Journal of Engineering Science and Technology
Vol. 7, No. 5 (2012) 563 - 573
© School of Engineering, Taylor's University

A STUDY ON PREDICTING PATTERNS OVER THE PROTEIN SEQUENCE DATASETS USING ASSOCIATION RULE MINING

LAKSHMI PRIYA. G.[1,*], SHANMUGASUNDARAM HARIHARAN[2]

[1]Oxford Engineering College, Tiruchirappalli, Tamil Nadu, India
[2]Department of CSE,TRP Engineering College (SRM Group), India
*Corresponding Author: glpriya.jjcet@gmail.com

Abstract

Data Mining has recently increased its popularity of solving crucial problems in the field of biological science. Nowadays a large quantity of data and information about biological issues and its environments has been accessed by individual, organization, business, family or institution. A critical problem in biological data analysis is to classify the biological sequences and structures based on their critical features and functions. Protein is one among the important factor and acts as the constituents of all living organisms. Protein plays the most predominant role for causing viral diseases like viral fever, fluid diseases, poliomyelitis, hepatitis, swine flu, tumor, etc. The proposed system focus onto discover the most dominating amino acids which causes the chromaffin tumor. Finally the dominating patterns were predicted from a clustered protein sequence Succinate dehydrogenase - DHSB_HUMAN protein chain with 100% confidence threshold and the results were quite promising in the field of bio-medicine.

Keywords: Data mining, Association rule, Chromaffin tumor, Protein sequence.

1. Introduction

Data Mining is a collection of techniques for efficient automated discovery of patterns from large data sets. The patterns must be actionable so that they may be used for decision making process [1]. It's not necessary to follow a typical data mining process such as requirement analysis, data selection and collection, cleaning and preparing data, data mining exploration and validation, implementing, evaluating and monitoring or result visualization. Data Mining employs a number of techniques which includes supervised classification, cluster analysis, association rule mining, web data mining, etc.

A case study is carefully carried out for better understanding the applications of data mining and its process. The prediction of the functional behavior of proteins has been an important challenge in modern functional Proteomic. Such prediction is made possible by the motifs in protein chains. Since more than one motif may exist within a protein chain, the correlation between protein properties and their motifs may not be always obvious.

From the literature, various data mining approach [2] has been developed for classifying the protein sequence. The recent trends in data mining applications are being carried out in the following disciplines, such as astronomy, banking and finance, climate, crime prevention, Direct mail service, health care, telecommunications, etc. Data mining techniques* that are likely to become important in the future and it determines *interestingness* of a discovered pattern in the data. Other techniques that are likely to receive more attention in the future are text and web content mining, Bioinformatics and multimedia data mining.

In other word it has been described as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data [3] and the science of extracting useful information from large datasets [4] or database. As datasets have grown in size and complexity, there has been a shift away from direct hands-on data analysis toward indirect, automatic data analysis using more complex and sophisticated tools. The modern technologies of computers, networks and sensors have made data collection and organization much easier. However, the captured data need to be converted into information and knowledge to become useful.

Data mining is the entire process of applying computer-based methodology, including new techniques for knowledge discovery from data. The greatest achievement in Bioinformatics is, without any doubt, the decoding of the DNA. Knowledge of the genetic code enables a better understanding of the mechanism for the creation of all proteins necessary for a cell. However knowing how a protein is created and it does not provide any information on how it will be used. The greatest challenge in modern bioinformatics [5] is the extract and reliable modeling of protein behavior and functionality.

1.1. Amino acids

Amino acids play central roles both as building block of proteins and as intermediates in metabolism [6]. The 20 amino acids that are found within proteins convey a vast array of chemical versatility. The precise amino acid content, and the sequence of those amino acids, of a specific protein, is determined by the sequence of the bases in the gene that encodes that protein.

1.2. General structure of standard amino acids

The general structure of an α-amino acid is illustrated in Fig. 1, which represents the amino group towards the left, the carboxyl group on the right and R a side chain to each amino acid[1]. The central carbon atom, called C_α, is a chiral central carbon atom (with the exception of glycine) to which the two termini and the R-group are attached. Amino acids are usually classified by the properties of the side chain into four groups. The chemical structures of the 20 standard amino acids, along with their chemical properties, are catalogued in the list of standard amino acids.

*Amino acids are the building blocks from which proteins are constructed.

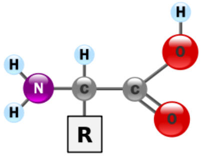


Fig. 1. Structure of an a-Amino Acid.

1.3. Protein structure

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 proteins are depicted in Fig. 2.

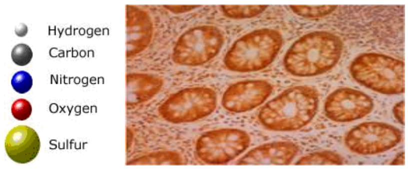


Fig. 2. Elements of Protein.

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 [7] of nucleotides in the gene that codes the protein. Table 1 displays the 280 amino acid sequence that forms the P21912 protein chain [8]. Each amino acid in the protein sequence [9] is represented by its formal abbreviation.

Table 1. Protein Sequence Chain of P21912 [8].

Protein Sequence			
10	**20**	**30**	**40**
MAAVVALSLR	RRLPATTLGG	ACLQASRGAQ	TAAATAPRIK
50	**60**	**70**	**80**
FAIYRWDPD	KAGDKPHMQT	YEVDLNKCGP	MVLDALIKIK
90	**100**	**110**	**120**
NEVDSTLTFR	RSCREGICGS	CAMNINGGNT	LACTRRIDTN
130	**140**	**150**	**160**
LNKVSKIYPL	PHMYVIKDLV	PDLSNFYAQY	KSIEPYLKKK
170	**180**	**190**	**200**
DESQEGKQQY	LQSIEEREKL	DGLYECILCA	CCSTSCPSYW
210	**220**	**230**	**240**
WNGDKYLGPA	VLMQAYRWM	DSRDDFTEER	LAKLQDPFSL
250	**260**	**270**	**280**
YRCHTIMNCT	RTCPKGLNPG	KAIAEIKKMM	ATYKEKKASV

2. Related Work

Mining association rules between sets of items in large databases [10] uses the large database of customer transactions. Each transaction consists of items purchased by a customer in a visit. An efficient algorithm that generates all significant association rules [11] between items in the database has been presented. The algorithm incorporates buffer management and novel estimation and pruning techniques. Present results using this algorithm were applied to sales data to develop their sales which show the effectiveness of the algorithm obtained from a large retail company.

2.1. Data mining in bioinformatics

PROTEAS: A Finite State Automata based data mining algorithm for rule extraction in protein classification has proposed an important challenge in modern functional proteomics is the prediction of the functional behavior of proteins [2]. A data mining approach for a motif-based class of proteins has been discussed. A new classification algorithm that includes rules and exploits finite state automata has been introduced. First, data are modeled by the terms of prefix tree acceptors, which are later merged into finite state automata. Finally, a new algorithm was proposed, for the induction of protein classification rules from finite state automata. The data mining model is trained and tested using various protein and protein class subsets, as well as the whole dataset of known proteins and protein classes. Results indicate the efficiency of the technique compared to other known data mining algorithm.

2.2. Extraction of knowledge on protein--protein interaction

Extraction of knowledge on protein--protein interaction by association rule discovery has proposed protein--protein interactions, systematically examined using the yeast two hybrid methods [12]. Consequently, a lot of Protein--protein interaction data is currently being accumulated. Nevertheless, general information or knowledge on protein--protein interactions is poorly extracted from these data. Thus they have been trying to extract the knowledge from the protein--protein interaction data using data mining. A data mining method is proposed to discover association rules [13] related to Protein--protein interactions. To evaluate the detected rules by the method, a new scoring measure of the rules is introduced. The method allowed detecting popular interaction rules such as, an SH3 domain binds to a protein-rich region. These results indicate that the method may detect novel knowledge on protein--protein interactions.

3. Implementation

In the proposed system a viral disease is taken and the associative patterns among the amino acids are identified using a data mining technique. To generate the strong association rules from the amino acids which cause the disease is taken into consideration with 90% and above as its confidence value and support count may range between 2 to 5. Based on the strong association rules, 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.

3.1. Methods

To facilitate subsequent discussion, the main symbols used throughout the paper and their definitions are summarized in Table 2. The main objective in the proposed algorithm Apriori is to produce the frequent items from a data set *D* of *n* protein sequences.

Table 2. Symbols and Definitions.

Symbols	Definitions
D	Data set of protein sequences to find the frequent item sets.
C_k	Candidate item sets of size *k*
L_k	Frequent item sets of size *k*
n	Total number of items in transactional dataset
TID	Transaction Identification Number

3.2. Apriori algorithm

Apriori is a classic algorithm for learning association rules. Apriori is designed to operate on databases containing transactions (TIDs). Apriori uses a bottom up approach, where frequent subsets are extended one item at a time and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a tree structure to count candidate item sets efficiently. It generates candidate item sets of length k from item sets of length k–1.

Then it prunes the candidates which have an infrequent sub pattern. According to the downward closure lemma, the candidate set contains all frequent k-length item sets. After that, it scans the transaction database to determine frequent item sets among the candidates. Candidate generation generates a large number of subsets. Bottom-up subset exploration finds any maximal subset S only after all 2 |S|-1 of its proper subsets.

3.3. Apriori itemset generation

Items may be generated as follows: (i) A frequent itemset L_k is an itemset whose support is greater than some user-specified minimum support. (ii) A candidate itemset C_k is a potentially frequent itemset.

3.4. Implementation of apriori algorithm

For implementing the proposed system, 280 attributes of proteins are considered. Each attribute is related to one of the amino acid. The steps used in the algorithm implementation are as follows:

3.4.1. Partitioning the attributes

After selecting the training set (protein Sequence) divide the 280 attributes into 28 intervals.

3.4.2. Assigning transaction IDs

Each interval is given a unique TID[2] as shown in Table 3.

3.4.3. Finding the frequent items using candidate generation

Apriori is an influential algorithm for mining frequent itemsets for Boolean association rules. To improve the efficiency of searching, the candidate itemsets C_k are stored in a hash tree. The leaves of the hash tree store itemsets while the internal nodes provide a roadmap to reach the leaves. Each leaf node is reached by traversing the tree whose root is at depth 1.

Each internal node of depth d points at all the related nodes at depth. The name of the algorithm is based on the fact that the algorithm uses prior knowledge

TID: Transactional IDentification number in the protein sequence P21912.

of frequent itemset properties. Apriori employs an iterative approach known as a *level-wise* search, where k-itemsets are used to explore (*k+1*)-itemsets.

First, the set of frequent 1-items are found. This set is denoted as $L_{1.}$ L_1 is used to find $L_{2,}$ the set of frequent 2-itemsets, which is used to find L_3 and so on, until no more frequent k-itemsets can be found. The finding of each L_k requires one full scan of the data base. Figure 3 shows the pseudo-code for apriori algorithm.

Table 3. Transactional Protein Datasets for Chromaffin Tumor.

TID	List of Transactional_Ids									
T10	M	A	A	V	V	A	L	S	L	R
T20	R	R	L	P	A	T	T	L	G	G
T30	A	C	L	Q	A	S	R	G	A	Q
T40	T	A	A	A	T	A	P	R	I	K
T50	K	F	A	I	Y	R	W	D	P	D
T60	K	A	G	D	K	P	H	M	Q	T
T70	Y	E	V	D	L	N	K	C	G	P
T80	M	V	L	D	A	L	I	K	I	K
T90	N	E	V	D	S	T	L	T	F	R
T100	R	S	C	R	E	G	I	C	G	S
T110	C	A	M	N	I	N	G	G	N	T
T120	L	A	C	T	R	R	I	D	T	N
T130	L	N	K	V	S	K	I	Y	P	L
T140	P	H	M	Y	V	I	K	D	L	V
T150	P	D	L	S	N	F	Y	A	Q	Y
T160	K	S	I	E	P	Y	L	K	K	K
T170	D	E	S	Q	E	G	K	Q	Q	Y
T180	L	Q	S	I	E	E	R	E	K	L
T190	D	G	L	Y	E	C	I	L	C	A
T200	C	C	S	T	S	C	P	S	Y	W
T210	W	N	G	D	K	Y	L	G	P	A
T220	V	L	M	Q	A	Y	R	W	M	I
T230	D	S	R	D	D	F	T	E	E	R
T240	L	A	K	L	Q	D	P	F	S	L
T250	Y	R	C	H	T	I	M	N	C	T
T260	R	T	C	P	K	G	L	N	P	G
T270	K	A	I	A	E	I	K	K	M	M
T280	A	T	Y	K	E	K	K	A	S	V

```
$L_1$ = find_frequent 1-itemsets ($D$);
            For ($k=2; L_{k-1} \neq \varphi; k$++
{           $C_k$ = apriori _gen ($L_{k-1}$);
For each transaction $t \in D$
{           $C_t$ =subset ($C_{k,}t$);
For each candidate $c \in C_t$
                        $c$.count++;
}$L_k$ = {$c \in C_k$ | $c$.count $\geq$ min_sup}
} return $L = U_k\ L_k$
```

Fig. 3. The Apriori Algorithm for Discovering Frequent Itemsets.

3.4.4. Generating association rules from frequent itemsets

Once the frequent itemsets from transactional protein datasets is loaded from the text file, it is straightforward to generate strong association rules. This can be performed using the following equation for measuring the confidence, where the conditional probability is expressed in terms of itemset support _count.

$$Confidence(\mathrm{A} \cup \mathrm{B}) = \mathrm{P}(\mathrm{B} \mid \mathrm{A}) \tag{1}$$

$$\mathrm{P}(\mathrm{B} \mid \mathrm{A}) = \frac{Support\_count(\mathrm{A} \cup \mathrm{B})}{Support\_count(\mathrm{A})} \tag{2}$$

where the *Support_count* (AUB) is the number of transactions containing the itemsets, AUB, *Support_count* (A) is the number of transactions containing the itemset A.

Based on this equation, association rules can be generated as follows:

- For each frequent itemset, *l*, generate all nonempty subsets of *l*.
- For every nonempty subset *s* of *l* the rule is *s* = (1 – *s*).

3.4.5. Generation of strong association rules for (P, I, K) & (P, N, L) amino acids

After the Apriori process, the frequent itemsets are retrieved as follows: {A, L, R}, {A, L, D}, {A, P, K}, {A, P, D}, {A, K, D}, {L, P, K}, {L, P, Y}, {L, P, D}, {L, P, N}, {L, I, K}, {L, I, Y}, {L, K, Y}, {L, K, D}, {L, Y, D}, {L, D, N}, {P, I, K}, {P, K, Y}, {P, K, D}, {P, Y, D} and {K, Y, D}}.

To generate strong association rules, Let X= {*P, I, K*} and Y= {*P, N, L*}, Find all the subsets of X and Y after omitting the empty set and the set itself. The subsets of X and Y are as follows,

$$X = \{\{P\},\{I\},\{K\},\{P,I\},\{I,K\},\{P,K\}\} \tag{3}$$

$$Y = \{\{P\},\{N\},\{L\},\{P,N\},\{N,L\},\{L,P\}\} \tag{4}$$

The resulting association rules are illustrated in the above Table 4, with the existence of strong association among the amino acids which causes the chromaffin tumor. The association rule may be considered to be strong enough in this case. Because the confidence threshold of P ^ N → L and P ^ I → K (100 %) is greater than the minimum confidence threshold that is 90%. Hence Proline (P) and Asparagine (N) are strongly associated with Leucine (L) and similarly Proline (P) and Isoleucine (I) is strongly associated with Lysine (K). The results have been obtained from the known clustered sequence of unknown patterns.

Table 4. Generation of Strong Association Rules Using Associative Measures.

Association Rule	Confidence *c*	Compare *c* with minimum confidence	Result
L → P ^ N	20%	20 %<90%	Rejected
P → L ^ N	35%	35% < 90%	Rejected
N → L ^ P	45%	45% < 90%	Rejected
P ^ N → L	100%	100% > 90%	**Accepted ***
L ^ N → P	71%	71% < 90%	Rejected
L ^ P → N	55%	55% < 90%	Rejected
P → I ^ K	35%	35% < 90%	Rejected
I → P ^ K	31%	31% < 90%	Rejected
K → P ^ I	20%	20% < 90%	Rejected
I ^ K → P	62%	62% < 90%	Rejected
P ^ K → I	50%	50% <90%	Rejected
P ^ I → K	100%	100% > 90%	**Accepted ***

4. Results and Discussion

Experiments were conducted on an Intel Pentium4 processor based machine, having a clock frequency of 2.4 GHZ and 512 MB of RAM. The Apriori algorithm has been implemented using Microsoft Visual C++ 6.0. Experimental results were obtained using the transaction datasets. In the proposed system the number of iterations is not fixed. The algorithm terminates when no further successful extensions are found.

4.1. Analysis and discussion

Before loading the Protein data sets, divide their 280 attributes into 28 intervals. Load the transaction Protein datasets for obtaining the frequent itemsets. The process terminates when no further successful iteration were found. The results are obtained from the frequent itemsets L_k. In the proposed system, the *threshold Support_count* value may be fixed between the range 2 to 5 and Confidence is fixed above 90%.

Figures 4 to 6 illustrate the frequent itemsets generation. Figure 7 shows the associative pattern discovery process. To predict the frequent patterns, select the table *aprior1*. *Text* by clicking the *browse menu* in order to execute the table which consists of *clustered protein Sequence P21912* with *280 amino acids* and the minimum support count specified as 5. It has the facility of avoiding the eliminated itemsets by clicking the check box in order to generate frequent patterns using the Apriori process. Here the report files Count1 has the frequent itemsets L_1, Count2 has the frequent itemsets L_2 and Count3 has the frequent itemsets L_3. From the generated frequent itemsets L_3, the strong association rules are applied in order to find the dominating amino acids.

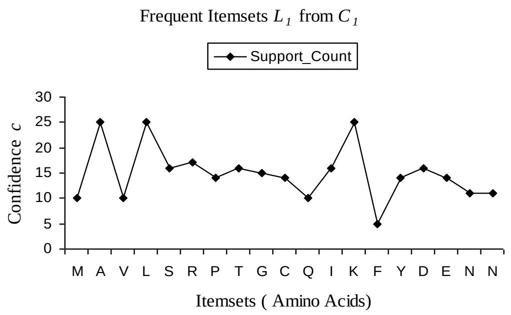


Fig. 4. Frequent 1- itemsets or L_1.

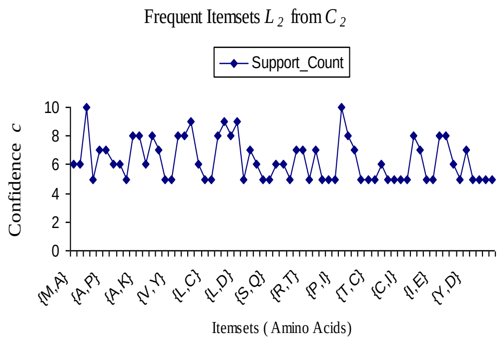


Fig. 5. Frequent 2-itemsets or L_2.

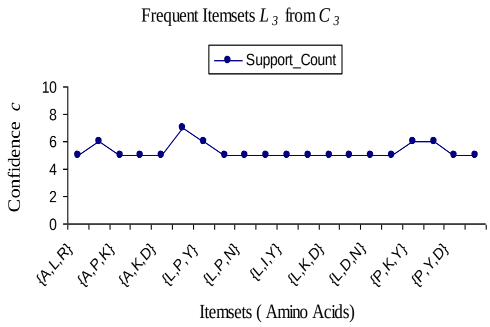


Fig. 6. Frequent 3-itemsets or L_3.

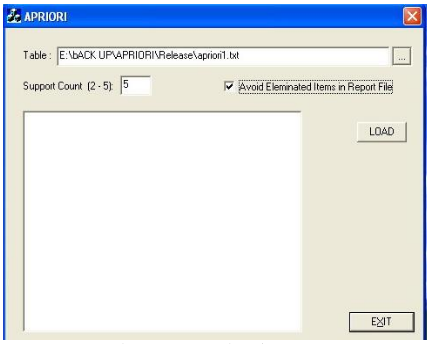


Table : E:\bACK UP\APRIORI\Release\apriori1.txt

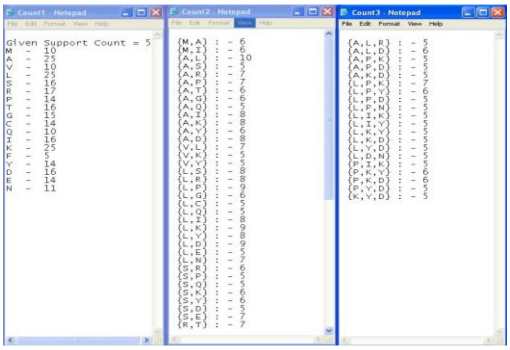
Support Count (2 - 5): 5

☑ Avoid Eleminated Items in Report File

Fig. 7. Apriori Process.

4.2. Experimental results

Experimental results using the existing algorithm on Chromaffin Tumor* is summarized in Table 4*. The resulting patterns illustrates the frequent itemsets generated from the transactional protein chain *Succinate dehydrogenase - DHSB_HUMAN*. These patterns were predicted using an Apriori process with the specified minimum *Support_count* as 5. The generated frequent itemsets can be viewed as reports as shown in Fig. 8. The Apriori process maintains three report files namely; Count 1, Count 2 and Count 3. Figure 9 illustrates the confidence measures after all the passes. Table 4 shows the confidence measures after applying the association rules for the obtained frequent itemsets from the report file Count 3.


Count1 - Notepad
File Edit Format View Help
Count2 - Notepad
File Edit Format View Help
Count3 - Notepad
File Edit Format View Help

```
Given Support Count = 5
M   -   10
A   -   25
V   -   10
L   -   25
S   -   16
R   -   17
P   -   14
T   -   16
G   -   15
C   -   14
Q   -   10
I   -   16
K   -   25
F   -   5
Y   -   14
D   -   16
E   -   14
N   -   11
```

```
{M,A} : - 6
{M,I} : - 6
{A,L} : - 10
{A,S} : - 5
{A,R} : - 7
{A,P} : - 7
{A,T} : - 6
{A,G} : - 6
{A,Q} : - 5
{A,I} : - 8
{A,K} : - 8
{A,Y} : - 6
{A,D} : - 8
{V,L} : - 7
{V,K} : - 5
{V,Y} : - 5
{L,S} : - 8
{L,R} : - 8
{L,P} : - 9
{L,G} : - 6
{L,C} : - 5
{L,Q} : - 5
{L,I} : - 8
{L,K} : - 9
{L,Y} : - 8
{L,D} : - 9
{L,E} : - 5
{L,N} : - 7
{S,R} : - 6
{S,P} : - 5
{S,Q} : - 5
{S,K} : - 6
{S,Y} : - 6
{S,D} : - 5
{S,E} : - 7
{R,T} : - 7
```

```
{A,L,R} : - 5
{A,L,D} : - 6
{A,P,K} : - 5
{A,P,D} : - 5
{A,K,D} : - 5
{L,P,K} : - 7
{L,P,Y} : - 6
{L,P,D} : - 5
{L,P,N} : - 5
{L,I,K} : - 5
{L,I,Y} : - 5
{L,K,Y} : - 5
{L,K,D} : - 5
{L,Y,D} : - 5
{L,D,N} : - 5
{P,I,K} : - 5
{P,K,Y} : - 6
{P,K,D} : - 6
{P,Y,D} : - 5
{K,Y,D} : - 5
```

Fig. 8. Report Generation for the Frequent Itemsets L_1, L_2 and L_3.

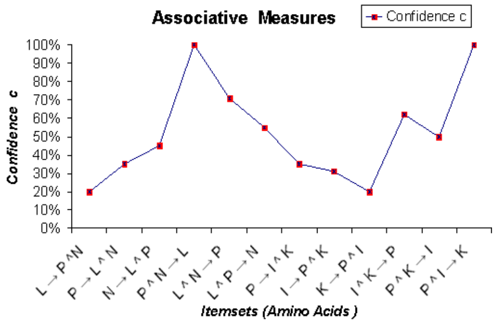


Fig. 9. Association Rules with Highest Confidence.

5. Conclusion and Future work

This case study concerns the understanding of associations between the amino acids behavior and patterns predicted from the known clustered sequence of unknown patterns. From the literature its clear that Apriori works in breath first search. The main objective is to predict the dominating amino acids for causing any kind of viral diseases which could be a similar kind of analysis applied to different problems. The aim of the analysis is to track the most influential patterns*. The chosen measures of importance determine the results of the analysis. For this study an existing prairie algorithm mining was utilized for evaluating the frequent itemsets using Candidate generation. This technique is time consuming process though it generates a candidate key for each and every pass. To overcome this problem, new techniques can be adopted to reduce the running time in the future. The proposed system is focused on finding the most dominating amino acids, which causes the viral disease named chromaffin tumor. The frequent itemsets are generated from a clustered protein sequence Succinate dehydrogenase - DHSB_HUMAN chain. Among the generated frequent itemsets few amino acids are found to be strongly associated. Using the results retrieved from the clustered protein sequence, a focus is given on which amino acids are more dominating by forming association rules. Further investigation will be involved by mining frequent itemsets without candidate generation and the results would be compared with the predicted results. Finally the predicted amino acids could be more beneficial in preparing medicines to cure the disease caused by the chromaffin cells.

Acknowledgement

The authors would like to express their heartfelt thanks to the management, head of IT department and colleagues of Oxford Engineering College, Trichy and TRP Engineering College, Trichy for providing a supporting environment. It is a great pleasure to thank the almighty and our family members for their encouragement in carrying out our research work in a successful manner.

References

1. Gupta, G.K. (2006). *Introduction to data mining with case studies*. PHI Learning Pvt. Ltd.
2. Psomopoulo, F.E.; and Mitkas, P.A. (2004). PROTEAS: A finite state automata based data mining algorithm for rule extraction in protein classification. *Proceedings of the* 5th *Hellenic Data Management Symposium* (*HDMS*), 118-126.
3. Han, J.; Kamber, M.; and Pei, J. (2011). *Data mining: Concepts and techniques*. (3rd Ed.), Morgan Kaufmann.
4. Agrawal, R.; and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. *Proceedings of the* 20th *International Conference on Very Large Data Bases*, *VLDB'*94, 487-499.
5. Luscombe, N.M.; Greenbaum, D.; and Gerstein, M. (2001). *What is Bioinformatics? An introduction and overview*. Yearbook of Medical Informatics.
6. The biology project, Biochemistry. *The Chemistry of amino acid*. http://www.biology.arizona.edu/biochemistry/problem_sets/aa/aa.html.
7. Bosu, O.; and Thukral, S.K. (2007). *Bioinformatics: Databases, tools and algorithms*. Oxford University press.
8. Protein sequence chain of P21912. http://www.uniprot.org/uniprot/P21912.
9. Wu, K.-P.; Lin, H.-N.; Sung, T.-Y.; and Hsu, W.-L. (2003). A new similarity measure among protein sequences. *IEEE proceedings of the Computational Systems Bioinformatics* (CBS'03), 347-352.
10. Agrawal, R.; Imielinski, T.; and Swami, A.M. (1993). Mining association rules between sets of items in large databases. *Proceedings of the* 1993 *ACM SIGMOD International Conference on Management of Data* SIGMOD, 207-216.
11. Kotsiantis, S.; and otiris and Kanellopoulos, D. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1), 71-82.
12. Oyama, T.; Kitano, K.; Satou, K; and Ito, T. (2002). Extraction of knowledge on protein--protein interaction by association rule discovery. *Bioinformatics*, 18(5), 705-714.
13. Mannila, H.; Toivonen, H.; and Verkamo, A.I. (1994). Efficient algorithms for discovering association rules. *AAAI*-94 *Workshop on Knowledge Discovery in Databases*, 181-92.