A New Hybrid Model to find The Dominant Pattern of Amino Acid Sequence to using Data Mining

Dimpal Prajapati, Prof. Riya Parmar
Computer Science & Engineering, LDRP-ITR, Gujarat, India

***Abstract*- Data Mining is the process of extracting or mining the patterns from very large amount of biological datasets. Utilization of Data mining algorithms can reveal biological relevant associations between different genes and gene expression. In Data Mining, several techniques are available for predicting frequent patterns. One among the technique is association rule mining algorithm; which can be applied for solving the crucial problems faced in the field of biological science. From the literature, various algorithms have been employed in generating frequent patterns for distinct application. These algorithms have some limitations in predicting frequent patterns, such as space, time complexity and accuracy. We consider the problem of discovering association rules between items in a large database of sales transactions. We present two new algorithms for solving this problem that are fundamentally different from the known algorithms. Bioinformatics is an interdisciplinary research area that is the interface between the biological and computational sciences. Bioinformatics deals with algorithms, databases and information systems, web technologies, artificial intelligence and information and computation theory, structural biology, software engineering, data mining, image processing, modeling and simulation, discrete mathematics, control and system theory, circuit theory, and statistics.**

***Index Terms*- Data mining, Bioinformatics, Partition Method, Apriori, Genetic Algorithm**

I. INTRODUCTION

Data mining refers extracting or "mining" knowledge from large amount of data. It is defined as "the process of discovering meaningful new correlations, patterns and trends by digging into large amounts of data stored in warehouses". Data Mining is called as Knowledge Discovery in Databases (KDD).As data sets have grown in size and complexity, the modern technologies of computers, networks and sensors have made data collection and organization much easier. However, the captured data needs to be converted into information and knowledge to become more useful. Data Mining is the entire from process of applying computer-based methodology, including new techniques for knowledge discovery from data.[1]

Data Mining approaches seem ideally suited for Biological Data Mining, since it is data-rich, but lacks a comprehensive 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.[1]

Data mining is the extraction of interesting relations and patterns hidden in the datasets. It combines database technologies, statistical analysis and machine learning[23].Today data mining techniques are innovatively utilized in numerous fields like industry, Commerce and Medicine. The data results generated by data mining techniques is highly priced by the professionals [24].

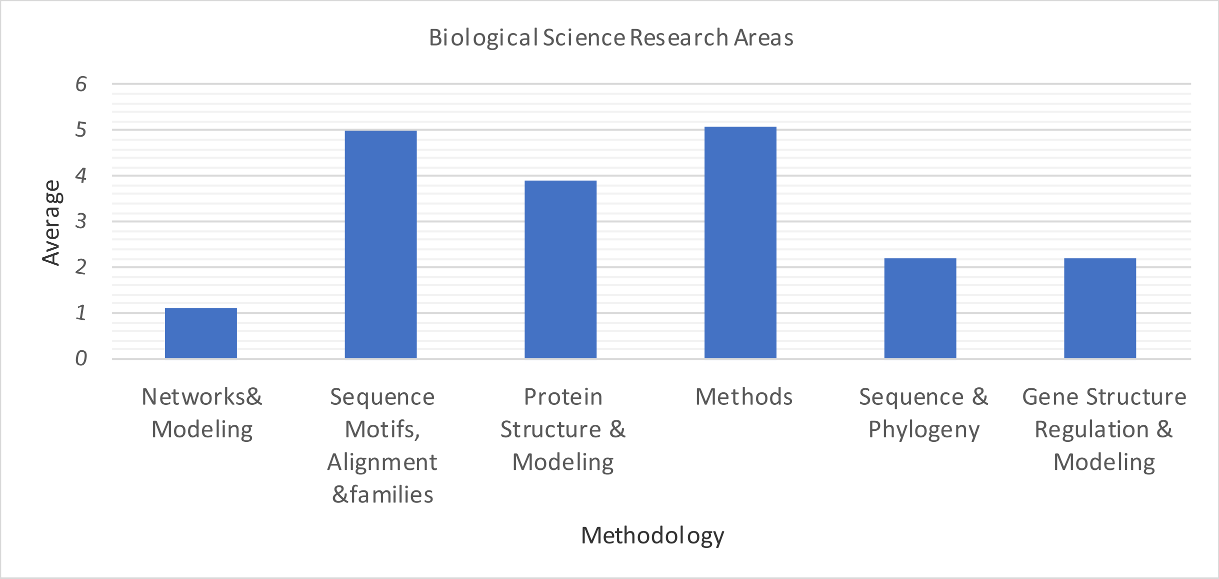

Chart-1: Average trends in Biological science research areas[1]

A crucial challenge in the future of bioinformatics involves putting that data to work. Now life scientists hope to plan large experiments, 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 average trends in bioinformatics research areas are shown in the following Fig.1.

In the past two decades the challenges faced by the research pupil in the field of biomedical is for an explosive growth of biomedical data (i.e., ranging from those collected in pharmaceutical studies and cancer therapy) investigations to those identified from genomics and proteomic research by discovering sequential patterns, gene functions, and protein interactions. The rapid progress of biotechnology and bio data analysis methods has led to the emergence and fast growth of promising new field coined as Bio Data Mining 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.

1.1 Amino Acids

Amino acids play central roles both as building blocks of proteins and as intermediates in metabolism. The chemical properties of the amino acids of proteins determine the biological activity of the protein. Proteins not only catalyze all or most of the reactions in living cells, they control virtually all cellular process. The general structure of an _-amino acid is depicted in Fig.2, Which represents the amino group on the left, the carboxyl group on the right and R a side chain to each amino acid..

1.2 Protein Stucture

Proteins are an important class of biological macromolecules present in all biological organisms, made up of elements such as carbon, hydrogen, nitrogen, phosphorous, oxygen and sulfur. The elements of a protein and the tertiary structure of protein are depicted in Fig. 3. There are four distinct aspects of a protein structure such as Primary structure, Secondary structure, Tertiary structure and Quaternary structure.

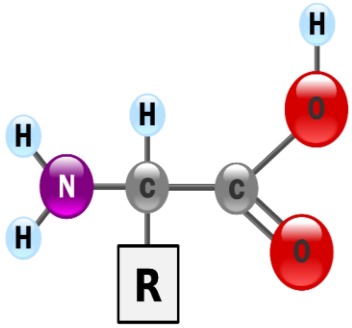

Figure 2. Structure of α- amino Acid[1]

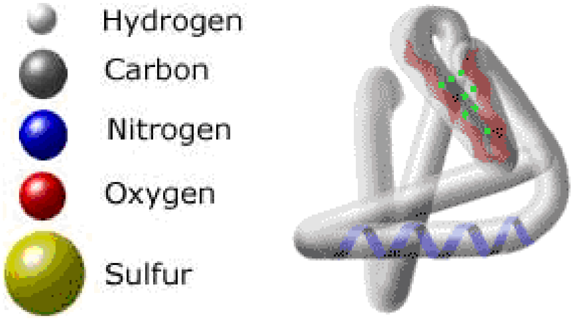

Figure 3. The elements and tertiary structure of a protein[1]

1.3 General Structure and functions of Amino Acids

Amino acids are the building blocks of proteins. 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. Amino acids combine in a condensation reaction that releases water and the new amino acid residue that is held together by a peptide bond. Proteins are defined by their unique sequence of amino acid residues; amino acids can be linked in varying sequences to form a vast variety of proteins. Twenty standard amino acids are used by cells in protein biosynthesis, and these are specified by general genetic code. The Table I illustrates the list of essential amino acids and Table II shows the list of non- The one-letter and three-letter codes for amino acids used in the knowledgebase are those adopted by the commission on Biochemical Nomenclature of the IUPAC-IUB

TABLE I. LIST OF AMINO ACIDS

Amino Acid	3-Letter	1-Letter
Arginine	Arg	R
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Threonine	Thr	T
Tryptophan	Trp	W
Valine	Val	V

TABLE II. LIST OF AMINO ACIDS

Amino Acid	3-Letter	1-Letter
Alanine	Ala	A
Asparagine	Asn	N
Aspartate	Asp	D
Cysteine	Cys	C
Glutamate	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Proline	Pro	P
Serine	Ser	S
Tyrosine	Tyr	Y

1.4 Chapter Organization

Chapter 1 deal with introduction to Data mining concepts, functionalities, and applications of data mining in the field of Biological science its protein structure and amino acids. Chapter 2 deal with introduction to bioinformatics. Chapter 3 refines the technique of finding a frequent item set of DNA dataset. Chapter 4 the recovery concludes and future investigation on mining frequent itemsets as the outlined research work.

II. BIOINFORMATICS

The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems. It was primary used since late 1980s has been in genomics and genetics, particularly in those areas of genomics involving large-scale DNA sequencing.[3]

Bioinformatics is the application of computer technology to the management and analysis of biological data.[2][3] The result is that computers are being used to gather, store, analyze and merge biological data. [4]The ultimate goal of bioinformatics is to uncover the wealth of biological information hidden in the mass of data and obtain a clearer insight into the fundamental biology of organisms. This new knowledge could have profound impacts on fields as varied as human health, agriculture, the environment, energy and biotechnology.[25] Over the past few decades rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. The primary goal of bioinformatics is to increase the understanding of biological processes.

2.1 Task Of Bioinformatics

Different biological problems considered within the scope of bioinformatics involve the study of genes, proteins, nucleic acid structure prediction, and molecular design with docking. A broad classification of the various bioinformatics tasks is given as follows.[2]

1. Alignment and comparison of DNA, RNA, and protein sequences.
2. Gene mapping on chromosomes.
3. Gene finding and promoter identification from DNA sequences.
4. Interpretation of gene expression and micro-array data.
5. Gene regulatory network identification.
6. Construction of phylogenetic trees for studying evolutionary relationship.
7. DNA structure prediction.
8. RNA structure prediction.
9. Protein structure prediction and classification.
10. Molecular design and molecular docking.

2.2 Application of Bioinformatics

Bioinformatics has found its applications in many areas. It helps in providing practical tools to explore proteins and DNA in number of other ways. Bio-computing is useful in recognition techniques to detect similarity between sequences and hence to interrelate structures and functions. Another important application of bioinformatics is the direct prediction of protein 3-Dimensional structure from the linear amino acid sequence. It also simplifies the problem of understanding complex genomes by analyzing simple organisms and then applying the same principles to more complicated ones. This would result in identifying potential drug targets by checking homologies of essential microbial proteins. Bioinformatics is useful in designing drugs..

The aims of Bioinformatics are:

1. To organize data in a way that allows researchers to access existing information and to submit new entries as they are produced
2. To develop tools and resources that aid in the analysis and management of data.
3. To use this data to analyze and interpret the results in a biologically meaningful manner.
4. To help researchers in the pharmaceutical industry in understanding the protein structures to make the drug design easy.

The first three trends can be viewed as instances of pattern matching. However, pattern matching in biology differs from its counterpart in computer science. DNA strings contain millions of symbols, and small local differences may be tolerated. The pattern itself may not be exactly known, because it may involve inserted, deleted, or replacement symbols. Regular expressions are useful for specifying a multitude of patterns and are ubiquitous in bioinformatics. However, what biologists really need is to be able to infer these regular expressions from typical sequences and establish the likelihood of the patterns being detected in new sequences.[2]

Two other algorithmic trends relevant to this discussion are related to micro-arrays and biologists' interest in computational linguistics. Recall that the main goal of analyzing micro-array data is to establish relationships among gene behavior, possible protein interactions, and the effects of a cell's environment. From a computer science perspective, that goal is amounted to the generation of parts of a program (flowchart) from data. This was also an early goal of program synthesis. However, it should be stressed that biological data is vast and noisy, spurring development of new heuristic based techniques is required.[2]

III. PROPOSED SYSTEM AND ARCHITECTURE

In this survey, using a three method,

1) Asocial Rule Mining Method
2) Optimization Method

3.1 Apriori Algorithm

The classical apriori algorithm is one of the most conventional and significant data mining algorithms which are used for identifying the relationships among attributes of transactions using binary values. However, mostly the real life applications have transactions with quantitative values which are not binary.[4]

In Proposed method,

1) Minimum Support – It's the percentage of task-relevant data transactions for which the pattern is true.

2) Minimum Confidence threshold – It's the measure of certainty associated with the pattern.

3) By Applying a algorithm To Generate The Frequent Item-set.

4) Extract the Dominant pattern From the dataset to apply a asocial rule mining method.

Advantages

1. Uses large itemset property
2. Easily parallelized
3. Easy to implement

3.2 Genetic Algorithm

GAs is executed iteratively on a set of coded solutions, called population, with three basic operators: selection/reproduction, crossover, and mutation. They use only the payoff (objective function) information and probabilistic transition rules for moving to the next iteration. Of all the evolutionarily inspired approaches, Gas seem particularly suited to implementation using DNA, protein, and other bioinformatics tasks [18]. This is because GAs are generally based on manipulating populations of bit-strings using both crossover and point-wise mutation.

Advantages

1. Several tasks in bioinformatics involve optimization of different criteria (such as energy, alignment score, and overlap strength), thereby making the application of Gas more natural and appropriate.

2. Problems of bioinformatics seldom need the exact optimum solution; rather, they require robust, fast, and close approximate solutions, which GAs are known to provide efficiently.

3. GAs can process, in parallel, populations billions times larger than is usual for conventional computation. The usual expectation is that larger populations can sustain larger ranges of genetic variation, and thus can generate high-fitness individuals in fewer generations.

4. Laboratory operations on DNA inherently involve errors. These are more tolerable in executing evolutionary algorithms than in executing deterministic algorithm

3.3 Proposed Architecture

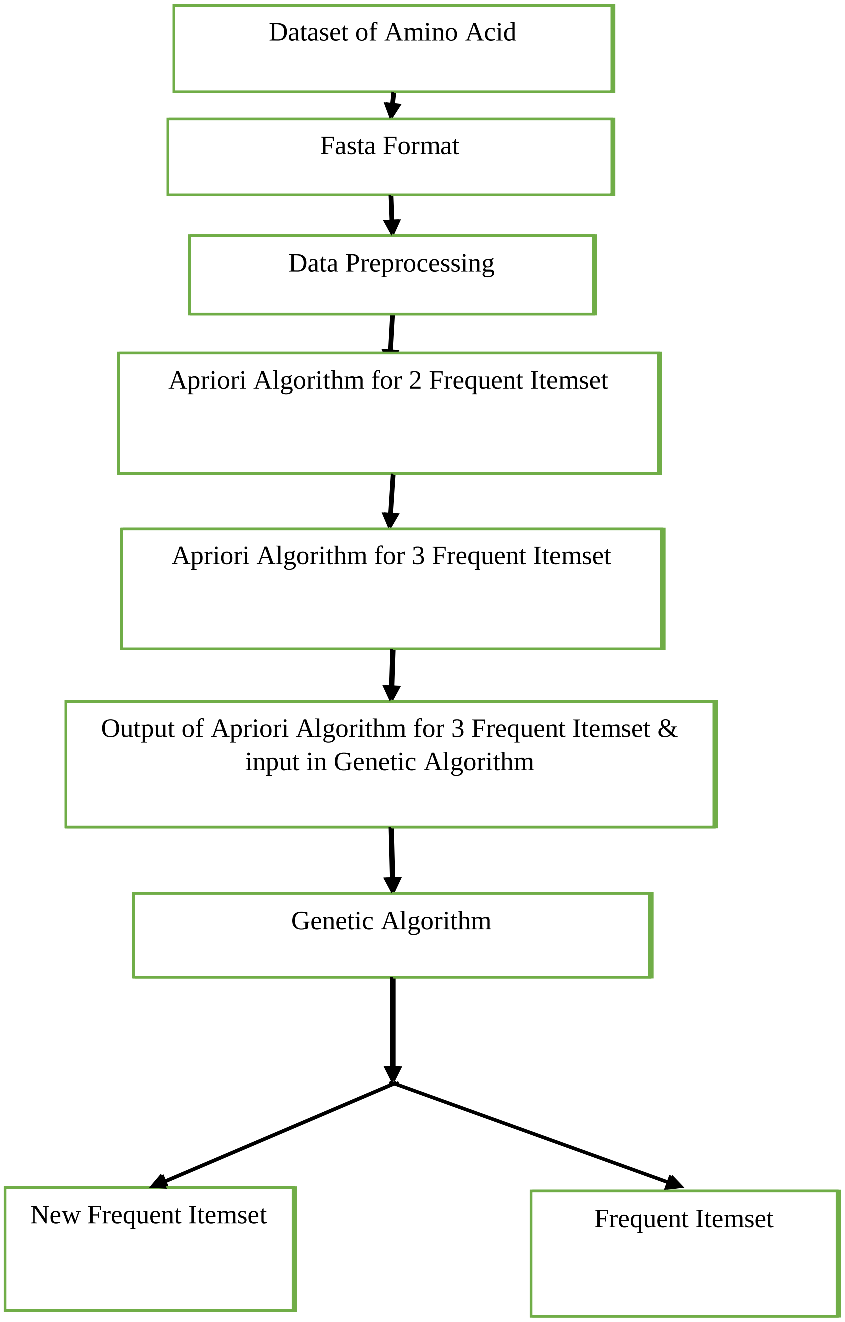
Dataset of Amino Acid
Fasta Format
Data Preprocessing
Apriori Algorithm for 2 Frequent Itemset
Apriori Algorithm for 3 Frequent Itemset
Output of Apriori Algorithm for 3 Frequent Itemset & input in Genetic Algorithm
Genetic Algorithm
New Frequent Itemset
Frequent Itemset

Figure 4. Proposed work

IV. RESULTS

4.1 Graph of 2 Frequent Itemsets:

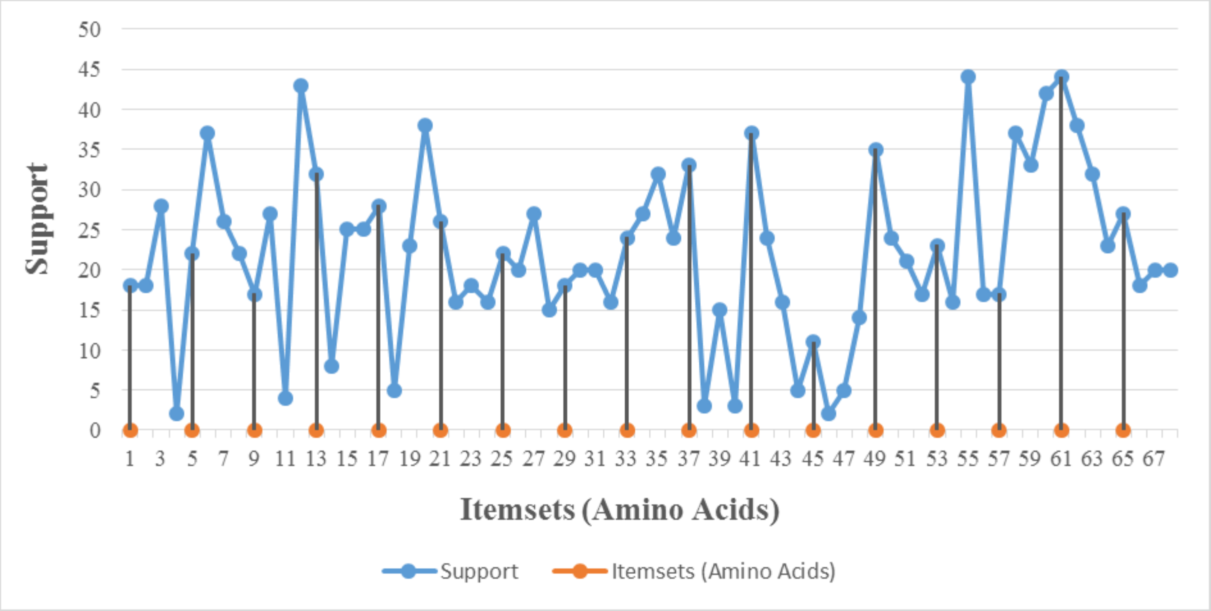

Figure 5. Graph of 2 Frequent Itemsets

4.2 Graph of Genetic Algorithm Frequent Itemsets:

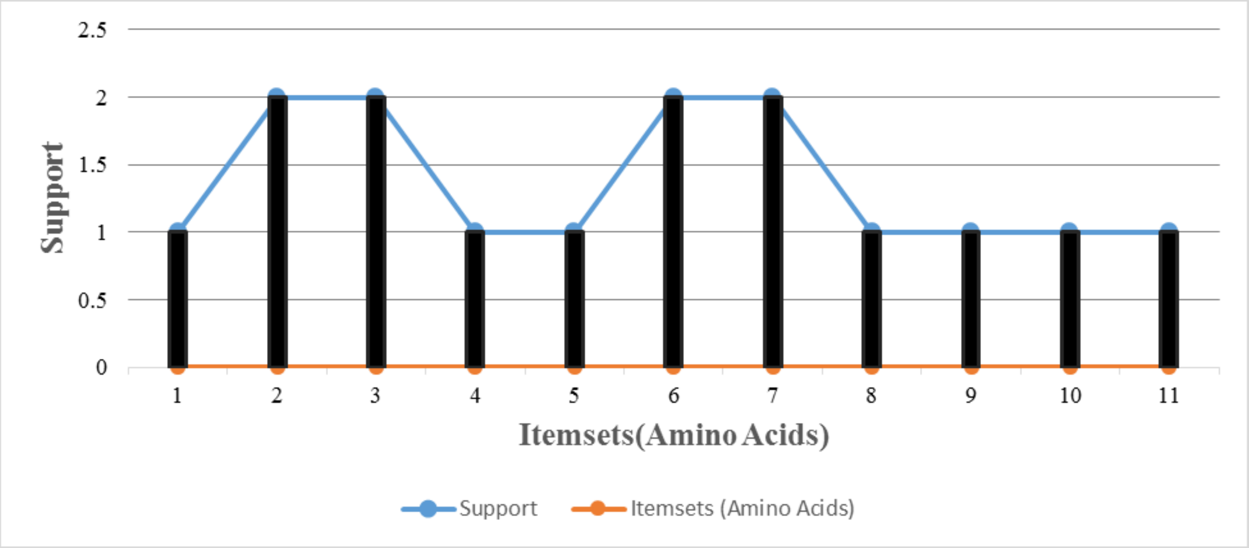

Figure 6. Graph of Genetic Algorithm Frequent Itemsets

4.3 Graph of Match Found and Match Not Found Frequent Itemsets:

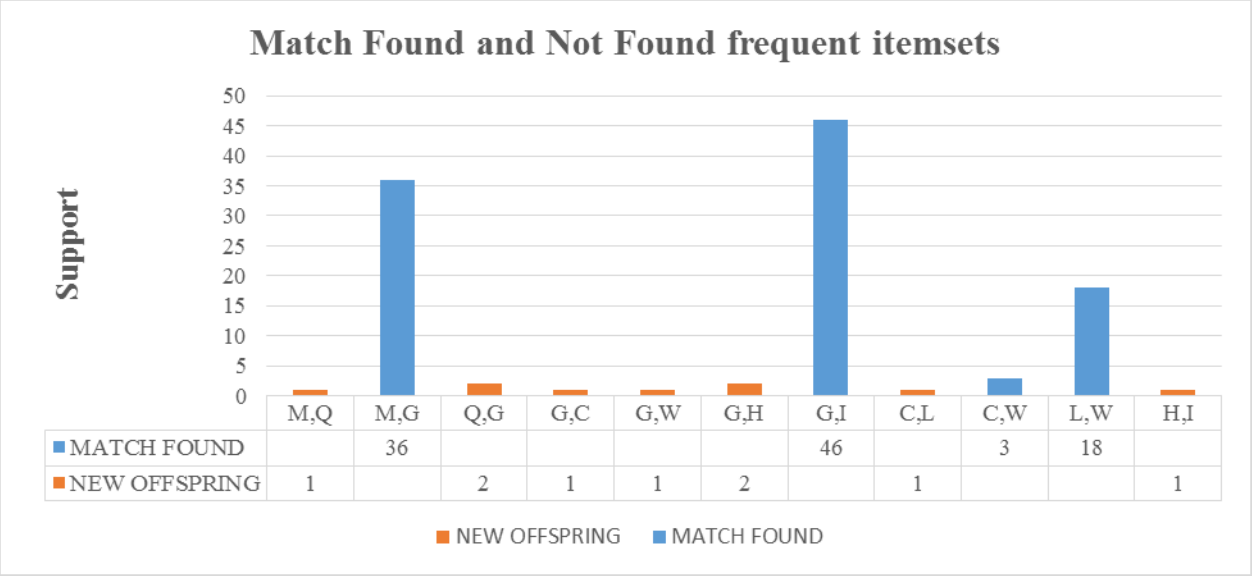

Figure 7. Graph of Match Found and Match Not Found Frequent Itemsets

V. CONCLUSIONS

From the study of literature, it is known that an efficient algorithm is required to predict frequent pattern. The survey concludes that frequent itemsets could be generated from a clustered protein sequence which causes the viral disease in human. Various protein sequences could be applied on the proposed system which is being developed for identifying the dominating amino acids. Further investigation will be involved by mining frequent itemsets with and without candidate generation and their results will be compared with the predicted results

VI. FUTURE WORK

In future this work could be extended to other protein sequence which causes other viral disease like flue, Dengue Fever, viral fever, swine flu, etc. Finally this work helps and it is more beneficial in preparing medicines to cure the disease caused by these viral infections during the case of emergency.

REFERENCES

[1] G. Lakshmi Priya,Department of Computer Science & Engineering, J. J College of Engineering and Technology, Tiruchirapalli, Tamil Nadu, India, 978-1-4673-0671-3/11/$26.00©2011 IEEE

[2] IJCEM International Journal of Computational Engineering & Management, Vol. 16 Issue 1, January 2013 38, ISSN (Online): 2230-7893,www.IJCEM.org

[3] Khalid Raza / Indian Journal of Computer Science and Engineering, Vol 1 No 2, 114-118

[4] K. Rajeswari, Mahadev Shindalkar, Nikhil Thorawade, Pranay Bhandari / International
Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622,www.ijera.com Vol. 3, Issue 4, Jul-Aug 2013, pp.132-136

[5] Han, Jiawei, and Micheline Kamber. Data mining: concepts and techniques. Morgan Kaufmann, 2006.

[6] R. Lemmon and M. C. Milinkovitch (2002), "The metapopulation genetic algorithm: An efficient solution for the problem of large phylogeny estimation", Proc. Nat. Acad. Sci., Vol. 99, No. 16, pp. 10516–10521.

[7] Shapiro, J. C. Wu, D. Bengali and M. J. Potts (2001), "The massively parallel genetic algorithm for RNA folding: MIMD implementation and population variation", Bioinformatics, vol. 17, no. 2, pp. 137–148.

[8] Chen, L. H. Wang, C. Kao, M. Ouhyoung and W. Chen (1998), "Molecular binding in structure- based drug design: A case study of the population based

annealing genetic algorithms", In Proc. IEEE Int. Conf. Tools with Artificial Intelligence, pp. 328–335.

[9] Zhang (1994), "A genetic algorithm for molecular sequence comparison", In Proc. IEEE Int. Conf. Systems, Man, and Cybernetics, Vol. 2, pp. 1926–1931.

[10] E. Clark and D. R. Westhead (1996), "Evolutionary algorithms in computer aided molecular design", J. Comput.-Aided Mol. Design, Vol. 10, No. 4, pp. 337–358.

[11] Goldberg (1989), "Genetic Algorithms in Optimization, Search, and Machine Learning", Reading, MA: Addison-Wesley.

[12] Hatzigeorgiou, A. G. Reckzo, M. (2004), "Signal peptide prediction on DNA sequences with artificial neural networks", Biomedical Circuits and Systems.

[13] J. Setubal and J. Meidanis (1999), "Introduction to Computational Molecular iology", Boston, MA: Thomson.

[14] J. M. Yang and C. Y. Kao (2000), "A family competition evolutionary Algorithm for automated docking of flexible ligands to proteins", IEEE Trans. Inf. Technol. Biomed., Vol. 4, No. 3, pp. 225–237.

[15] J. W. Fickett (1996), "Finding genes by computer: The state of the art", Trends Genetics, Vol. 12, No. 8, pp. 316–320.

[16] Leping Li 1,*, Yu Liang 1 and Robert L. Bass (2007), "GAPWM: a genetic algorithm method for optimizing a position weight matrix", Bioinformatics, 23(10): 1188-1194.

[17] P. Baldi and P. F. Baisnee (2000), "Sequence analysis by additive scales: DNA structure for sequences and repeats of all lengths", Bioinformatics, Vol. 16, pp. 865–889.

[18] S. B. Needleman and C. D. Wunsch (1970), "A general method applicable to the search for similarities in the amino acid sequence of two proteins", J. Mol. Biol., Vol. 48, pp. 443–453.

[19] S. Schulze-Kremer (2000), "Genetic algorithms and protein folding. Methods in molecular biology", Protein Structure Prediction: Methods and Protocols, Vol. 143, pp. 175–222.

[20] T. Hou, J. Wang, L. Chen and X. Xu (1999), "Automated docking of peptides and proteins by using a genetic algorithm combined with a tabu search", Protein Eng., Vol. 12, pp. 639–647.

[21] T. F. Smith and M. S. Waterman (2001), "Identification of common Informatics: Edmonton", AB, Canada: IMIA, pp. 83– 100.

[22] T. Murata and H. Ishibuchi (1996), "Positive and negative combination effects of crossover and mutation operators in sequencing problems", Evol.Comput., Vol. 20–22, pp. 170–175.

[23] Thuraisingham, B., 2000. A primer for understanding and applying data mining, IT professional.IEEE Computer Society, pp. 28–31.

[24] Tang, T.I., Zheng, G., Huang, Y.,Shu, G., Wang, P., 2005. A comparative study of medical data classification methods based on decision tree and system reconstruction analysis. IEMS 4 (1), 102–108.

[25] WWW.123seminarsonly.com

Swati Jain received the BE degree in Computer Science Engineering, and the ME degree in Computer Engineering, from the RGPV university Bhopal, in 2007 and 2010 respectively. Currently she is an assistant professor at the department of Computer Science, Takshshila Institute of Technology, Jabalpur [M.P.], India. Her research interests include soft computing, artificial intelligence, neural network, fuzzy logic, genetic algorithm, computer graphics, data structure and software engineering. She is a member of the ISTE.

Abhishek Pandey BE [I.T.], ME[C.S.E] is an assistant professor at the Department of Computer Science, Takshshila Institute of Technology, Jabalpur[M.P.], India. His research interests include computer graphics, data structure and software engineering.