International Journal of ChemTech Research
CODEN (USA): IJCRGG, ISSN: 0974-4290, ISSN(Online):2455-9555
Vol.12 No.03, pp 227-232, 2019

Introduction to Hidden Markov Model and its Biological Applications

S Narendra Kumar[1]*, Lingayya Hiremath[2], Ajeet Kumar Srivastava[3], Praveen Kumar Gupta[4], Jyothsana R[5], Rithika Pravin Iyer[6], Ruchika Pravin Iyer[7]

[1, 2, 3,4]Assistant Professor, Department of Biotechnology, RV College of Engineering, Bengaluru, India
[5,6,7] B.E students, Department of Biotechnology, RV College of Engineering, Bengaluru, India

Abstract : Hidden Markov Model (HMM) is a stochastic model where all the states are hidden and emit outputs that are observable. HMM is intertwined with artificial intelligence that allows it to be applied to latest technologies like speech and handwriting recognition. It has proved to be an extremely valuable tool in the field of bioinformatics and has been extensively used in sequencing, alignment, homology prediction and protein secondary structure prediction. In order to understand HMM, we have used a simple problem based on land used for pigeon pea cultivation throughout years which has been illustrated using MATLAB and studied using matrix multiplication and probability concepts. This paper also outlines the technique used for studying gene prediction using HMM. It heavily draws content from the forward and backward recursions which were given by Ruslan L Stratonovichlt in 1960. It is a nexus of complicated ideas and calculations, including various algorithms like Viterbi and Forward algorithm which can be implemented using various programming languages such as C, C++ and tools like Matlab. Even though the concept is about 60 years old,it is one of the most detailed algorithms available to date. Due to ever increasing applications, it has become a constant in the curriculum for students of different backgrounds like computer science, electronics and bioinformatics.
Keywords : HMM, Markov chain, Transition diagram, Forward algorithm, Viterbi algorithm, Gene prediction.

Hidden Markov Model:

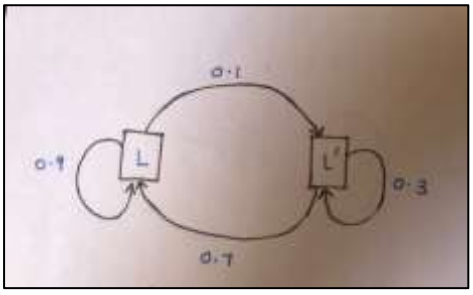
Hidden Markov Models has proven to be very useful as it can be used to predict outcomes for a set of ordered observations that possess a hidden structure. The problem for predicting genes is frequently solved using the concept of Viterbi algorithm[3,4,8,12]. The various states are exon, intron and the splicing region. Using the transition diagram we can easily understand this procedure where the algorithm is used to get the paths and

DOI= http://dx.doi.org/10.20902/IJCTR.2019.120329

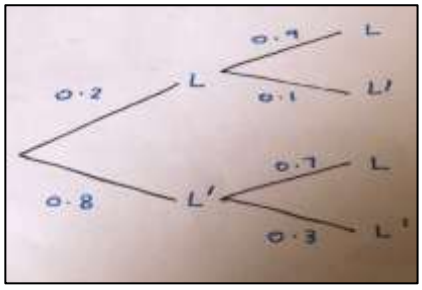
the Viterbi probability[5,6]. HMM has a strong statistical foundation and includes efficient learning algorithms that enable us to solve issues related to bioinformatics. Its advantages include: a.) Efficient learning algorithms – learning can take place directly from raw sequence data. b.) Modularity- It can be combined into larger HMM. It can handle inputs of variable length. It can be combined into libraries. c.) Transparency- the model itself can help increase understanding[14].Hidden Markov Model (HMM) is a statistical model in which the system is assumed to be a Markov process with unobserved (i.e. hidden) states[2]. It can be represented as the simplest dynamic Bayesian network.[1] Where the states are visible, the state transitional probabilities are the only parameters[7]. While, in the Hidden Markov model, the states are not visible, but the output is.The word "Hidden" refers to the states of the model and not its parameters. Even if, the parameters are known the model still comes to be known as hidden Markov Model.

Implementation of HMM Algorithm Using Matlab:

Problem: The land under pigeon pea (toor dal) production is 20%. The government wants to increase this and decides to provide incentives to those producing pigeon pea. After some research, it was concluded that someone who currently cultivates pigeon peawill continue to do so in the future,with a probability of 90%. Someone who does not cultivate pigeon pea will switch to producing some other crop with a probability of 70%.Q1)What is the percentage of land under pigeon pea cultivation after 1year (take 5-11 months)?Q2)What is the land under cultivation after 3 years? Q3) Will all the land be under pigeon pea production?Q4) Does every Markov chain have a unique stationary matrix?Q5) If a Markov chain has a unique stationary matrix will the successive state matrices always approach this stationary matrix?



Transition State



Tree Diagram

Let, the land under cultivation of pigeon peas be L and the land under another crop cultivation be L´. The time taken to grow the crop is 1 year. From the transition state diagram, we can conclude that we have two states. L means we are cultivating pigeon peas and the other L´ were we are not cultivating pigeon peas.The probability of continuing to cultivate pigeon peas after the incentive goes live is 0.9.The probability of switching to pigeon pea cultivation after the incentive goes live is 0.7.By using total probability concept, we can conclude that the probability of switching to other crop cultivation is (1-0.9)* 0.1 and the probability of continuing to produce crops other than pigeon peas are (1-0.7) *0.3.The transition diagram denotes the probability of staying and switching from one state to another.Using the rule of multiplication of probability, we can find out the land under pigeon pea cultivation:(0.2×0.9) +(0.8×0.7) = 0.74 or 74% .The land not under pigeon pea cultivation can be found out in a similar fashion as:(0.2×0.1) +(0.8×0.3) = 0.26 or 26%. But this is a tedious task and hence we can use matrix multiplication for the same effect.Let P be the transition matrix. It contains the probabilities of switching from one to another state or continuing in it.

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{bmatrix}$$

Let S0 be the initial state distribution matrix that denotes the initial land cultivation for pigeon peas.

S0 = [0.2 0.8].

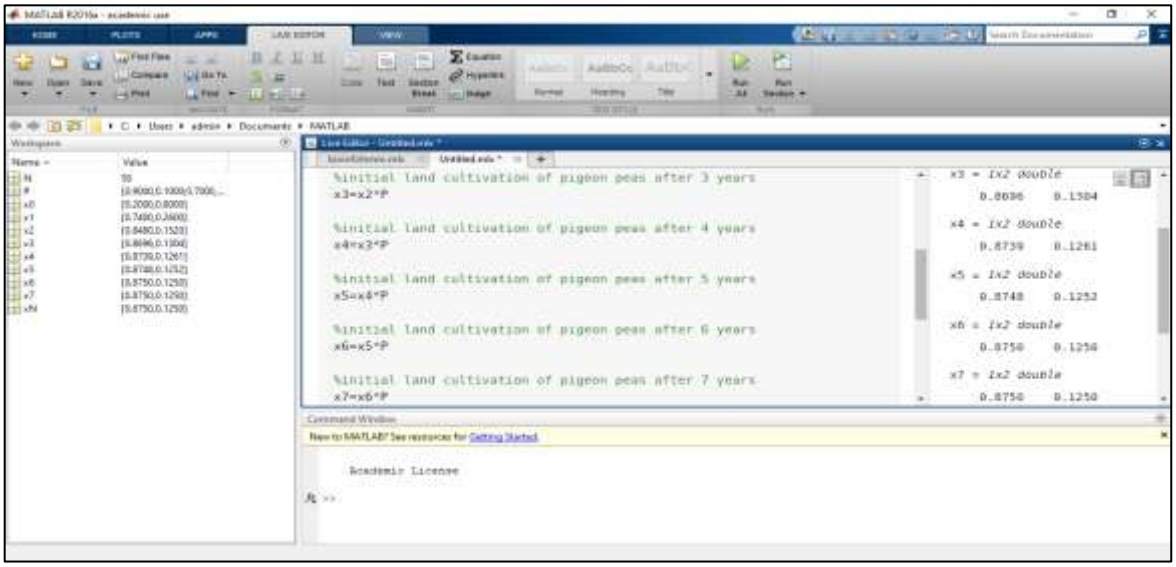
P×S0 = S1 the land under cultivation of pigeon peas after 1 year:

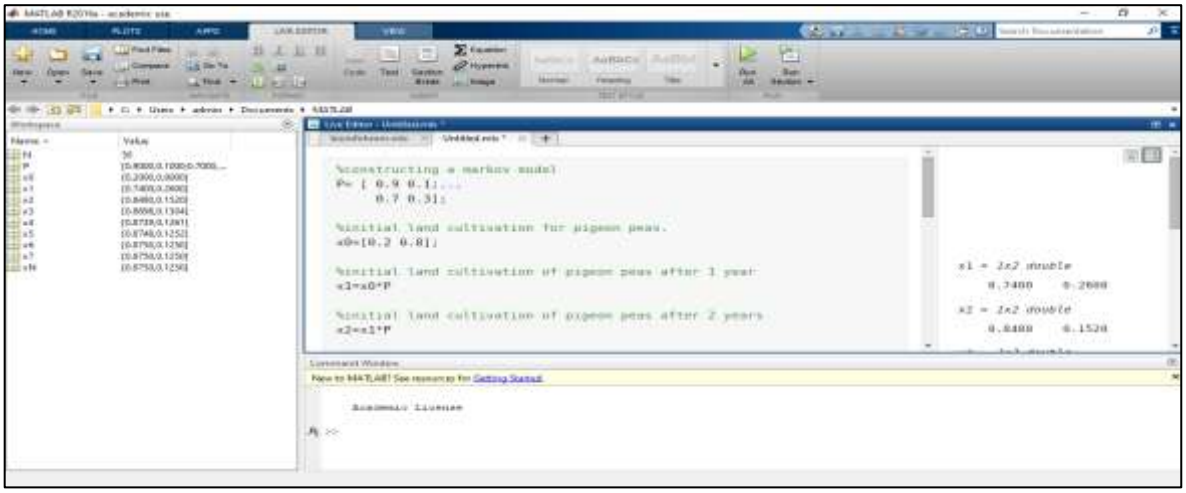
$$S1 = [0.2\ 0.8]\begin{bmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{bmatrix} = [0.74 \quad 0.26]$$

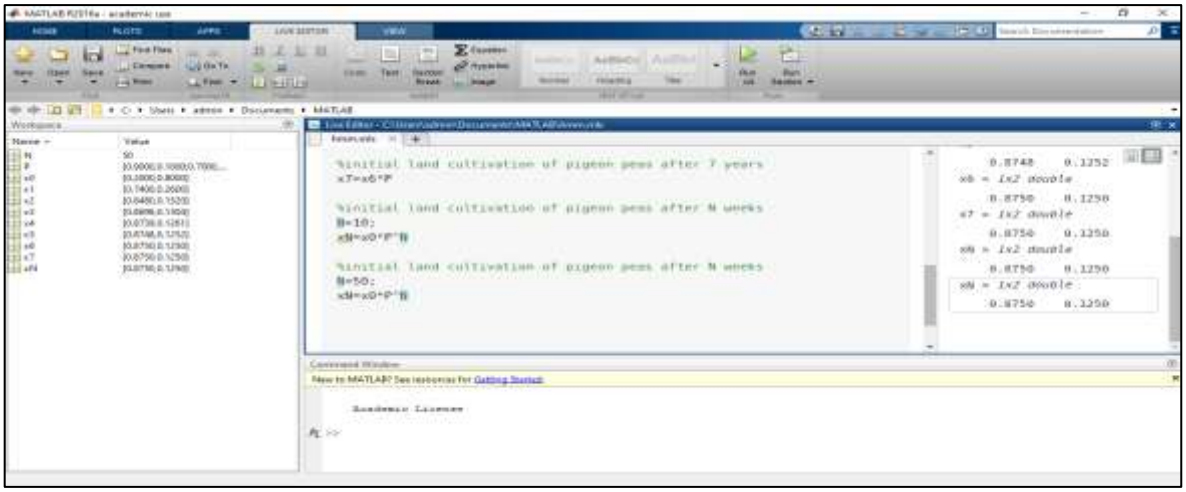
Assuming that if the probabilities in P remain valid over a long time, what happens to the initial state matrix (land for cultivation)? The lands under cultivation in the following years are: S2=S1*P (YEAR 2),=S2= [0.8480 0.1520], S3=S2*P (YEAR 3)= S3= [0.8696 0.1304], S4=S3*P (YEAR 4)= S4= [0.8739

0.1261],S5=S4*P (YEAR 5)=S5= [0.8748 0.1252],S6=S5*P (YEAR 6)=S6= [0.8750 0.1250],S7=S6*P (YEAR 7)= S7= [0.8750 0.1250]

This can be done using MATLAB code as follows:





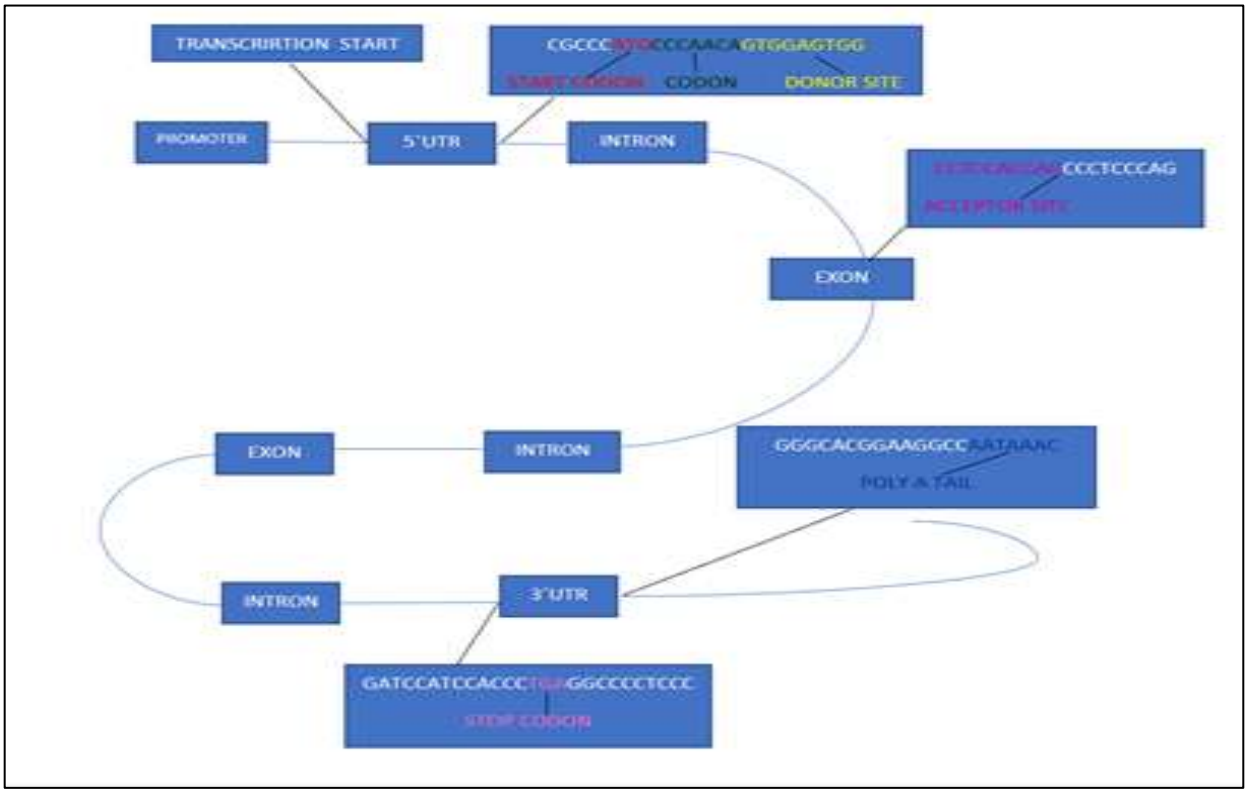


It was observed that after the 7th year, the land cultivation of pigeon peas approached to a standstill. The S7 initial matrix is called a stationary matrix and the system is said to be at steady state. There was no loss

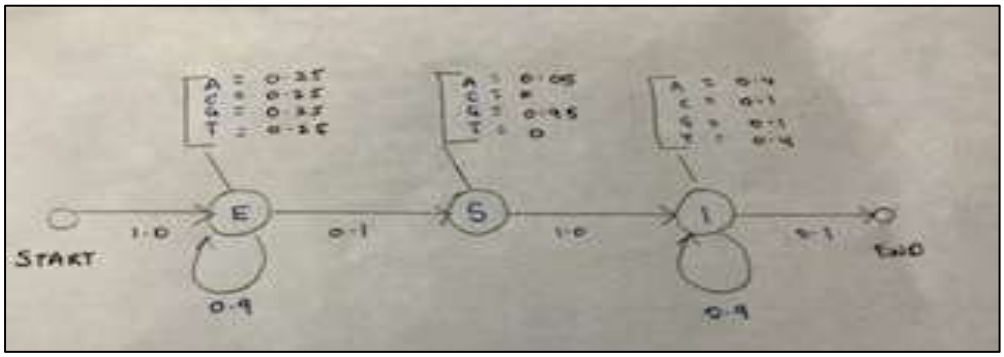
or gain indicating that not all the land (100%) would be under pigeon pea production. Regular Markov Chain was studied to understand the concept of stationary matrix and steady state. Regular Markov Chain: A transition matrix P is regular if some power of P (i.e. P^n) has only positive entries. (A Markov chain is a regular Markov Chain if its transition matrix is regular

Biological Example:

Gene finding refers to the process of identification of protein coding segments of the given DNA. Gene searching algorithms are used to find proteins[8,10,11,13]. This is not an easy task as the structure of the gene has varying degrees ofcomplexity. This process will help in genomic data annotation that obtained from genomesequencing. It also helps gain insight on the process of translation, post translationalmodifications and splicing.In gene finding there are some important biological rules to be considered: 1)Translation starts with a start codon (ATG).2)Translation ends with a stop codon (TAG, TGA, TAA).3) Exon can never follow an exon without an intron in between.4) Complete genes can never end with an intron[14].From the transition diagram, we can conclude the following: The probability of starting is 1.0, when we encounter Exon we notice that p (A) =p(C) =p (G) =p (T) =0.25 i.e. there is equal probability of any of the above bases being present in that position. The probability of continuing inE (exon) state is 0.9 and switching is 0.1. The region for splicing has G in higher concentration thus, p (A) =0.05, p(C) =P (T) =0 and p (G) =0.95. The probability of moving on is 1.0. I(intron) state has p(A)=p(T)=0.4 and p(C)=p(G)0.1. The probability of continuing is 0.9 and switching is 0.1[9,14]. The states are = {start, exon, splicing, intron}. Can be written as {S, E, D, I}. Observations are= {A, C, G, T}.



The Diagram of a Gene and the Transition State Diagram



The transition probability P= { Start: {S=0, E=1, D=0, I=0}
Exon: {S=0, E=0.9, D=0.1, I=0}
Splicing: {S=0, E=0, D=0, I=1}
Intron: {S=0, E=0, D=0, I=0.9/0.1} }

This indicates the probability of each step. When it starts the p(event) =1 because here is no way to go back but at the exon state it can come back i.e. 0.9 or move on. At the splicing stage one has to move on. At the intron stage one can again come back or move on.

The emission probability Q = { Start: {A=0, C=0, G=0, T=0}
Exon: {A=0.25, C=0.25, G=0.25T=0.25}
Splicing: {A=0.05, C=0, G=0.95, T=0}
Intron: {A=0.4, C=0.1, G=0.1, T=0.4} }

The end goal is to find the total probability.Let the input sequence be:

CCCGCCATGCACGAGACACCAGCCACCTAG .Where ATG is the start codon, CACGAGACACCA is the exons, G is the splicing point, CCACC is the intron and TAG are a stop codon.CACGAGACACCAGCCACC is the region to find exon and intron. Using the concept, we get,Start (1) +C (0.9×0.25) +A (0.9×0.25) +C (0.9×0.25) +G (0.9×0.25) +A (0.9×0.25) +G (0.9×0.25)+A (0.9×0.25) +C (0.9×0.25) +A (0.9×0.25) + C (0.9×0.25) +C (0.9×0.25) +A (0.1×0.25) +G (1×0.95) + C (0.9×0.1) +C (0.9×0.1) +A (0.9×0.4) +C (0.9×0.1) +C (0.1×0.1)=5.09.

Conclusion:

Though a very difficult and time-consuming approach,the Hidden Markov model proved to be very useful in solving the above problems in asatisfactory manner. One needs to have sufficient knowledge in the domains of bioinformatics, statistics, core genetics and programming to fully understand its use.Itsapplicability is limited to only those problems that have independent states. Other disadvantages include thatit doesn't hold good for RNA folding algorithms when the relationships are local. HMM often has a number of unstructured parameters.A fully connected transition diagram can lead to over-fitting; to overcome situations like this one can use other types of HMM such as factorial or hierarchical HMMs. HMM is unable to capture higher order correlation among amino acids in a protein molecule. The Hidden Markov Model in spite of the cons is still a very valuable tool in solving problems in multiple domains.The reading we obtained was 5.09.

References:

1. Z. Ghahramani,"An Introduction to Hidden Markov Models and Bayesian Networks," International Journal of Pattern Recognition and Artificial Intelligence, vol. 15, no. 1, pg. 9–422001.
2. Meyer, I.M. and Durbin, R. (2002) Comparative ab initio prediction of gene structures using pair HMMs. Bioinformatics, 18, 1309-1318.
3. Munch K, Krogh A. Automatic generation of gene finders for eukaryotic species. BMC Bioinform. 2006; 7:263
4. Byung-Jun Yoon, "Hidden Markov Models and their Applications in Biological Sequence Analysis", Curr Genomics. 2009 Sep; 10(6): 402–415.
5. Mathe, C., Sagot, M.-F., Schiex, T. and Rouze, P. (2002) Current methods of gene prediction,their strengths and weaknesses. Nucleic Acids Research, 30, 4103-4117.
6. L. R . . .Rabiner B. H. Juang," An Introduction to Hidden Markov Models", IEEE ASSP MAGAZINE JANUARY 1986,pg 1-12

7. Durbin R, Eddy S, Krogh A, Mitchison G. Biological Sequence Analysis. Cambridge, UK: Cambridge University Press; 1998.
8. Pachter L, Alexandersson M, Cawley S. Applications of generalized pair hidden Markov models to alignment and gene finding problems. J. Comput. Biol. 2002; 9:389–399
9. Srabanti Maji and Deepak Garg,Gene Finding Using Hidden Markov Model, 2012 Journal of Applied Sciences, 12: 1518-1525.
10. Birney, E., 2001. Hidden markov models in biological sequence analysis. IBM J. Res. Dev., 45: 449-454.
11. Mario S,Stephan W,"Gene prediction with a hidden Markov model and a new intron submodel Bioinformatics, Volume 19, Issue suppl_2, 27 September 2003, Pages ii215–ii225.
12. MarioStanke, "Gene Prediction with a Hidden Markov Model", G¨ottingen, 2003,pg 1-104.
13. Bronaˇ Brejova´ Bioinformatics research group," Hidden Markov models in gene finding", David R. Cheriton School of Computer Science University of Waterloo, February 27,pg 1-21.
14. Mathe Zoltan and Korosi Zoltan Lecture series notes " Hidden Markov Model in Bioinformatics" , Scoala Gimnaziala Augustin, Romania 2006.

*****