Predicting the Lung Cancer from Biological Sequences

Kalaiyarasi R.
Department of computer Science and Engineering
Jayaram College of Engineering and Technology, Trichy, Tamilnadu, India

Prabasri S.
Department of computer Science and Engineering
Jayaram College of Engineering and Technology, Trichy, Tamilnadu, India

Abstract ——**Cancer is the most important cause of death for both men and women. The occurrence of lung cancer has increased rapidly and become the most common cancer in men in most of the countries. Bio Data mining is the process of extracting non-trivial, implicit, previously unknown and potentially useful information or patterns from large amount of biological sequences. There are several data mining techniques are used one among them association rule mining is one of the most popular technique which is currently used in the field of biological science. Its purpose is to find the association among the item sets in the biological datasets. With the wide spread of lung cancer causing death all over, here comes the necessity to predict the dominant amino acids for causing the lung cancer and to take immediate preventive actions.. In Association rule mining, several algorithms are available for predicting frequent patterns. But few algorithms have certain drawbacks such as time complexity, space complexity and cost. These drawbacks could be rectified by our new proposal. Therefore a new data mining technique offers the greatest promise at this time for fighting the lung cancer**

***Keywords*-Data Mining, Amino acids, Lung cancer, Association Rule Mining**

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). . 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. Data mining functionalities mainly focuses on classification, prediction, clustering and association analysis. Classification is the process of finding a set of models which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. Prediction is the process of finding the missing numerical values or increase/decrease trends in time related data. The major idea is to use a large number of past values to consider probable future values. Clustering is the organization of data in classes. In clustering, class labels are unknown and it is up to the clustering algorithm to discover acceptable classes.

A.Association Rule Mining

Let I = {i_1, i_2,…..,i_m} be a set of items. Let D, the task relevant data, be a set of database transactions where each transaction T is a set of items such that T $\subseteq$ I. Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if A $\subseteq$ T. An association rule is an implication of the form A=> B, where A $\subseteq$ I, B $\subseteq$ I and A ∩ B = Φ. A is called antecedent, while B is called consequent; the rule specifies A implies B.There are two important basic measures for association rules, support(s) and confidence(c).Support(s) of an association rule is defined as the percentage/fraction of records that contain (A U B) to the total number of records in the database. The count for each item is increased by one every time the item is encountered in different transaction T in database D during the scanning process. It means the support_count does not take the quantity of the item into account.

The rule A => B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain A U B. That is,

$$Support\ (A \Rightarrow B) = \frac{Transactions\ containing\ both\ A\ and\ B}{Total\ number\ of\ transactions} \qquad (1.1)$$

Confidence(c) of an association rule is defined as the percentage/fraction of the number of transactions 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 transactions in D containing A which also contain B. That is,

$$Confidence\ (A \Longrightarrow B) = \frac{Transactions\ containing\ both\ A\ and\ B}{Transactions\ containing\ A} \quad (1.2)$$

Rules that satisfy both a minimum support threshold (min sup) and a minimum confidence threshold (min conf) are called strong.

B.Bio Data Mining

Bio Data Mining 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 to those identified in genomics and proteomic research by discovering sequential patterns, gene functions, and protein-protein interactions. The rapid progress of biotechnology and bio data analysis methods has led to the emergence and fast growth of a promising new field is Bio Data Mining.

C.Applications of Data Mining in Biological Science

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.

D.Amino acids

Amino acids are the building blocks (monomers) of proteins. 20 different amino acids are used to synthesize proteins. The shape and other properties of each protein are dictated by the precise sequence of amino acids in it. The chemical properties of the amino acids of proteins determine the biological activity of the protein.

Amino acids are at the basis of all life processes, as they are absolutely essential for every metabolic process. Among their most important tasks are the optimal transport and optimal storage of all nutrients (i.e. water, fat, carbohydrates, proteins, minerals and vitamins). The majority of diseases such as obesity, high-cholesterol levels, diabetes, insomnia, erectile dysfunction or arthritis can essentially be traced back to metabolic disturbances. The Figure 1.1 shows the Amino Acid structure. Each amino acid consists of an alpha carbon atom to which is attached

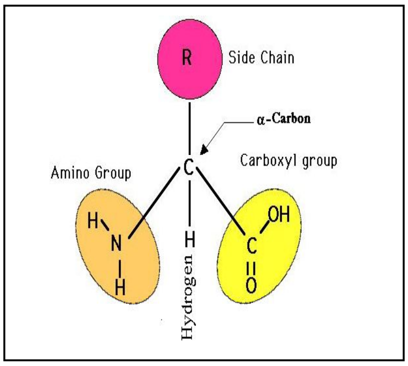

Figure 1.1 Amino Acid structure

Twenty standard amino acids are used by cells in protein biosynthesis, and these are specified by the general genetic code. These 20 amino acids are biosynthesized from other molecules, but organisms differ in which ones they can synthesize and which ones must be provided in their diet.

The Table 1.1 shows the List of Essential Amino Acids and Table 1.2 shows the List of Non-Essential Amino Acids. Non-essential amino acids are those which can be produced from other amino acids and substances in the diet and metabolism.

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 1.1 List of Essential 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

Table 1.2 List of Non-Essential Amino Acids

E.Proteins

Over 10,000 different proteins are found in the human body for maintaining life.75% of human body weight consists of protein. Eukaryotes synthesize amino acids. Amino acid synthesizes protein and protein synthesizes essential nutrients.

Proteins are organic compounds made of amino acids arranged in a linear chain and folded into a globular form. Proteins are an important class of biological macromolecules present in all biological organisms, made up of such elements as carbon, hydrogen, nitrogen, phosphorus, oxygen, and sulfur. All proteins are polymers of amino acids.

The various levels of protein structure are shown below in Figure1.2.

- Primary structure
- Secondary structure
- Tertiary structure
- Quaternary structure

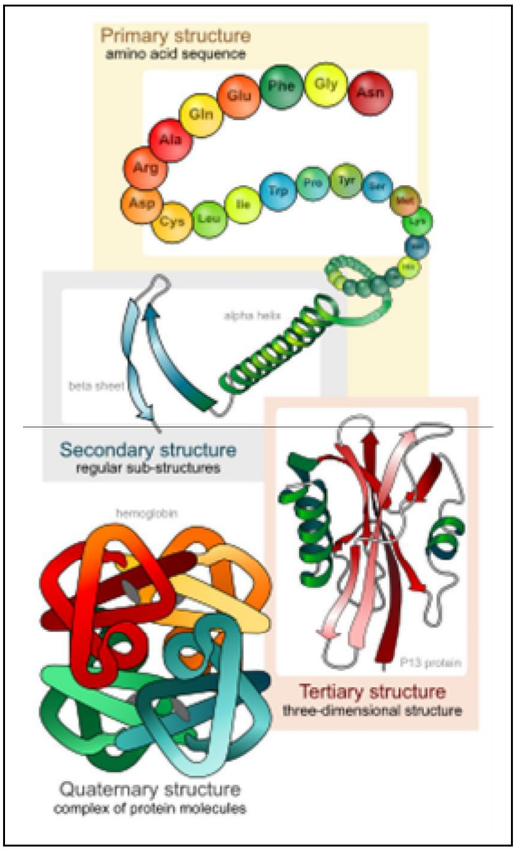

Figure 1.2 Levels of Protein Structure

II. LITERATURE SURVEY

Literature survey is the act of studying the existing system and analyzing the need for the system to be reengineered. It is the first and foremost stage in the design of software and is defined as "a study of the operations or a set of connected elements and of the inner connection between the elements".

In 2010, Anandhavalli et al [1] has proposed two major goals for analyzing massive genomic data: (i) To determine how the expression of any particular gene might affect the expression of other genes, (ii) To determine what genes are expressed as a result of certain cellular conditions using association and clustering concepts. selected an efficient algorithm to facilitate these analysis, the number of passes were not a major factor to be considered. Finally the author has concluded that the number of genes in one single transaction was very large.

In 2014, Chien-Hung Huang et al [7] has proposed differentially expressed genes are identified via an expression dataset generated from lung adnecarcinoma tumor and adjacent non-tumor tissues using protein-protein interaction network , micro-array data analysis and cluster analysis with up and down regulated communities. Input taken from 18 patients cancerous and non-cancerous tissues with 41 samples and they findings systematic strategy to discover potential drugs and target genes for lung cancer. From which eight drugs from drugBank and three drugs from NCBI. Less number of patients considered for analysis and takes more time and cost to consume the entire process.

In 2014, Shoon Lei Win et al. [8] proposed survivability prediction and micro-array gene expression to examine the thousands of genes simultaneously. Dimensionality reduction is applied for selection and discretization process. Taking the input as the physical sample from lung cancer affected patient's body then mining the micro-array gene expression based on entropy-based gene selection, entropy-minimization discretization and prediction finally the result will be used to predict whether the patient will survive or will not survive.. This approach to predict the lung cancer survivability with accuracy of 92.31% .It provides less computational complexity and more generalization capability of the system. This study showed by inadequate in predicting lung cancer development and clinical outcome.

In 2013, Parag Deoskar et al. [10] proposed the SPACO (Support Based Ant Colony Optimization) technique for lung cancer symptoms .Input data taken from UCI repository and relevant patterns are extracted then choose frequent symptoms by support count value. The support value decides ant and pheromone value. Each trail updated the pattern of cancer symptom with pheromone value. It concludes either increasing or decreasing the prediction level of patient and help in detecting lung cancer and improves the accuracy.

In 2014, Dr.D.P.Shukla et al [9] reviewed about various data mining techniques such as clustering, classification, regression, association rule mining, CART(Classification And Regression Tree) are widely used in Healthcare domain. This approaches used to improve the quality of prediction and diagnosis with respect to various diseases like cancer, cardio vascular abnormalities and others are using various algorithms for example genetic algorithm, association rule mining, k-means clustering, Naïve Bayesian classification applied in huge volume of medical data

In 2009, Mao et al. [2] explores the application of Apriori-Gen algorithm on the disease association study which has a large data set and tries to find the association among multiple Single Nucleotide Polymorphisms (SNPs) that may be responsible for the disease. The association of multi-SNP combination can be measured by risk ratio and odds ratio. The goal of disease association study is to assess accumulated information targeted to find interaction of multi-SNPs which are associated to complex diseases with significantly high accuracy and statistical power. The proposed method was applied to analyze the genetic data of the type 2 diabetes. This will save a huge amount of money and time for diagnosis and can be done before the onset of diseases. Therefore, treatment could start earlier to prevent or delay the occurrence of disease.

In 2007, Martinez et al. [3] has proposed GENMINER approach for extracting the association rules from genomic data. GENMINER was a co-clustering and Bi-clustering approach that integrates gene annotations and gene expressions to discover intrinsic associations among both data sources based on co-occurrence patterns. GENMINER follows the four steps of the ARD process: (i) data selection and pretreatment, (ii) frequent itemsets extraction, (iii) association rules generation, (iv) interpretation of extracted rules. It uses the NORDI algorithm for minimal non-redundant rules extraction. GENMINER approach takes less time and memory fro the correlated data compared to the Apriori ARD algorithm.

III.LUNG CANCER

The body is made up of human body up of trillions of living cells. Normal body cells grow, divide into new cells, and die in orderly way but the cancer cell growth is different from normal cell growth. Instead of dying, the cancer cells keep on growing and form new cancer cells. In most cases the cancer cells form a tumor.

More people die of lung cancer of colon, breast and prostate cancers combined. From the survey of 2014 about 2, 29, 210 new cases and 1, 59,260 deaths from lung cancer.

Several risk factors are such as smoking tobacco-including cigarettes, cigars and pipes, second hand smoke ,Radon, Having family member with lung cancer, arsenic in drinking water and etc.Some of the symptoms of lung cancer are headache,dizziness,chest pain,hoarseness,Jaundice and etc[9].Some of the lung cancers cause a group of very specific symptoms. These are often described as syndromes such as horner [9] and etc.Several treatments are available such as surgery, Radiofrequency ablation,Chemotherapy,Radiation therapy, Targeted therapy. Among these treatments the tumor can be removed however the cancer cell keep on growing and form new cancer cells. Lung cancer research is going on now in medical centers around the world. Lung cancer research for the most part focus to eradicate cancer cells or stop them growing

IV.CONCLUSION AND FUTURE WORK

From the literature survey, we narrow down for solving the crucial problems in biomedical applications. In the biomedical field huge volume of data sets are available. Several algorithms are for finding the frequent patterns from the biological sequences and several other techniques are used for predicting the lung cancer. But, it takes more time complexity and also gives less efficiency. The proposed model uses the efficient frequent pattern algorithm to mine the most frequent patterns from the given input dataset. The proposed model used to find the most dominating amino acid sequence to block the cancer cell's growth from the clustered protein sequence .Finally, the predicted amino acids could be more beneficial in preparing medicine to cure the lung cancer. Single drug kill cancer cells or stop them from further growing and it may help some people live longer .current research in lung cancer is looking at various protein sequence such as ALK Tyrosine Kinase, Histone deacetylase and Ral Protein Sequence which are used to block the growth of cancer cell further.

REFERENCES

[1] Anandhavalli, IACSIT, IAENG, Ghose, Gauthaman (2010) "Association Rule Mining in Genomics" International Journal of Computer Theory and Engineering, Vol. 2, No. 2

[2] Weidong Mao, Jinghe Mao (2009) "The Application of Apriori-Gen Algorithm in the Association Study in Type 2 Diabetes".

[3] Ricardo Martinez, Claude Pasquier, Nicolas Pasquier (2010) "GENMINER: Mining Informative Association Rules from Geenomic Data"IEEE International Conference on Bioinformatics and Biomedicine.

[4] Sandro Da Silva Camargo (2002) "MiRABIT: A New Algorithm for Mining Association Rules", Proceedings of the 12th International Conference of the Chilean Computer Science Society.

[5] Savasere, Omiecinski, and Navathe (2004) "Mining for Strong Negative Associations in a Large Database of Customer Transactions", Proceeding of the 14th International Conference on Data Engineering IEEE Computer Society Press, pp.494-502.

[6] Vaibhav Sharma, Sufyan Beg (2010) "A Probabilistic Approach to Apriori Algorithm" IEEE International Conference on Granular Computing.

[7] Chien-Hung Huang, Min-You Wu, Peter Mu-Hsin Chang, Chi-Ying Huang, Ka-Lok Ng,(2014),"In silico identification of potential targets and drugs for non-small cell lung cancer", IET Systems Biology,Vol.8,Issue 2,2014.

[8] Shoon Lei Win, Zaw Zaw Htike, Faridah Yusof, Ibrahim A. Noorbatcha, ,"Gene expression mining for survivability of patients in early stages of lung cancer", International Journal of Bioinformatics and Biosciences,Vol.4,No.2,June 2014.

[9] Dr. D. P. Shukla, Shamsher Bahadur Patel, Ashish Kumar Sen,"A literature review in Health Informatics using Data Mining Techniques", International Journal of Software and Hardware Research in Engg,ISSN.NO:2347-4890,Vol 2,Feb 2014.

[10] Parag Deoskar, Dr. Divakar Singh, Dr. Anju Singh," An Efficient Support Based Ant Colony Optimization Technique for Lung Cancer data", International Journal of Advanced Research in Computer and Communication Engineering,Vol.2,Issue No.9,Sep 2013.

[11] Monali Dey and Siddharth Swarup Rautaray(2014),'Study and Analysis of Data mining Algorithms for Healthcare Decision Support System', (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5,470-477, ISSN:0975-9646.

[12] Noel Blessy R. and Mohamed Amanllah K. (2013), 'Oral cancer detection using apriori algorithm', International Journal of Advanced Research in Computer and Communication Engineering,Vol. 3, Issue 7.