A Study on Nucleotide Cancer Liver Cells with DNA Binding Using Hidden Markov Model

M. Mayilvaganan  
Associate Professor: Department of Computer Science  
PSG College of Arts and Science  
Coimbatore, Tamil Nadu, India

R. Rajamani  
Assistant Professor: Department of Computer Science  
PSG College of Arts and Science  
Coimbatore, Tamil Nadu, India

Abstract— The work will be focused and analyzed about performance of DNA gene liver cancer database and normal liver cell data set from ncbi DNA data set. Each amino acid can have character variables and also assigned numeric number and its corresponding pair combination of sequence are represented in a graph. The nucleotide sequence of biological databases is growing long terms of quantity, memory and complexity, managing these databases is becoming very complex. In this paper focuses Hidden Markov Model (HMM), has increased on the Pattern recognition domain primarily because of its strong mathematical basis and the ability to adapt to unknown of nucleotide sequence of normal and cancer affected liver cells as are pictorially represented by finite state machine. It is a finite automaton with a fixed number of states which are trained to maximize the probability of the observation sequence by using viterbi algorithm and forward algorithm. The proposed HMM system is validated with two different nucleotide values for analyse the performance and get the simulated output using viterbi and forward algorithms implemented in Mat Lab Tool.

Keywords— Hidden Markov Model; Viterbi algorithms; Forward algorithms; Pub Chem of liver and Cancer DNA dataset;

I. INTRODUCTION

Clustering can be one of the indispensable unsubstantiated learning methods. Every other section of this kind, it overcomes with formative a structure in an concatenation of unlabeled data. A cluster is a widespread collection of objects that are analogous to one another with in the same cluster and are different to the objects in other clusters. A cluster of data objects can be aggregated treated as one group and so may be measured as a form of data compression.

Although many sophisticated methods for detecting regulatory interactions cluster analysis remains a useful routine in array analysis.

Hidden Markov model (HMM) is a finite-state machine corresponds with the doubly stochastic process include at minimum pair of levels of uncertainty: a random observation process corresponds with each hidden-state, and a Markov chain, which analyze the no of occurrence relations among the layers in which how likely one state is to follow another.

II. DATA FOR RESEARCH

A. Overview of Original Genetic Code Data

The nucleotide research data set used for proposed methodology in DNA sequences which are taken from Pubchem. The information of sequence DNA data source are relevant to chemical and bioactivity manner. The nucleotide sequence in DNA is stored as a code made up of A,G,C,T chemical codes. A-Adenine, G-Guanine, C-Cytosine, T-Thymine. Human DNA consists of three billion bases and more than 99 percent are same in all people. This paper contains the descriptions of Homo sapiens (human). This database contains 40000 gene sequences. The following data shows the normal DNA nucleotide data set sequence.

`aaaagttgcgg cgagacagct ggctaatgct tgtaactccc gactttcgg aggcgtgagtt caggagacttg agacatctct gcgtaactgg ttttccaaaa aatatataaa
acacacgttc gcgccggtg cctgtaggcc cagctactcg gggagctg aagcgttgag atggctgaccc cccctttca a g`
**B. Liver Cancer DNA Nucleotide Data Set**

The cancer affected nucleotide research data set used for proposed methodology in DNA sequences which are taken from Pubchem. The nucleotide sequence in DNA is stored as a code made up of A,G,C,T chemical codes. A-Adenine, G-Guanine, C-Cytosine, T-Thymine. This database contains more than 40000 gene sequences. The following data shows the normal cancer affected liver DNA nucleotide data set sequence.

2281 ggattgtcag agaacaagtc ctatcgtgg cegggtgga aacagaggagggagagc
2341 gcctgcgcg gggcatcag gattggctta aacaaccttc cctggggtgagc
2401 tgaggaag tggctaggt gaaactctaat gaaactctaat gaaactctaat
2461 tttaggaag tgaataagct gtgggaactc gggccctgtc cttggggtgagc
2521 ccgctggtg tggagaacta tggagaacta tggagaacta tggagaacta
2581 cgggtggtg tggagaacta tggagaacta tggagaacta tggagaacta
2641 aataaatcgg cgtctttatg tgcacactgt tgcacactgt tgcacactgt

gctgccacactc cctccgaagcc ecatcttgtc gcactggtg cactcttgtc

cctcatttg cttgcctgg cgtgcatggct tggcaacctt cttggaaatcc

ttgagggca gcatctgtgt cttgtctaac ttggtatccc
gagaaatcag tggatttctt
tgtagagt acttcatgcc atgtactttg ttccccttta
tgtaatttt tgccccttta ttcatctctc
tggccaactc cctccaagcc ccatcgtcttt
tcttcctatt gcctcgactc
tttgtaacctc ttgagctcag

gtcgtagat acttcatgcc atgtactttg ttccccttta ttcatctctc
tgggaatata cttttgcttc
tggggctctt gagcagcttg ctttagcctg
tgagctcag tgcctggagc cgctccctca
tggggat
tagcaactcttc cccc

gctctggagc cgctccctca
tggggctctt gagcagcttg ctttagcctg
tgagctcag tgcctggagc cgctccctca
tggggat
tagcaactcttc cccc

gctctggagc cgctccctca
tggggctctt gagcagcttg ctttagcctg
tgagctcag tgcctggagc cgctccctca
tggggat
tagcaactcttc cccc

gctctggagc cgctccctca
tggggctctt gagcagcttg ctttagcctg
tgagctcag tgcctggagc cgctccctca
tggggat
tagcaactcttc cccc

gctctggagc cgctccctca
tggggctctt gagcagcttg ctttagcctg
tgagctcag tgcctggagc cgctccctca
tggggat
tagcaactcttc cccc

gctctggagc cgctccctca
tggggctctt gagcagcttg ctttagcctg
tgagctcag tgcctggagc cgctccctca
tggggat
tagcaactcttc cccc

**III. EXISTING SYSTEM**

Association rule mining is one of the traditional data mining methodology, which finds associated item sets from a large number of data set occurrence. Apriori recognize the patterns with frequency above the smallest amount value as threshold and establish rules that express occurrence relationships between nodes in frequent item sets [2]. It is used for data diminution or pre-processing to diminish the amount of the attribute to be discussed. The output is to make strong association rule with respect to the data which is used for analyzing the data compression. The data pre-processing in FSA-Red processed with a reduction techniques such as attribute selection, row selection and feature selection. Feature selection will erase all the unwanted attribute, closed with attribute selection to reduce the non value attributes which is no need to be measured.

**IV. PROPOSED SYSTEM**

The following fig.1 illustrates the proposed system architecture for two nucleotide sequence of normal liver cells and cancer affected liver cells. In the following architecture, data analysis is the first phase in which data are analysed with nucleotide sequence. After analyzing the data, the HMM process constructs the finite state machine model for two data sets. Using the viterbi algorithm and forward algorithm, the score or probability can be estimated for all possible alignment of amino acids in a nucleotide sequence. Using the simulated output, the performance analysis of viterbi and forward algorithm are analysed.
B={a_{ij}} = P(A_{x_i}/A_{x_{i-1}} = i)  \hspace{1cm} (1)

A. Viterbi Algorithm and Forward Algorithm

It is an energetic programming procedure that can calculate approximately the best positions and its probability without going through all possible combinations. This algorithm is used to produce numerous alignments of a pair of sequence. All the amino acids are coordinated to a certain match state are placed each other to form the same position alignments. If any series do not have equal to any state, a break character is added.

Each amino acids sequence of nodes starts with begins state (G) and an end state (E). Each amino acids in an HMM has a match state (A), insert state (N) and delete state (D) with position specific probabilities for transitions into each states from previous node. Forward algorithm is used to calculate the aggregate over all paths individually.

For modeling the amino acid sequence, the following steps are produced. HMM can be visualized as Finite state machine.

1. Collect the set of sequence of amino acids
2. Define a grammar for sequence set G={x_1, x_2, x_n).
3. Develop a model, to generate typical sequence from the class of trained data sequence.

Finite state machine pass through a set of states and produce some output whether the machine has reached a goal state or when machine transfer from one state to another state. A gene, and the outputs will be DNA bases. The states will correspond to different regions within or near to genes.

In HMM sequence of observations x = x_1, x_2, ……x_T and an HMM λ and to find the best sequence of states S=S1,S2, ST to explain x, given λ. There are different ways in order to determine the best state set S maximizing the total likelihood of the outputs and states, given the model such as

Max Pr(x, S / λ)

The rules of the conditional probability such as

Pr(x, S / λ) = Pr (x / S, λ) Pr[S / λ]

Each element of the DNA gene, Mij, will record the best possible sequence of states for the first j observations such that the jth state is q_i. That is

Mij = max (si---sj) = Pr[x_1, …x_j ^ S1,….Sj | λ ^ Sj = q_i)

Algorithms for Hidden Markov Models

1. Given an experimental set of output x and an HMM λ = (Q, P, Σ, π, B), find the affordable state string S to produce x from λ.

2. Determine the probability of generating x from λ.

3. Determine the parameters of (P, B and π) that maximize the probability of producing x from λ.

V. SIMULATION RESULT AND DISCUSSION

Using the HMM model, the amino acid pair sequence and unpaired sequences are formed using finite state machines (FSM). FSM model generates the goal state by passing the set of amino acid node to pair of nodes and process various constraints such as append state, delete state and match state. If match is found the sequences are paired otherwise gap is formed between each pair of amino acid. It is represented in the fig.3.

![Fig. 3 Pair combination of sequence](image)

FSM model generates the output by passing the set of amino acid pair of node and process various constraints such as append state, delete state and match state. If the match is not found, gap is formed between each sequence of amino acid. It is represented in the fig.4.

![Fig. 4 Simulation result and discussion](image)
The following Fig 4 shows the performance precision of two nucleotide cancer dataset. The Fig 6 represents the performance of HMM algorithm achieves high performance with respect to memory, time and speed as 93.3% with compared to normal and cancer affected nucleotide sequence.

Fig. 5 DNA Binding combination with distance matrix

VI. CONCLUSION

The proposed tool that extracts the from gene data files by means of a diversity of selectable algorithms and criteria. The program integrate numerous taking out methods which allow the efficient extraction of rules, while allowing the care of the mine to be specific at the users discretion. The program also allows the results to be displayed through a range of graphical representations. Such representations can often help to review the facts being analyzed by given that a concise conceptualization of the data under scrutiny. This
paper uses viterbi algorithm and forward algorithms and uses other algorithms to improve this approach. This was applied in biological application in DNA data set; future work can be carried out in other biological nucleotide industry.

REFERENCES


