IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 5, Ver. III (Sep – Oct. 2014), PP 01-07
www.iosrjournals.org

HMM and its application in MSA

[1]Vijay Kumar Verma, [2]Biresh Kumar, [3]Ram Krishna Kumar
[1]Asst. Professor Dept. of CSE NSIT, [2]Research Scholar NIT Patna, [3]Asst. Professor Dept. of Mathematics NSIT

Abstract: *This document gives an insightinto HMM and its application in context of Multiple sequence alignment (MSA). Various computational approaches proliferatedin response to resolve the complexities of Human Genome sequence. An HMM is probabilistic competitive approachthat can be applied to resolve the key issues like Gene prediction, multiple sequence alignment, pattern recognition .This paper describe HMM approach of MSA and its significance.*

Keywords: *HMM, Multiple sequence alignment, Pattern recognition,Maximum likelihood,R,Viterbi algorithm ,Back Propagation.*

I. Introduction

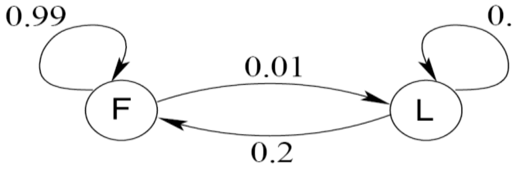
HMM(Hidden Markov Model) is a probabilistic finite automaton models that were first applied in the speech recognition [1] field later seeing its potential applied to several domains and areas like protein and DNA sequence, pattern recognition [5]. The type of questions that we can ask using HMM is: Does the sequence belong to a particular family or assuming the sequence belong to certain family what we can say about its internal structure. In fact, HMM is best suited for the problem like search and alignment for biological sequence analysis [9].HMM are very valuable in computational biology as it allow a search or alignment algorithm based on firm probability and its flexibility to train the parameter s with known data.

Let's explore the basic component of HMM and Mathematical annotations.

II. Components and Mathematical annotations

A. Components

-Observed variables (Emitted Symbols)
-Hidden variables
-Relationship between these two (transitionprobabilities)
The goal is to find the most likely explanation for the observed variables.
Demonstration:
Using basic example of dishonest casino:



F-fair die
L-Loaded die

Probab(F(1))=Probab(F(2))=Probab(F(3))=Probab(F(4))=Probab(F(5))=Probab(F(6))=1/6
Probab(L(1))=Probab(L(2))=Probab(L(3))=Probab(L(4))=Probab(L(5))=1/10
Probab(L(6))=1/2

In general **Probab(X(n))**:probability of emitting n in state X.
These are **emission probabilities.**
Transition probabilities (probability of switching from one state to another):
Probab(F→L)=0.01
Probab(L→F)=0.2
Probab(F→F)=0.99
Probab(L→L)=0.8

Transition table:

	F	L
F	0.99	0.01
L	0.2	0.8

Observable: the series of die tosses might result into sequences like:
341526664523...

Hidden: what are the probable series of states that result into the above observable sequences:
FFFFLLLFFFLLL...

What we are interested in :

a) When was a fair die used

b) When was a loaded one used

Making Inference:

Maximum Likelihood: Determine which explaination(path of hidden sequences) is most likely.

Total Probability: Considering all path that could have produced the observable sequence.

B. Mathematical Annotatoins

X is the sequence of symbol emitted by model(HMM)

$X_i$is sequence of symbol emitted at time i.

A path $\pi$is a sequence of states. $\pi_i$isith state in π

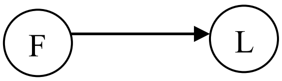
Transition probability$a_{kr}$is probability of making transition from state **k** to state **r.**

i.e. probability of state **r** given state **k**

$$a_{kr}= Pr(\pi_i = r \mid \pi_{i-1=k})$$

i.e. probability of r at state sequence position i given probability of k just before the(i-1) state sequence r.

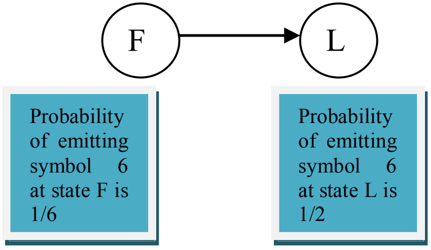
$a_{FL} = 0.01$



Emission probabilityek(b)is the probability of observable sequence **b** at state **k**.

ek(b)=Pr(X_i=b| $\pi_{i\,=}$k)

eF(6)=1/6 **eL(6)=1/2**


F
L
Probability of emitting symbol 6 at state F is 1/6
Probability of emitting symbol 6 at state L is 1/2

Formally HMM can be defined as 5-tuple structure

M= {**π, a_{kr} ,O,e,q_0**}

Where,

π=total no. of states

$a_{kr=}$transition probabilities from state k to r

O=total no of observed sequences

E= emission probability

q_0= beginning state

so dishonest casino can be mathematically formulated as

M= {**π, a_{kr} ,O,e,q_0**}

π={F,L}

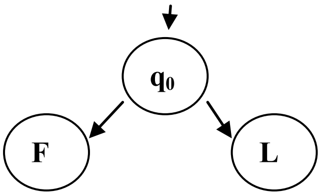
$a_{kr=}${F-->F(0.99),F-->L(0.01),L-->L(0.08),L-->F(0.02)}

O: one probable string is (626) as outcomes of three times dice throw

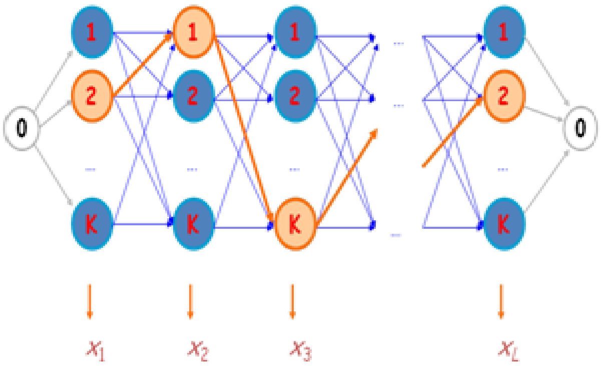
e=

eπ(i)	**1**	**2**	**3**	**4**	**5**	**6**
F	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**	**1/6**
L	**1/10**	**1/10**	**1/10**	**1/10**	**1/10**	**1/2**

q_0=inital state

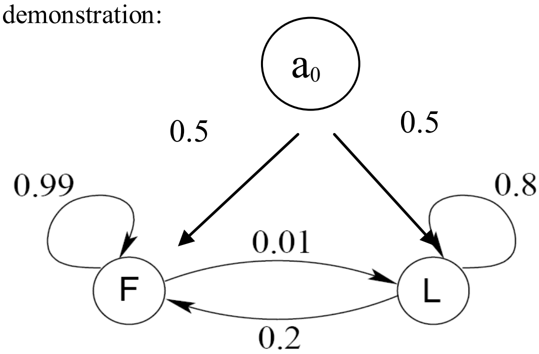


I. A **PARSE** OF COMPLETE SEQUENCE IN GENERIC WAY CAN BE ANNOTATED AS:



$$\Pr(\mathbf{X},\pi)=\mathbf{a_0\pi_1}\prod_{i=1}^{L}\mathbf{e\pi_i\,(X_i).a\pi_i\pi_{i+1}}$$

demonstration:



Case X= $\{X_1,X_2,X_3\}$ =(6,2,6)

Given observed sequence 626 what are the most likely hidden states sequences that resulted into the observed sequence one.

Guess 1:π=FFF

Pr(X, π)=a_{0F}.**eF(6)**.a_{FF}.**eF(2)**.a_{FF}.**eF(6)**

eF(6)=1/6	**eF(2)=1/6**	**eF(6)=1/6**
a_{0F}=**0.5**	a_{FF}=**0.99**	a_{FF}=**0.99**

=0.5*1/6*0.99*1/6*0.99*1/6=**0.00227**

Guess 2:π=LLL

Pr(X,π)=a_{0L}.**e_L(6)**.a_{LL}.**e_L(2)**.a_{LL}.**e_L(6)**

eL(6)=1/2	**eL(2)=1/10**	**eL(6)=1/2**
a_{0L}=**0.5**	a_{LL}=**0.8**	a_{LL}=**0.8**

=0.5*1/2*0.8*1/10*0.8*1/2=**0.008**

Guess 3:π=LFL

Pr(X,π)=a_{0L}.**e_L(6)**.a_{LF}.**e_F(2)**.a_{FL}.**e_L(6)**

eL(6)=1/2	**eF(2)=1/6**	**eL(6)=1/2**
a_{0L}=**0.5**	a_{LF}=**0.2**	a_{FL}=**0.01**

=0.5*1/2*0.2*1/6*0.01*1/2=**0.0000417**

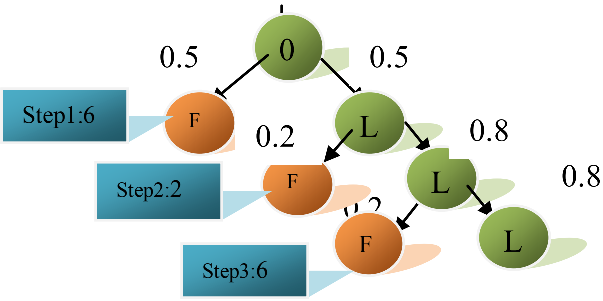
From the above three cases it is clear that the most probable path to produce the emitted symbol (626) is LLL.The most likely pathπ* (the best possiblestate sequence) is defined as:

π*=argmaxPr(π|x)
π

To calculate π* traverse all possible way of emitting symbol x.GivenO(observed sequences) find
π*=argmaxPr(π|O)
The goal is to maximizePr(π*|O)by choosing best π*
π*is bounded by search space $|\pi|^{|O|}$
in case of dishonest casino for observed sequence (626)

Search space is 2^3 as the total no of state is 2(F,L) and sequence length is 3.

TREE ANNOTATION OF BEST HIDDEN PATH CALCULATIONFOR(626):



Step 1:	a_{0L}.**e**$_L$**(6)>** a_{0F}.**e**$_F$**(6): 0→F block**
Step2:	a_{LL}.**e**$_L$**(2)>**a_{LF}.**e**$_F$**(2): L→F block**
Step3:	a_{LL}.**e**$_L$**(6)>**a_{LF}.**e**$_F$**(6): L→F block**

The resulting path sequence that leads to 626 with highest probability is right slanted LLL.

Viterbi Algorithm:

is a dynamic programming algo that allow us to compute the most probable path.

Then $$v_k(i) = \text{Prob. of path} \langle \pi_1, \cdots, \pi_i \rangle \text{ most likely to emit } \langle x_1, \ldots, x_i \rangle \text{ such that } \pi_i = k$$

$$v_k(i) = e_k(x_i) \max_r \left(v_r(i-1) a_{rk} \right)$$

Initialization: i=0

$$v_0(0) = 1, \quad v_k(0) = 0 \text{ for } k > 0$$

Recursion: i=1 to L for each state k

$$v_k(i) = e_k(x_i) \max_r \left(v_r(i-1) a_{rk} \right)$$

Termination:

$$\Pr(x, \pi^*) = \max_k \left(v_k(L) a_{k0} \right)$$

Demonstration:

$V_F(1)=e_F(X_1).\max_r(V_0(0).a0F)$=1/6*1*1/2
$V_L(1)=e_L(X_1).\max_r(V_0(0).a0L)$=1/2*1*1/2
$V_F(2)=e_F(X_2).\max_r(V_F(1).aFF)$=1/6*1/12*0.99
$V_F(2)=e_F(X_2).\max_r(V_L(1).aLF)$=1/6*1/4*0.2
$V_L(2)=e_L(X_2).\max_r(V_L(1).aLL)$=1/10*1/4*0.8
$V_L(2)=e_L(X_2).\max_r(V_F(1).aFL)$=1/10*1/12*0.01
$V_F(3)=e_F(X_3).\max_r(V_F(2).aFF)$=1/6*1/6*1/12*0.99*0.99
$V_F(3)=e_F(X_3).\max_r(V_L(2).aLF)$=1/6*1/10*1/4*0.8*0.2
$V_L(3)=e_L(X_3).\max_r(V_L(2).aLL)$=1/2*0.02*0.8
$V_L(3)=e_L(X_3).\max_r(V_F(2).aFL)$=1/2*0.1375*0.01

	X_1	X_2	X_3
	6	2	6
F	1/12	0.01375	0.0022
L	¼ ←	0.02 ←	0.8

Basic operation in HMM:

Exploration	Approach applied	Complexities
Calculating **P(Qt=Si|O1O2O3...Ot)**	Forward-Backward	O(TN2)
Inference: Computing $\pi^*=\text{argmaxPr}(\pi|O)$	Viterbi-Decoding	O(TN2)

R implementation of Dishonest Casino:

Package=HMM 1.0

initHMM(States, Symbols, startProbs=NULL, transProbs=NULL, emissionProbs=NULL)

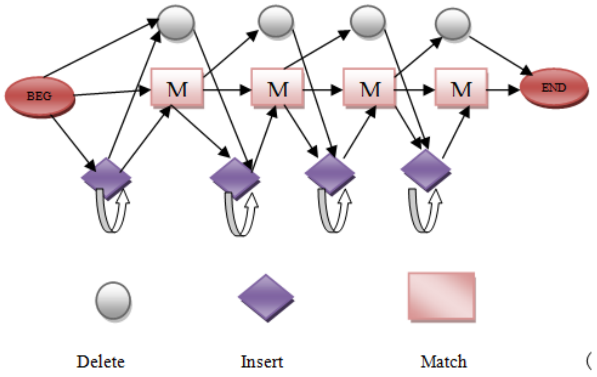
States	Vector with the names of the states.
Symbols	Vector with the names of the symbols.
startProbs	Vector with the starting probabilities of the states.
transProbs	Stochastic matrix containing the transition probabilities between the states.
emissionProbs	Stochastic matrix containing the emission probabilities of the states.

```
>hmm = initHMM(c("F","L"), c("1","2","3","4","5","6"), transProbs=matrix(c(.99,.2,.01,.8),2),
emissionProbs=matrix(c(1/6,1/10,1/6,1/10,1/6,1/10,1/6,1/10,1/6,1/10,1/6,1/2),2))
>print(hmm)
>observations = c("6","2","6")
>viterbi = viterbi(hmm,observations)
>print(viterbi)
>logBackWardProb=backward(hmm,observations)
>print(exp(logBackWardProb))
```

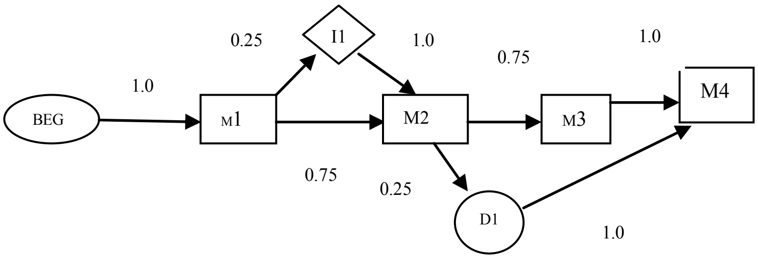
C. HMM application in MSA(Multiple Sequence Alignment)

HMM can be used to model a family of sequences (Protein or Nucleotide)[10]. For a given alignment (aligning sequences of a family), the HMM model will generate probabilistic values for each position. Then we can use the model's probabilities to compute the probability of a sequence belonging the family represented by the model.

Schematic Representation:



1) Deriving the Model (Example Demonstration):



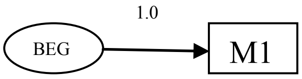
Protein Family Sequences:

N	-	F	L	S
N	-	F	L	S
N	Q	F	L	S
Q	-	N	-	T

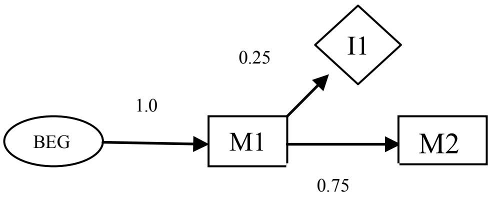
Identify the Match State, Delete State, Insertion State:

Beg-M1-M2-M3-M4-End
Beg-M1-M2-M3-M4-End
Beg-I1-M2-M3-M4-End
Beg-M1-M2-D1-M4-End

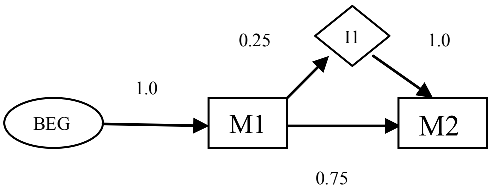
Model the Sequences:
=>From BEGIN state there is only one trasition that lead to Match state.



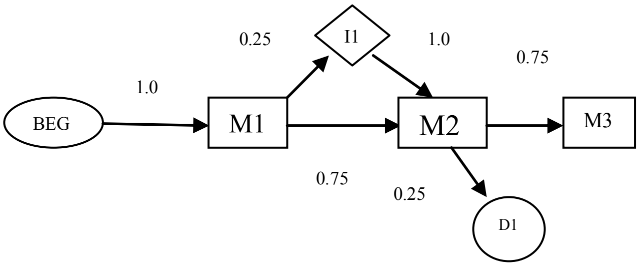
⇨ From Match state out of 4 one goes to Insert state and 3-goes to Match state.



=>From Insert state transition goes to Match state.



⇨ From M2 state out of 4, 3-goes to M3 state and one goes to delete state.



⇨ From M3 state the only transition left is to M4 state and from Delete state transition lead to M4 state.
=>Lastly from M4 it terminate with end state

Problem area that can be explored using HMM:

I. Modal Evaluation(What is the probability of the observation, use forward algorithm)
II. Path Decoding(Best state sequence for the observation,use Viterbi algorithm)
III. Model Training (Estimating model parameter use Baum- Welch algorithm)

Some Application Area of HMM:
I. Online Handwriting Recognition
II. Speech Recognition

III. Protein Sequence and Gene Sequence alignment
IV. Gene Predictin
V. Gesture Recognition etc.

III. Conclusions

This paper represent a basic introduction of hmm and its application in context of multiple sequence alignment. However hmm, encompasses a rich class of variable length probability distribution for classification purpose they may not precisely represent the true conditional distribution.

References

[1] Lawrence R. Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77(2) p.257-286, 1989

[2] Bystroff, C., V. Thorsson& D. Baker (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J MolBiol, 301, 173-90.

[3] Chen, X., M. M. Hoffman, J. A. Bilmes, J. R. Hesselberth& W. S. Noble (2010) A dynamic Bayesian network for identifying protein-binding footprints from single molecule-based sequencing data. Bioinformatics, 26, i334-42.

[4] Durbin, R., S. R. Eddy, A. Krogh & G. Mitchison. 1998. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge, UK: Cambridge University Press.

[5] Henderson, J., S. Salzberg& K. H. Fasman (1997) Finding genes in DNA with a Hidden Markov Model. Journal of Computational Biology, 4, 127-141.

[6] Humburg, P., D. Bulger& G. Stone (2008) Parameter estimation for robust HMM analysis of ChIP-chip data. BMC Bioinformatics, 9, 343.

[7] Husmeier, D. & F. Wright (2001) Detection of recombination in DNA multiple alignments with hidden Markov models. Journal of Computational Biology, 8, 401-427.

[8] Krogh, A., I. S. Mian& D. Haussler (1994b) A hidden Markov model that finds genes in E. coli DNA. Nucleic Acids Res, 22, 4768-78.

[9] Gribskov. 31..Liithy, R. & Eisenberg. D. (1990). Profile analysis. MethalaEnzymZ. 183, 146159.

[10] Barton. G. J. (1990). Protein multiple sequence alignment and flexible pattern matching. Method8 Enzymol.183, M3428.

[11] Allison. L..Wailace. C. S. &Yee, C. X. (1992).

[12] Finite-state models in the alignment ofmacromolecules. J. Md. Evd. 35, 77-89.

[13] S. Theodoridis y K. Koutroumbas, Pattern Recognition,AcademicPress, 1999.