Estimation of Distance complexity in amino acids between Normal and Cancer Liver Cells using Data Mining Techniques

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 Data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Clustering algorithm used to find groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups . This paper comprises of two database such as normal liver cells and cancer affected cells. After analyzing the cancer cells, there is a need to determine the distance between normal and cancer affected cells. 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 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. The extracted rules and analyzed results are graphically demonstrated. The performance is analyzed based on the different no of instances and confidence in DNA sequence data set.

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

I. Introduction

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. 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. The probability of the each amino acid sequence is found by aggregating the probabilities in the last column. In the context of HMM, the forward algorithm is used to calculate a ‘belief state’ the probability of a state at a fixed time, given the history of evidence.

It is understood that there is a hidden process that generates a sequence of amino acids residues, where chance plays an important role in determining the exact sequence being produced. 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.

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.

aaaaagtcggc cggacacagt ggctcatgcc tgtaatccca gcacttcggg aggctgaagt

ccaggagactg agaccatcct ggctaacatg gtgaaacccc atctctacta aaaatacaaa

aacaaaatta gccaggcatg gtggcgggcg cctgtagtcc cagctactcg ggaagctgag

gcaggaggaat ggcgtgaacc cgggaggcgg agcttgcagt gagctgagat tgcaccactg

ctactccagcc ttggcgacag agtgagactc cgtctcaaaa aaaaaaaaaa aaaaaagtcc agcgagcatct cgaaacaaac aggctcaact ccaaatcctt cactgtccac taacaagtac

actccatgtct tgctggatgg gagcacatgt agctccacaa tacttttggt cacacaggcc

tattgaatct taacttcctt attcatcccc tctcactatt ctcacctctg tggacttaat

tcagttccct gtttctcctg tggatcactg cattaggctc cttaccattt tcttcctgcc

attaactttg cccccttca agtcaccctt cactgagttt cttcactatc tttccaaaaa

g tgtaaatctt agcacaacag gctgcagctt aaagtccttt agtgactccc cgtagctcag

taggatgaggt tctcatttcg gagtatttac agttcttgtc tatctctgtg gcctcgactc

cgtccccactct cctccaagcc ccatttcctt gactgggcag cactccttgt tcttcctatt

ccttatgctg tttcctgcct ctagccccgt gcgtttgtac ttcccactgc tggaacattc

agttctctcctt tccctttccc cgctcctgat ccttcagagt ctaataccca cctctctggg

aggccacatg agctcactgg acaggtgctc ctctgtgtgc aaacatcact gtgcatggct

gctgttagagt acttcatgcc atgtaatttt tgcccctta ttcatctctc ccctcatttg

tctggaaatcc tgtgagggca gcatctgtgt cttgtctaac ttggtatccc tgacacctaa

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.

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. After estimating the score of each amino acids, there is a need to determine the nearest distance between each amino acids with respect to normal and cancer cells. Using the simulated output, the performance analysis of viterbi and forward algorithm are analysed.

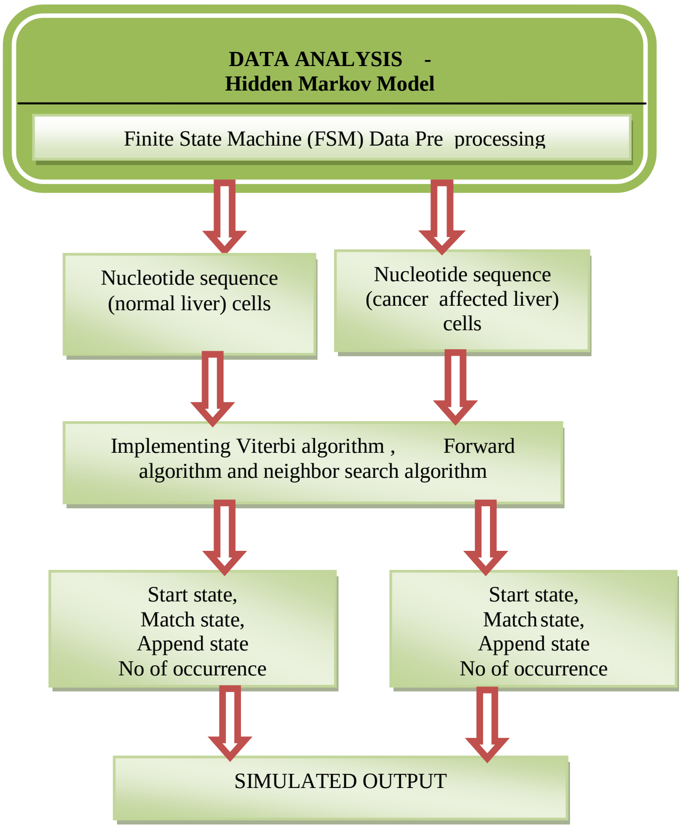
DATA ANALYSIS -
Hidden Markov Model
Finite State Machine (FSM) Data Pre processing
Nucleotide sequence (normal liver) cells
Nucleotide sequence (cancer affected liver) cells
Implementing Viterbi algorithm , Forward algorithm and neighbor search algorithm
Start state, Match state, Append state No of occurrence
Start state, Match state, Append state No of occurrence
SIMULATED OUTPUT

Fig.1 Process Flow Diagram

Fig.1 shows the process flow of the proposed methodology. The efficiency of proposed work is considered with the accessible techniques. The Count and position of gene sequences are retrieved using vetribi algorithm. The HMM representation states the gather probability of a concealed data variable and pragmatic discrete random data. It assumes the hidden data xi th in the (xn-1) th variable is not depend to previous variables and the current observed data depend only on current hidden data.

Let A_x be a discrete hidden data with Z possible values. We assume the equation (1), P(A_x/A_x-1) is independent of time m, which leads to the definition of the time independent stochastic matrix.

$$B=\{a_{ij}\} = P(A_{xi}/A_{xn-1} = i) \quad (1)$$

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.

Neighbor Search Algorithm

Nearest neighbor search (**NNS**), also known as proximity search, similarity search or closest point search, is an optimization problem for finding closest (or most similar) points. Closeness is typically expressed in terms of a dissimilarity function: the less similar the objects, the larger the function values. Formally, the nearest-neighbor (NN) search problem is defined as follows: given a set *S* of points in a space *M* and a query point *q* ∈ *M*, find the closest point in *S* to *q*. Donald Knuth in vol. 3 of *The Art of Computer Programming* (1973) called it the post-office problem, referring to an application of assigning to a residence the nearest post office. A direct generalization of this problem is a *k*-NN search, to find the *k* closest points.

```
1: {find an Approximate Nearest Neighbor to query point q}
2: r = 0 {radius of ball which has been completely explored}
3: δ = ∞ {distance to nearest point seen so far}
4: enqueue (bounding box) {queue is ordered by distance of nearest point in box to q}
5: while δ ≥ r do
6: dequeue box B, containing representative point p
7: r = d(q, B)
8: if d(q, p) < δ then
9: p becomes best choice seen so far and δ = d(q, p)
10: end if
11: for all children B0 of B containing points in P do   12: enqueue (B0 )
13: end for 14: end while 15: return p
```

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 and their distance are estimated otherwise gap is formed between each pair of amino acid. It is represented in the fig.2.

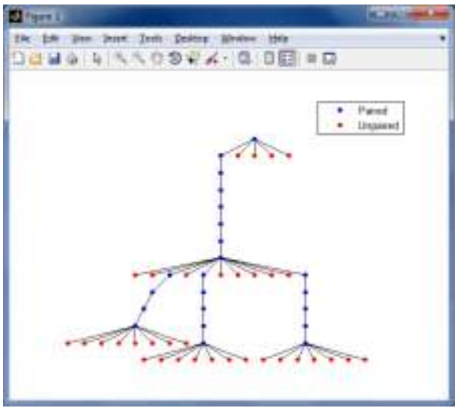

Fig. 2 Pair combination of distance sequence:

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.

The following fig.3 shows the perfromance precision of two nuclетide cancer dataset. The following diagram 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..

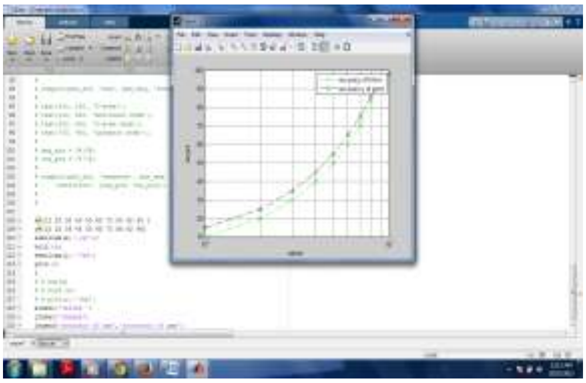

Fig.3 Accuracy Calculation of HMM algorithm

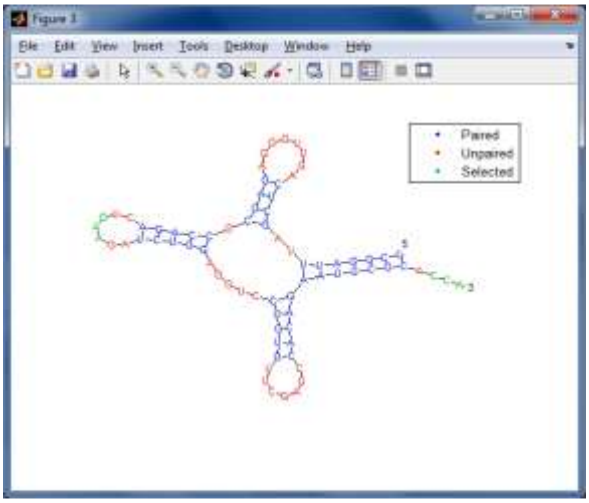

Fig 4. distance estimation in cancer cells

The following Graph (4), and (5) shows the perfromance estimation of HMM algorithm with two different nucleotide DNA data set. In the graph red colour represents the performance of algorithms with respect to time 0.946msec and blue represents the memory 3GB with compared to normal nucleotide liver DNA liver cell and cancer affected liver cell in DNA

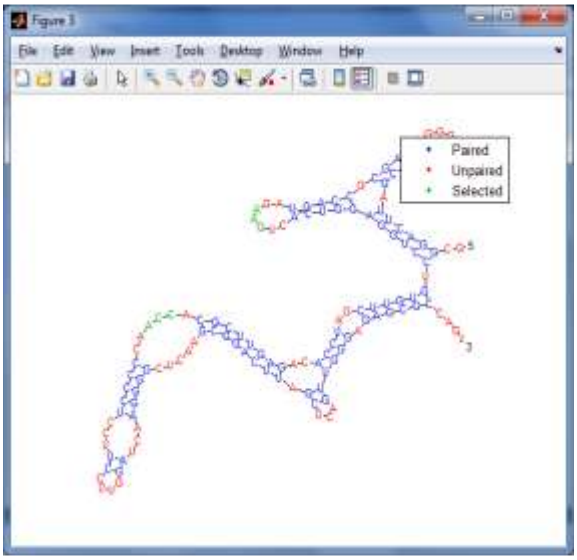

Fig 5 distance estimation of selected sequence between amino acids

VI. Conclusion

The proposed tool that extracts the from gene data files using a variety of selectable algorithms and criteria. The program integrates several mining methods which allow the efficient extraction of rules, while allowing the thoroughness of the mine to be specified at the users discretion. The program also allows the results to be displayed through various graphical representations. Using the nearest neighbour distance algorithm, distance are estimated between cancer and normal cells. Using the clustering sequence alignment technique in data mining, the pair of sequence alignment within nucleotide, character search alignment and combination pair of cells are analysed with running time and memory efficiency. In future, the research work will be extended into following direction. Using the Hidden Markovo Model, the finite state machine model will be generated and analysed for nucleotide sequence for cancer affected liver cells and normal liver cells .

REFERENCES

[1] M.Anandavalli , M.K.Ghose , K.Gouthaman ,"Association Rule Mining in Genomics",International journal of computer Theory and engineering, Vol.2,No.2 April,2010.

[2] Bayardo, Roberto J., Jr.; Agrawal, Rakesh; Gunopulos, Dimitrios (2000). Constraint-based rule mining in large, dense databases Data Mining and Knowledge Discovery (2): 217–240.

[3] Donald, "Introduction to Data Mining for Medical Informatics," Clin Lab Med, pp. 9-35, 2008.

[4] JunoWatada,KeisukeAoki, Masahiro Kawano, Muhammad SuzuriHitam, Dual Scaling Approach to Data M Journal of Advanced Computational Intelligence Intelligent Informatics , Vol. 10, No. 4, pp. 441-447, 2006.

[5] Jiawei Han and MichelineKamber,"Data Mining Concepts and Techniques." San Francisco, CA: Elsevier Inc, 2006.

[6] Irene M. Mullins et al., "Data mining and clinical data repositories: Insights from a667,000 patient data set," Computers in Biology and Medicine, vol. 36, pp. 1351-1377, 2006.

[7] Liao.S & M. Embrechts I. -N. Lee, "Data mining techniques applied to medical information," Med. Inform , pp. 81-102, 2000.

[8] Piatetsky-Shapiro, G.& myth P. &Uthurusamy, R. Fayyad, "From Data Mining toKnowledge Discovery: An Overview," in Advances in Knowledge Discovery and DataMining, 1996.

[9] Webb, Geoffrey I. "Efficient Search for Association Rules", Proceedings of the Sixth ACM SIGKDD International Conference Knowledge Discoveryand Data Mining (KDD-2000), Boston, MA, New York.

[10] R. Zhang, Y, Katta, "Medical Data Mining,"Data Mining and Knowledge Discovery, pp. 305-308, 2002.