Mining Association Rules on DNA Sequences

Tin Tin Nwe
Computer University (Mandalay)
tintinnwe123@gmail.com

Abstract

Database mining is the process of extracting interesting and previously unknown patterns and correlations from data stored in Database Management System (DBMS). Association rule mining is the process of discovering items, which tend to occur together in transactions. Biological data mining becomes an essential part of bioinformatics. We identify DNA sequence pattern and obtain association rule from these frequently occurred DNA sequence item sets. A linear string or sequence of DNA is translated into sequence of amino acids. In this system, frequent item sets will be generated from DNA sequences datasets using FP-tree. We outline mining sequential patterns. The association rules we employ have the ability to extract the frequent pattern in particular genetic disease. The rules of interest are those whose set of frequent patterns are strongly associated to occur genetic disease

1.Introduction

All DNA sequences are comprised of four basic building blocks called: Adenine (A), Cytosine(C), Guanine (G), and Thymine (T). These four nucleotides are combined to form long sequences. A gene comprises hundreds of individuals nucleotides arranged in a particular order. These can be ordered and sequenced in an almost unlimited numbers of ways to form distinct genes. Changes in DNA sequences can result in disease. Mutation in several different genes can also lead to some clinical consequence such as Alzheimer's disease. Most diseases are not triggered by a single gene but by a combination of genes acting together. Association analysis method used to help determine the kinds of genes that are likely to co-occur in target symbols. Finding such frequent pattern plays an essential role in mining association. The association rule technique has to be applied to gene-expression data. Typical example such rules are: if gene a expresses then there is a good chance that gene b also expresses. This is an unsupervised data-mining technique.

2. Related work

The past decade has been an explosive growth in genomics, and biomedical research. Example range from the identification and comparative analysis of the genomes of human and other species (by discovering sequencing patterns, gene functions, and evolution paths) to the investigation of genetic networks and protein pathways, and the development of new pharmaceuticals and advances in cancer therapies. Biological data mining has become an essential path of a new research field called bioinformatics. We outline only a few interesting topics in this field, with an emphasis on genomic and proteomic data analysis.

The paper is organized as follows: Section 1 describes the Introduction and section 2 presents the related work of this system. Section 3 introduces Theory Background of the system and introduction about DNA mutation and DNA translations .Section 4 shows the system design and how to process data sequences. Section 5 presents experimental results over the DNA sequence dataset and section 7 conclude this paper and shows future works.

3. Theory Background

Many studies have focused on the comparison of one gene to another. However most disease are not triggered by a single gene but by a combination of genes acting together. Association analysis methods can be used to help determine the kinds of genes that are likely to co-occur in target samples. We proposed fp-growth algorithm for mining frequent pattern and mining association rules over sequences.

3.1. FP-growth

An interesting method in this attempt is called frequent-pattern growth, which adopts a divide-and-conquer strategy. First it compresses the database representing frequent items into a frequent-pattern tree which retains the itemset association information. It then divides the compressed database into a set of conditional database (a special kind of projected database), each associated with one frequent item or "pattern fragment," and mined each such database separately.

The first scan of the database is which derived the set of frequent items (1-itemsets) and their support counts (frequencies). An FP tree is constructed follows. First, create the root of the tree, labeled with "null". Scan Database D a second time. In general, when considering the branch to be added for a transaction the count of each node allowed a common prefix is incremented by 1, and nodes for the items following prefix are created and linked accordingly.

To facilitate tree traversal, an item header table is built so that each item points to its occurrence in the tree via a change of node-links. The tree obtained after scanning all of the transaction with the associated node links. The tree obtained after scanning all of the transaction with the associated node link. In this way, the problem of mining frequent pattern in database is transformed to that of mining the FP-tree.

The FP-tree is mined as follow. Start from each frequent length-1 pattern (as an initial suffix pattern), construct its conditional pattern base(a "sub database," which consists of the set of prefix paths in the FP-tree co-occurring with the suffix pattern), then construct its(conditional) FP-tree, and perform mining recursively by the concatenation of the suffix pattern with the frequent patterns generated form conditional FP-tree.

3.2. Mining sequential patterns

The information age has caused an explosive growth in the amount of data produced. Mining for knowledge from this data has been shown useful for many purposes ranging from finance, marketing, and bio-informatics. Sequential pattern mining is an important data mining method with broad applications that can extract frequent sequences while maintaining their orders. It will extract pattern that appear more frequently than a user-specified minimum support while maintaining their item occurrence order. Addressing the same issue of knowledge discovery from data, we study the problem of mining recurrent rule, having the following form:

"Whenever a series of precedent events occurs, eventually another series of consequent events occurs"

3.3. Generating association rules

Once the frequent itemsets from transactions in a Database D have been found, it is straightforward to generate strong association rules from them (where strong association rules satisfy both minimum support and minimum confidence).

Confidence (A ⇒B)=P (A\B) = support_count (A U B)/support_count (A) (3.1)

Association rule can be generated as follows:

- For each itemsets L, generate all nonempty subsets of L.
- For every nonempty subsets s of l, output the rule

"S ⇒ (*l* - *s*) " if support count (l) / support count(s) ≥ min_conf, where min_conf is the minimum confidence threshold.

The strong association rules over DNA sequences are generated by using equation (3.1). Before generating association rules all frequent patterns and their support counts are computed by FP-growth algorithm.

3.4. Mining association rules on sequential Data

We are interested in dealing directly with sequences. The total orders can be seen as sequences, and then an association rule is a fragment of order implying another fragment of order, so that the antecedent is a subsequence of the consequent. The association rule mining depends on the calculation of generators for each closed set of sequences.

Table 1.Inuput DNA sequences

SeqID	DNA Sequences
D1	gatcctccat atacaacggt
D2	ccgacatgag cagtta
D3	ctgcatctga gccgctgaa

4. DNA mutation

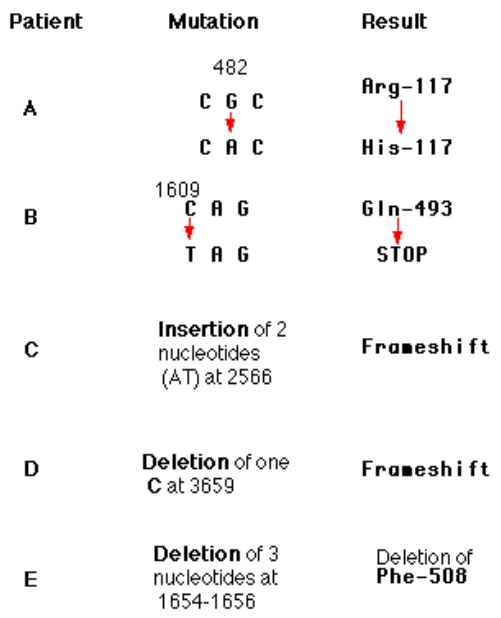

Figure 4.1 Mutations in DNA Sequence

Mutation is a permanent change in the DNA sequence of a gene. Mutation in a gene' DNA sequence can alter the amino acid sequence of the protein encoded by the gene. The DNA sequence is interpreted in groups of three nucleotides bases, called codons. We can think about the DNA sequence of a gene as a sentence made up entirely of three-letters words. For example: thesunwashotbuttheoldmandidnotgethishat then split this sentence into three letters words. The sun was hot but the old man did not get his hat. This sentence represents a gene. Each letter corresponds to nucleotides base, each word represents a codon. We can mutate the reading frame of this sentence by inserting or deleting letters within the sentence. Sample DNA sequence: AAA, AAC, ACC, ACA, CAA, CCA, CAC, and CCC are translated into corresponding Amino acids asparagines, glutamine, histidine, lysine, proline, and threonine. For example if a DNA nucleotides sequence is changed from ATCGACTAGCT to ATCGACTTGCT? Changing an A to T in your sequence may or may not alter the amino acid produced, depending on the reading frame. Figure 4.1 shows the DNA mutation and changes of amino acid. These mutation records are going to be the input of the system.

Mutations are often more harmful because they tend to change the amino acid produced according to the positions. We describe an approach to detect the presence of abnormal alleles in those genetic diseases in which frequency of occurrence of the same mutation is high. And in others in which multiple mutations cause the disease and the sequence variation in an affected member of a given family is known.

4.1. DNA Translation

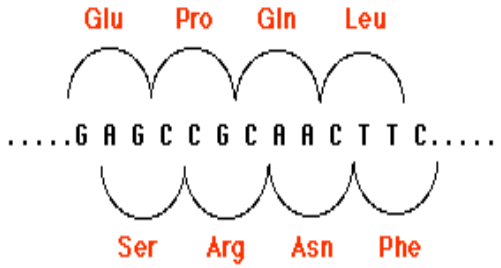

Figure 4.2 DNA Translations to Protein

Translation is the process in which the RNA blueprint is read and the information used to build a protein molecule. Each group of three nucleotides in our mRNA blueprint is called a codon and encodes for one amino acid. In other words, our cells read the nucleic acid triplet code and build proteins based on the 3-nucleotide 'words'. Figure 4.2 shows the DNA translation into proteins.

5. System Architecture

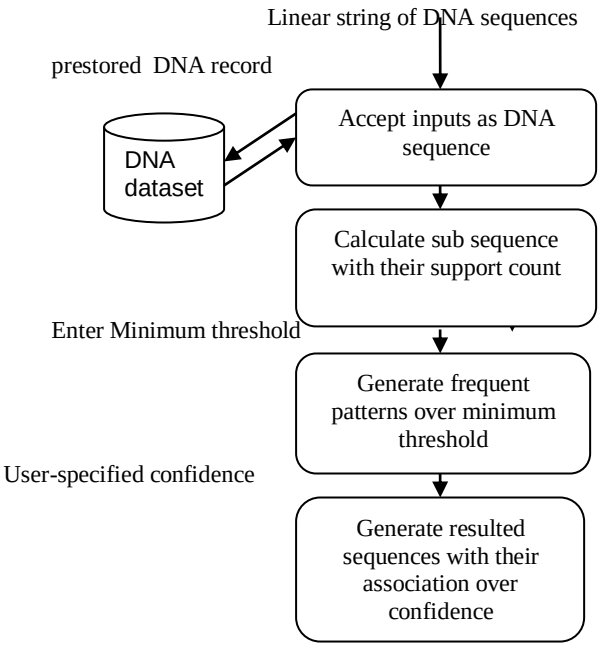

Figure 5.1. System Architecture

The system stores the user-input DNA sequences into the database and process previously stored DNA datasets as shown in table 1. Then it accepts

minimum supports and generates frequent itemsets whose support_count is over user-specified minimum support. After generating frequent itemsets, the system accepts the minimum confidence to extract the interesting patterns over user-specified minimum confidence. Finally we would have a rule such that <aca>→<cca>^ <ctg> , where the presence of fragment of <aca> predicts the presence of fragment of all the order in the consequent. The association rule is a fragment of order implying another fragment of order, so that antecedent is a subsequence of consequent. The presence of the sequence may suggest that the presence of the others to be expected. The resulted sequences must satisfy both minimum support and minimum confidence in order to generate strong association rules over these sequences.

5. Experimental Result

The system is implemented in C#.net and all experiments were found on Window Server2003. The Influenza A virus (H1N1virus) segment 4 sequences of 26 sequences file is used to generate frequent pattern and strong association rules over user-specified minimum support and minimum confidence. The system proposed frequent mining algorithm which mines frequent pattern in descending orders of support by specifying minimum support and minimum confidence. The proposed mining algorithm (fp-tree) has significant advantage. It helps improve the efficiency of mining process significantly. We have implemented the Fp-growth method, studied its performance in comparison with several influential frequent pattern mining algorithms in large databases. Our performance study shows that the method mines both short and long patterns efficiently in large database since we have accepted the large record of DNA sequence in various lengths as shown table 1

6. Conclusion and future work

The focus of this thesis is to emphasize FP-tree algorithm working on sequential database. The limitation is that it is unrealistic to construct fp-tree on main memory or very large database. If we can get the DNA sequence dataset for a particular genetic disease, we can extract the interesting association rules from these sequences and predict which kind of genetic disease can occur. We can extract interesting association rules over DNA sequences in specific disease and determine the kinds of genes likely to co-occur frequently .So that we can predict which kind of genetic disorder can occur if we can determine the strongly associated DNA sequences. Since the system explores no of possible frequent patterns, we can see most frequently patterns likely to occur in genetic disorder DNA sequences dataset.

7. References

[1].Jian Pei , "Pattern-Growth Methods For Frequent Pattern Mining",Simon Fraser University,June 13 2002

[2] Yu Hirate,Eigo IWAHASHI,"TF^2 P-growth: An Efficient Algorithm for Mining Frequent Patterns without any Thresholds", Waseda University

[3]Hengshan Wang,"Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan",Business School ,University of Shanghai for Science and Technology

[4]David Lo, Siau-Cheng Khoo, and Chao Liu ,"Efficient Mining of Recurrent Rules from a Sequence Database", Department of Computer Science, National University of Singapore

[5]Yu Hirate,"Generalized Sequential Pattern Mining with Item Intervals", Media Network Center ,Tokyo Japan

[6] C.I. Ezeife ,Yi Lu, Yi Liu,"PLWAP Sequential Mining : Open Source Code " ,Department of Computer Science ,Wayne State University Detroit, Michigan

[7] Jiawei Han and Micheline Kamber,"Data Mining Concept and Technique"