the Parent item sets

First of all some random frequent pattern set will be selected from generated frequent pattern. First red mark is selected form fig5. After that the crossover operation performed on the string that will be input to String input field. The result of crossover operation will be input to apriori algorithm. And again frequent pattern create. You will see in fig6. And second red line pattern match with the apriori algorithm

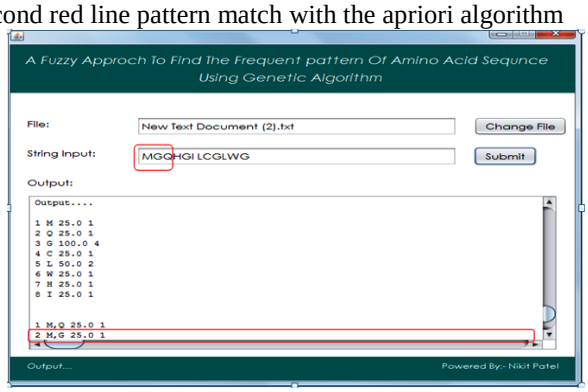


Fig 6: Crossover operation with Frequent item.

In this screenshot (fig 7) calculated frequent pattern by apriori algorithm. Red mark pattern and also remaining other pattern match with output of cross- over operation frequent pattern using fuzzy method.

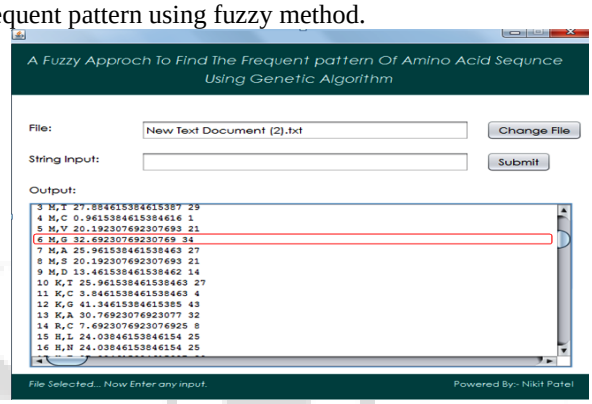


Fig 7: Calculation of frequent pattern by apriori algo.

F. Fuzzy method

The results from apriori algorithm module(fig 7) and results from genetic algorithm module(fig 6) are given to the fuzzy logic module for generation of required frequent patterns. In this fig 8 the match found item sets available in both module which has genetic algorithm module and apriori algorithm module. In this screenshot (fig 8) red mark rectangle pattern available in both modules. So this Pattern is very useful for to analysis include protein structure prediction, gene classification, cancer classification based on microarray data, clustering of gene expression data, statistical modeling of protein-protein interaction, etc. and also offspring also useful.

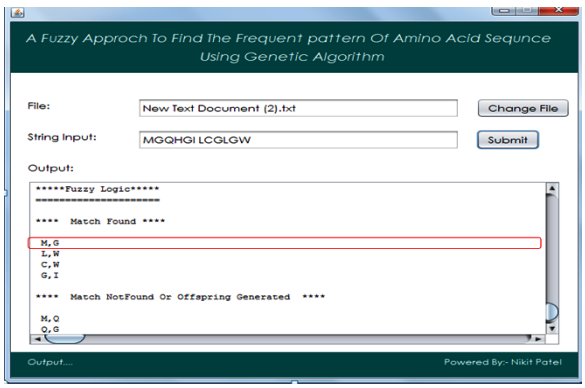


Fig 8: Comparison of end result of apriori and genetic algorithm.

V. CONCLUSION

Literature survey is the act of studying the existing system and analyzing the need for the system to be reengineered. In existing system only using association rule along with apriori algorithm using to predict the frequent item. And consider only confidence above 90% this frequent pattern is valid.

My approach is to predict the frequent pattern of amino acid using partition method, apriori algorithm, genetic algorithm and fuzzy logic. The results from genetic algorithm module and results from apriori algorithm module are given to the fuzzy logic module for generation of required knowledge of frequent patterns. And also optimize the frequent pattern. This Frequent Pattern is very useful to drug design, drug discovery etc. The main objective is predict frequent of amino acids could be more beneficial in preparing medicines to cure the disease.

In future this work could be extended to other protein sequence which causes for other viral disease like HIV, flu, Dengue fever, Viral fever, Swine flu, etc. Hence it is more beneficial in preparing medicines to cure the disease caused by the virus.

REFERENCES

[1] Jiawei Han, Hong Cheng, Dong Xin and Xifeng Yan, "Frequent pattern mining: current ststus and future directions", Data Mining Knowledge Discovery(2007) 15:55 -86.
[2] Lakshmi Priya. G., Shanmugasundaram Hariharan "A Study On Predicting Patterns Over The Protein Sequence Datasets Using Association Rule MINING", Journal Of Engineering Science And Technology Vol. 7, No. 5. (2012) 563 – 573
[3] Davnah Urbach And Jason H Moore, "Data Mining And Thev Evolution Of Biological Complexity", Biodata Mining 2011, 4:7.
[4] Lakshmipriya, Shanmugasundaram Hariharan," An Efficient Approach For Generating Frequent Patterns Without Candidate Generation", ACM, August 3-5, 2012.
[5] Khalid Raza,"Application Of Data Mining In Bioinformatics",Indian Journal Of Computer Science And Engineering August 2010,Vol 1 No 2, 114-118
[6] Amino Acid Http://En.Wikipedia.Org/Wiki/ Congenital_Disorders_Of_Amino_Acid_Metabolism
[7] Valerie Guralnik And George Karypis Parallel Formulations Of Tree-Projection-Based Sequence Mining Algorithm.
[8] Agrawal R. And Srikant R. (1994), "Fast Algorithms For Mining Association Rules", Proceedings Of The 20th International Conference On Very Large Data Bases. Pg.487–499, Santiago De Chile.
[9] Dr. Tryambak A. Hiwarkar,R. Sridhar Iyer, "New Applications Of Soft Computing, Artificial Intelligence, Fuzzy Logic & Genetic Algorithm In Bioinformatics", IJCSMC, Vol. 2, Issue. 5, May 2013, Pg.202 – 207.

