International Journal of Scientific and Innovative Mathematical Research (IJSIMR)
Volume 4, Issue 3, March 2016, PP 1-7
ISSN 2347-307X (Print) & ISSN 2347-3142 (Online)
www.arcjournals.org

Hidden Markov Model in Biological Sequence Analysis– A Systematic Review

R. Sasikumar
Assistant Professor
Department of Statistics
Manonmaniam Sundaranar University
Tirunelveli, Tamil Nadu, India
sasikumarmsu@gmail.com

V. Kalpana
Research Scholar
Department of Statistics
Manonmaniam Sundaranar University
Tirunelveli, Tamil Nadu, India
kalpanaselwyn@gmail.com

Abstract: *For biological sequence analysis Hidden Markov Model (HMM) have been used widely in many applications. It has provided solution for various biological sequence analysis problems. In this paper, we first elucidate the fundamentals of HMM, biological sequence analysis and description of the most important algorithms of HMM. This paper especially focusing on HMM and its various types like Profile Hidden Markov Models (PHMMs) and Pair Hidden Markov Models (Pair HMM). Then we have discussed the major bioinformatics applications on HMM in biological sequence analysis problem.*

Keywords: *Markov chain, Hidden Markov Model, Profile Hidden Markov Model, bioinformatics, sequence alignment, biological sequence analysis.*

1. INTRODUCTION

In many genome sequence projects, huge amount of sequence data is available in the successfully completed projects. The available information provides numerous biological functions in cells. Computational methods play a vital role in extracting the meaningful information from the huge data. Even though numbers of processing models and algorithms have been used in biological sequence analysis, HMM is one of the best model for the same. HMM were first used in speech recognition by Rabiner in the year 1989. The algorithms and mathematical foundations of HMM are clearly explained in speech recognition [1]. Many problems in biological sequence analysis have been resolved due to the success of HMMs in engineering. For biological readers, Durbin et.al 1998 [2] discussion is very useful. Churchchill et.al 1989 [3] explained the DNA sequence as a stochastic process the states of a Markov chain. Brown et.al (1993) [4] used HMM to derive for protein families. Asai et.al (1993) [5] introduced the prediction system of protein secondary structure by HMM. Baldi et.al (1994) [6] established the algorithm for the transition and emission parameters of HMM. Eddy (1995) [7] initiated simulated annealing method and produced multiple sequence alignments from unalign proteins or DNA sequences. Sojilander et.al (1996) [8] present a method for detecting the weak but significance protein sequence homology. Bateman et.al (1996) [9] presented HMM to detect Fibronectin type III domins in yeast. Birney et.al (1997) [10] developed code generating language for biological sequence comparison. Barrett et.al (1997) [11] analyzed scoring methods to compare probability of sequence generated by HMM. Dalgaard et.al (1997) [12] used HMM to identify the KlbA proteins which is used in the formation of surface- associated protein complexes. Sonnhammer et.al (1997) [13] apply the domineer algorithm to cluster an align protein sequences after removing Pfam-Adomains. Francesco et.al (1997) [14] apply the HMM to alpha class proteins even when to detectable primary amino acid sequence similarity is present. Ahola et.al (2003) [15] finding the efficient estimation of emission probabilities in profile HMM. Liang et.al (2007) [16] proposed Bayesian approach base calling for DNA sequence analysis using HMM. Madera (2008) [17] introduced Profile comparer (PRC) program for scoring and aligning profile HMM of protein families.

This paper is organized as follows: section 2 describes the notation, definitions and basic problems of HMMs. Then we explained the practical example of HMM. Section 3 provides on description on Profile HMM and their applications in sequence alignment. We also explained construction of Profile

HMM with example. Section 4 describes the Pair HMM and their applications in sequence alignment. Then we explained method of pair HMMs with example.

2. HIDDEN MARKOV MODEL

HMM is a stochastic model which is not directly observable. But describes the observable events that are depends on internal factors. The observable events are represented as a symbols, where the invisible factor involved in the observation is represented as a state. Among the two stochastic processes, one processes is called the hidden states and other is visible process or observable symbols. So it is also called a doubly embedded stochastic Process [1]. There are two layers, visible and invisible presents in HMM, which is very useful in many real world problems. Thus if $S=\{S_n, n=1,2,...\}$ is Markov process and $V=\{V_k, k=1,2,...\}$ is a function of S, then S is a Hidden Markov process that is observed through V and we can write $V_k = f(S_k)$ for some function f. In this way we can regard S as the state process that is hidden and V as the observation process that can be observed.

A HMM is usually defined as 5-tuple (S, V, A, B, Π) where

$S=\{S_n, n=1,2,...S_N\}$ is a finite set of N hidden states.

$V=\{v_1, v_2,...v_M\}$ is a finite set of M possible observed states.

$A=\{a_{ij}\}$ is transition probabilities of the state , where a_{ij} is the probability that the system goes from state S_i to state S_j.

$B=\{b_i(v_k)\}$ are the observation probabilities where $b_i(v_k)$ is the probability that the symbol v_k is emitted when the system is in state S_i

$\Pi=\{\pi_i\}$ are the initial probabilities that is π_i is the probability that the system starts in state S_t

3. FUNDAMENTAL PROBLEMS IN HMM

There are three main types of problems occurrence in the use of HMMs are

3.1 The Evaluation Problem

Given a model $\lambda=(A,B,\Pi)$ and an observation sequence $V=v_1,v_2,...,v_T$ of length T, to compute the probability that the model generated the observation sequence $P[V|\lambda]$.

$P[V|\lambda]$ is given by $P[V|\lambda]=\sum_U [V|U,\lambda]P[U|\lambda]$

Where, $U=u_1, u_2, u_3 \ldots u_t$ is a fixed sequence

$P[V|U,\lambda]$ is the probability of observation sequence V for the state sequence U and $P[V|\lambda]$ is the probability of the sequence U for a given model.

Assume that observations are independent; the two probabilities are given by

$$P[V|U,\lambda]=\prod_{t=1}^{T} P[V_t|U_t,\lambda]$$

We obtain

$$P[V|\lambda]=\sum_U P[V|U,\lambda]P[U|\lambda]$$

3.2 The Decoding Problem

Given a model $\lambda=(A,B,\Pi)$, compute the most likely sequence of hidden states that could have generated a given observation sequence.

$U^* = \{u_1^*, u_2^*, u_3^* ... u_t^*\}$ for a given observation sequence V= $\{v_1, v_2, v_3 ... v_t\}$

Let the function $\arg\max_x(y)$

$$\arg\max_u \{P[U \mid V, \lambda]\} = \arg\max_u [P[U, V \mid \lambda]$$

3.3 The Learning Problem

Given a sequence of observations, find an optimal model. The learning problem is usually solved by the Baum-Welch algorithm.

$$\lambda^* = \arg\max_\lambda \{P[V \mid \lambda]\}$$

Sometimes it is very difficult to find an optimal model, then we choose the model parameter P[V|λ] is locally maximized. This method is an iterative solution called Baum-welch algorithm.

3.4 A simple biological example

Eddy 2004 [7] explained, the detailed description of basic biological related example of HMM. In sequence analysis problem, the hidden state path can be identified by already given sequence, this sequence is developed many state paths but our interest is to find one path which has highest probability. For example, we assume a protein coding gene A, here we have to find the locations of exons and introns in the given sequence. Firstly, we predict the state sequence B in the HMM that best describes A then we have find out the best B, it is easily predict the locations of the exons and introns.

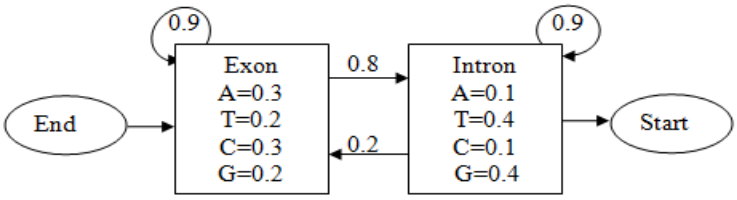

Figure 1. *Example of Hidden Markov Model in gene sequence problem*

In Figure 1, the square boxes represent exon and intron. Inside the box indicate that emission probability of 'A' 'T' 'C' and 'D'. The arrow key indicate that transistion probability of exon and intron.

4. PROFILE HIDDEN MARKOV MODEL

Sequence alignment is a method of writing one particular sequence on the top of other sequences where the remaining in one position are said to have a common basis. If the same letter comes in both sequences then this position has been placed in evaluation. If the letters vary it is assumed that the second come from a formal letter. Some sequences have variable in length but same sequences are explained through insertions or deletions in sequences. Hence, a letter or a total of letters may be paired up dashes in the following sequence to show such an insertion or deletion. Hence insertion in a one sequence usually shown as a deletion of the other one. PHMMs are remarkable types of HMM and much used in biological sequence analysis. Modern bioinformatics systems use the effective technique of MSA in all of their applications. The biomedical method and algorithms used in MSA have enormous importance in solving a series of related biological problems. The renowned and broadly used statistical method of characterizing the unique properties of the reaming of a genomic or proteomic pattern is the HMM approach. PHMM has proved to provide a better solution for MSA.

PHMM was introduced by Krogh (1994). PHMM represents a profile of MSA that specifies flow of the modeling sequence profile. It has a left to right structure does not contain any cycle. Hidden state present in Profile HMM are position specific. They are Match state M_k , Insert state I_k , Delete state D_k

4.1 Example of PHMM

The following example explains working of PHMM. Let us build a PHMM based on the multiple alignment described in Fig 2. We can know, the alignment which given has four columns where the base frequencies in the respected columns are vary from each other. To explain the symbol frequencies in the K^{th} column of the alignment K^{th} match state M_k of the PHMM is used. A symbol in a newly observed sequence matches the K^{th} symbol in the consensus sequence of the actual alignment said to be 'match state'. Finally, we came to know the length of the consensus sequence is similar to the number of match states in the resulting PHMM. The observed symbol frequencies in the K^{th} consensus column are reflected by the emission probability $e(x|M_k)$ at the K^{th} match state M_k. Let us consider, the new observed sequence which is shorter than the consensus sequence. In this, the consensus sequence has one or more bases which are not present in the observed sequence. Let us consider D_k is the K^{th} delete state that is used in the deletion of the K^{th} symbol in the actual consensus sequence. In delete states where symbols are missing as D_k is a silent state which is used to interconnect the neighboring states. We can get the PHMM after us adding the insert states and delete states to the ungapped HMM.

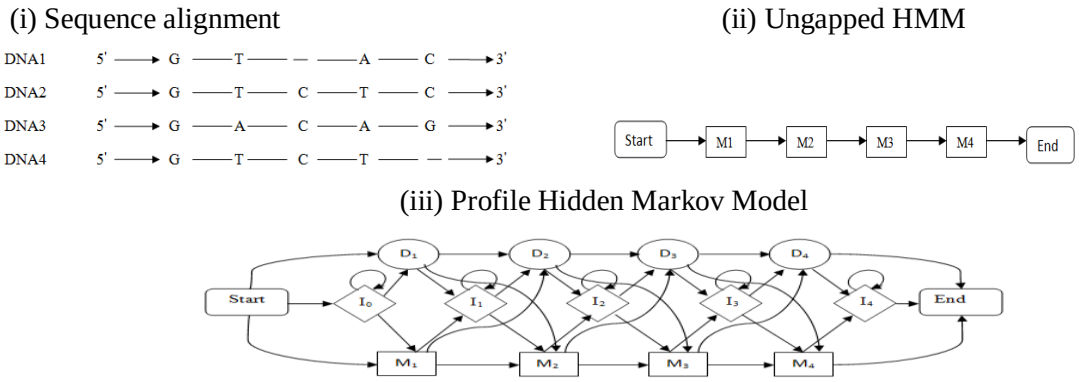

Figure (2). *Construction of PHMM*

4.2 Applications of PHMM

PHMM have been highly used for modeling and analyzing biological sequence as it is very convenience and effectiveness in marking sequence profiles. Recently there are two software packages available which are Sequence Alignment and Modelling system (SAM) and HMMER. These are used to build and trainPHMM easily. The software packages deliver very easy tools for applying PHMM to many sequence analysis problem. Baldi et.al (1994) [6] first proposed the modeling the characteristics of a number of protein families such as globins, immunoglobulins and kinases. Eddy (1995) [18] introduced the simulated annealing method and produced multiple sequence alignments from unaligned protein or DNA sequences. Difranceco et.al (1999) [19] finds the method for fold recognition from secondary structure predictions of proteins which is based on HMM. They used FORESST web server of the library of HMM of structural families. Smith et.al (2003) [20] communicates the HMM and optimized sequence alignments for related gene or proteins. Edger and Sjolander (2004) [20] Present a COACH aligns two multiple sequence alignments by constructing a profile HMM. Wistrand and sonnhammer (2004,2005) [21,22] improved PHMM discrimination by adapting transition probabilities and also discussed PHMM performance by assessment of critical algorithmic features in SAM and HMMER. Söding(2005) [23] illustrated PHMMs to detect the protein homology and sequence alignment of protein structure prediction, function prediction and evaluations Anne et.al (2006) [24] developed a jumping PHMM and applied the human immunodeficiency virus and hepatitis c virus to simulated recombined genome sequences.

5. PAIR HMM

Pair HMM finds sequence alignment and evaluates significance of aligned symbols. Original HMM generates only a single sequence where as Pair HMM generates aligned pair of sequence. Pair HMM compare two sequence where find out whether functionally related. To compare two biological sequences we have, to align them based on their similarity compute the alignment score and evaluate the statistical significance of the predictive alignment. The best alignment between sequences can be determined by defining a reasonable scoring scheme on based on that the alignment that maximizes the alignment score can be chosen.

5.1 Example of Pair HMM

Let us consider the two sequence A= $a_1a_2a_3a_4a_5$=AATTC and B=$b_1b_2b_3b_4b_5$=TTCAA. It is Illustrated in the figure 3. We assume that the hidden state sequence is X=$I_AI_AYYYI_BI_B$

A (seq)	=	A	A	T	C	-	-
B (seq)	=	-	-	T	C	-	-
X (states)		I_A	I_A	Y	Y	I_B	I_B

As we can see, a_1 and a_2 individually are emitted at I_A, hence they are not aligned to any based in B. The pairs (a_3,b_1), (a_4,b_2) and (a_5,b_3) jointly emitted at Y and the pairs are aligned to each other. Lastly b_4 and b_5 are individually emitted at I_B as unaligned bases. In this example, hidden state sequence X and the two observed sequence A and B have one- to –one relationship with each other.

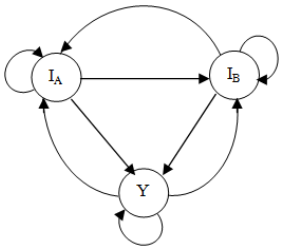

Figure 3. *Example of Pair HMM*

5.2 Applications of Pair HMMs

Meyer and Durbin (2002) [25] presented the comparative abinito prediction of gene structures using Pair HMM. Yasubumi (2003) [26] proposed the Pair HMM on tree structures. They explained extension of Pair HMM to aligning the tree structures and provide a probabilistic framework for aligning RNA sequences. Knudeen and miyamoto (2003) [27] used Pair HMM in optimal evolutionary for symbol insertions and deleations. Loytynoja and milinkovitch (2003) [28] presented multiple sequence algorithm uses a Pair HMM to find pairwise alignment and to find estimate their alignment reliability. Alexanderson et.al (2003) [29] used generalized Pair HMM to find cross-species gene and alignment. Chuong et.al (2005) [30] used multiple sequence alignment algorithm also make use of Pair HMM. Majoros et.al (2005) [31] proposed implementation of Pair HMM for comparative approach for gene prediction. Wang et.al (2005) [32] discussed another method called MCALIGN2. This method was faster and accurate global pairwise alignment of non-coding DNA sequence based on explicit models of indel evaluation. Matsui et.al (2005) [33] Pair HMM to align tree adjoining grammer trees. They presented the extension of Pair HMM to align RNAs with more complicated secondary structures, including pseudo knots.

6. CONCLUSION

The principle advantages of HMMs are very easy to use, minimum training sets are required and easy understanding of the structure of phenomenon. There are various gene prediction tools are available in tools which are the base of HMM gives higher accuracy. There are many software like SAM, HMMER available to use HMM in biological sequence analysis. The HMM can be easily applied to sequence alignment using software. It has one drawback that is it takes more time to get the result of biological sequence problems. In this paper, we have evaluated the present situation of the HMM in biological sequence analysis and prosperity of available applications. This review provides detailed types of HMM like PHMM, Pair HMM and explained their applications in biological sequence analysis.

REFERENCES

[1] Rabiner L.R, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proceeding of the IEEE, 77(2), 257-86 (1989).

[2] Durbin R, Eddy S, Krogh A, Mitchison G, Biological Sequences Analysis: Probabilistic Models of Proteins and Nucleic Acid, Cambridge University Press, Cambridge(1998).

[3] Churchchill G.A, Stochastic Models for Heterogeneous DNA sequences, Bull Math Biol, 51(1), 79-94(1989).

[4] Brown M, Hughey R, Krogh A, Mian I.S, SjSlander K and Haussler D, Using Mixture Priors to Drive Hidden Markov Models for Protein Families, Proceedings of the ISMB-93, 47-55 (1993).

[5] Asai K., Hayamizu S and Handa K, prediction of protein secondary structure by the Hidden Markov model, Bioinformatics, 9(2), 141-146 (1993).

[6] Baldi P, Chauvin Y, Hunkapiller T and McClure M.A, Hidden Markov Models of Biological Primary Sequence Information, Proc. Natl.Acsd.Sci. USA, 91, 1059-1063 (1994).

[7] Eddy S.R, What is Hidden Markov Model?, Nature Biotechnology, 22, 1315-1316 (2004).

[8] Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian I.S and Haussler D, Dirichlet mixtures: A method for improving detection of weak but significant Protein Sequence Homology, Applic.Biosci, 12 (4), 327-345.

[9] A.Bateman , C.Chothia, Fibronectin III Domain in Yeast Detected by a Hidden Markov Model, Current Biology, 6 (12),1544-1547.

[10] Birney E, Durbin R, Dynamite: A flexible code generating language for dynamic programming methods used in sequence comparison. ISMB proceedings (1997).

[11] Barrett C, Hughey R and Karplus K, Scoring hidden markov models, Bioinformatics, 13 (2), 191-199 (1997).

[12] Dalgaard J.Z, Moser M,J, Hughey R and Mian S, Statistical modeling, Phylogentic analysis and structure prediction of a protein splicing domain common to Inteins and Hedgehog Proteins, Journal Of Computational Biology 4(2), 193-214 (1997).

[13] Sonnhammer E.L.L, Eddy S.R, and Durbin R, PROTEINS: Structure, Function, and Genetics, 28, 405-420 (1997).

[14] Di Francesco V, Garnier J and Munson P.J, Protein Topology Recognition from Secondary Structure Sequences: Application of the HiddenMarkov Models to the Alpha Class Proteins, Journal of molecular biology, 267 (2), 446-463 (1997).

[15] Ahola v.,Aittotallio T., Uusipaikka E., and vihinen M., Efficient estimation of emission probabilities in Profile Hidden Markov models, Bioinformatics, 19 (18), 2359-2368 (2003).

[16] Liang K.C, Wang X, anastassiou D, Bayesian Base calling for DNA Sequence Analysis Using Hidden Markov Models, Transactions on Computational Biology and Bioinformatics, 4(3), 430-440 (2007).

[17] Madera M., Profile comparer: a program for scoring and aligning Profile Hidden Markov models, Bioinformatics, 24 (22), 2630-2631 (2008).

[18] S.R. Eddy, Multiple Alignment Using Hidden Markov Models, Proc of Third Int. conf.on Intelligent Systems for Molecular Biology, 3, 114-120 (1995).

[19] Di Francesco V.,Munson P.J., and Garnier J., FORESST: fold recognition from secondary structure predictions of proteins, Bioinformatics, 15 (2), 131-140 (1999).

[20] Edger R.c., Sjolander K., COACH:profile-profile alignment of protein families using Hidden Markov Models, Bioinformatics, 20 (8), 1309-1318 (2004).

[21] Wistrand M, and Sonnhammer E.L, Improving Profile HMM discrimination by adopting transition probabilities, journal of molecular biology, 338(4), 847-854 (2004).

[22] Wistrand M, and Sonnhammer E.L, Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER, BMC Bioinformatics, 6:99 (2005).

[23] Söding.J., Protein homology detection by HMM–HMM comparison, Bioinformatics,21 (7), 951-960 (2005).

[24] Schultz A.K, Zhang M, Leitner T, Kuiken C, Korber B, Morgenstern B and Stanke M, A Jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes, BMC Bioinformatics, 7:265 (2006).

[25] Meyer M.I and Durbin R, Comparative ab initio prediction of gene structures using pair HMMs, Bioinformatics, 18 (10), 1309-1318 (2002).

[26] Sakakibara Y, Pair hidden Markov models on tree structures, Bioinformatics, 19 (1), 1309-1318 (2003).

[27] Knudsen B and Miyamoto M.M, Sequence Alignments and Pair Hidden Markov Models Using Evolutionary History, journal of molecular biology, 333 (2), 453-460 (2003).

[28] Loytynoja A and Milinkovitch M.C, A hidden Markov model for progressive multiple alignment, 19 (12), 1505-1513 (2003).

[29] Alexandersson M, Cawley S, and Pachter L, SLAM: Cross-Species Gene Finding and Alignment with a Generalized Pair Hidden Markov Model, Genome research, 13 (3), 496-502 (2003).

[30] Chuong B. Do, Mahabhashyam M.S.P, Brudno M, and Batzoglou S, ProbCons: Probabilistic consistency-based multiple sequence alignment, Genome research, 15 (2), 330-340 (2005).

[31] Majoros W.H, Pertea M and Salzberg S.L, Efficient implementation of a generalized pair hidden Markov model for comparative gene finding, 21 (9),1782-1788 (2005).

[32] Wang J, Keightley P.D and Johnson T, MCALIGN2: Faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution, BMC Bioinformatics, 7:292 (2006).

[33] Matsui H, Sato K and Sakakibara Y, Pair stochastic tree adjoining grammars for aligning and predicting pseudoknot RNA structures, Bioinformatics, 21 (11), 2611-2617 (2005).

[34] Bynng_jun yoon, Hidden Markov Models and their applications in biological sequence analysis, current genomics, 10(6), 402-415 (2009).

AUTHORS' BIOGRAPHY

Dr.R.Sasikumar, is presently working as an Assistant Professor in the Department of Statisatics, Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India. He has been 10 years of Teaching and 13 years of Research experiences. His Research Areas of interest are Stochastic Process and its Applications, Statistical Quality Control and Reliability. He has published 21 (Twenty One) research papers in the Reputed International and National level Journals and Proceedings.

V. Kalpana, was completed M.Sc., (Statistics with computer applications) and M.Phil., (Statistics) Currently she has pursing Ph.D., in Statistics under the guidance of Dr. R. Sasikumar, Assistant Professor, Department of Statisatics, Manonmaniam Sundaranar University, Tirunelveli,Tamilnadu, India. Her research interest are Stochastic Models and Statistical Quality Control.