Hidden Markov Model and Applications

Manu Madhavan

Lecture 14

Recap

- Pattern identification and discovery

Outline

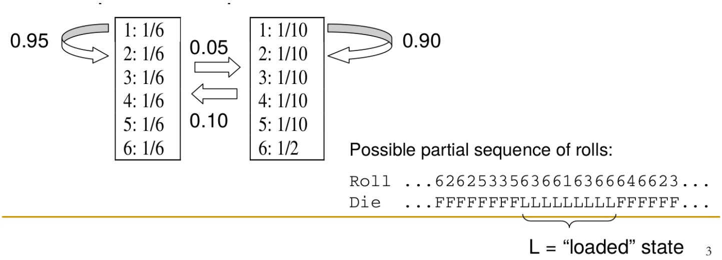
- What is HMM
- Applications of HMM- Profile HMM

Hidden Markov Model

- A casino usually uses a fair die, but sometimes (5% of the time) switches to a loaded die. Once using the loaded die, they usually keep using it (90% of the time).
- How do you know which die youre playing?
  - not sure, but have to look at many plays to see pattern
  - the “state” here is “hidden”
- How we can model this?

Hidden Markov Model

- A casino usually uses a fair die, but sometimes (5% of the time) switches to a loaded die. Once using the loaded die, they usually keep using it (90% of the time).
- How do you know which die youre playing?

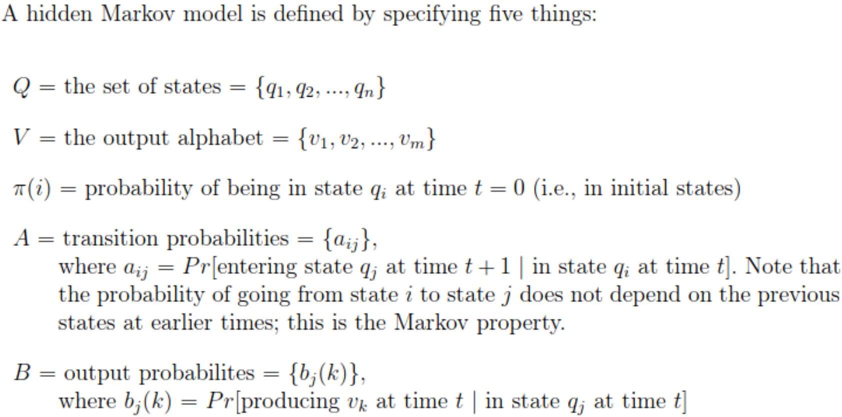


Possible partial sequence of rolls:

```
Roll  ...62625335636616366646623...
Die   ...FFFFFFFFLLLLLLLLLFFFFFF...
```

L = “loaded” state [3]

Hidden Markov Model



A hidden Markov model is defined by specifying five things:

Q = the set of states = $\{q_1, q_2, ..., q_n\}$

V = the output alphabet = $\{v_1, v_2, ..., v_m\}$

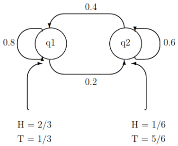
$\pi(i)$ = probability of being in state q_i at time $t = 0$ (i.e., in initial states)

A = transition probabilities = $\{a_{ij}\}$,
where $a_{ij} = Pr$[entering state q_j at time $t+1$ | in state q_i at time t]. Note that the probability of going from state i to state j does not depend on the previous states at earlier times; this is the Markov property.

B = output probabilites = $\{b_j(k)\}$,
where $b_j(k) = Pr$[producing v_k at time t | in state q_j at time t]

HMM-Example

Two biased coins, which we are flipping, and an observer is seeing the results of our coin flip



- P(q1q1q1q2q2q1q1)?
- P(HHTTTTH|q1q1q1q2q2q1q1)?
- What is the probability of the above output sequence and the above transition sequence?

HMM

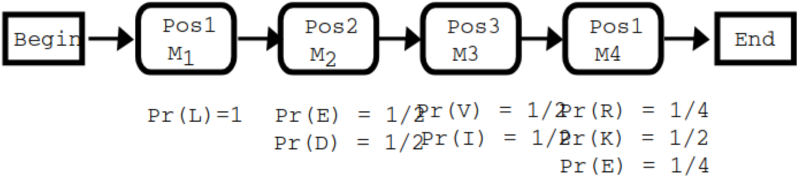
- What is the probability of the observed data O1,O2,...OT given the model? That is, calculate Pr(O1,O2,...OT|model).
- At each time step, what state is most likely? It is important to note that the sequence of states computed by this criterion might be impossible. Thus more often we are interested in what single sequence of states has the largest probability. That is, find the state sequence q1,q2,...,qT such that Pr(q1,q2,...,qT |O1,O2,...OT , model) is maximized.
- Given some data, how do we learn a good hidden Markov model to describe the data? That is, given the topology of a HMM, and observed data, how do we find the model which maximizes Pr(observations|model)?

HMM

- What is the probability of the observed data O1,O2,...OT given the model? **Forward Algorithm**
- At each time step, what state is most likely? **Viterbi Algorithm**
- Given some data, how do we learn a good hidden Markov model to describe the data? That is, given the topology of a HMM, and observed data, how do we find the model which maximizes Pr(observations|model)? **Baum-Welch-EM Algorithm**

Building HMM-Profile

- Consider the MSA:
  LEVK
  LDIR
  LEIK
  LDVE



Begin → Pos1 M_1 → Pos2 M_2 → Pos3 M3 → Pos1 M4 → End

Pr(L)=1

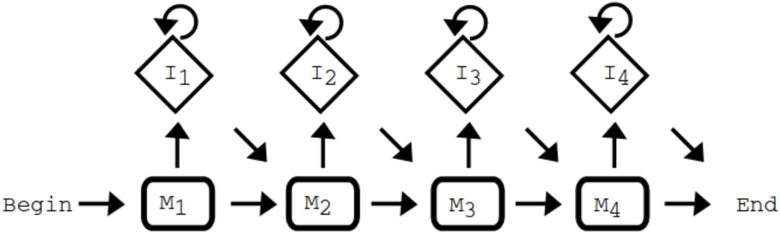
Pr(E) = 1/2
Pr(D) = 1/2

Pr(V) = 1/2
Pr(I) = 1/2

Pr(R) = 1/4
Pr(K) = 1/2
Pr(E) = 1/4

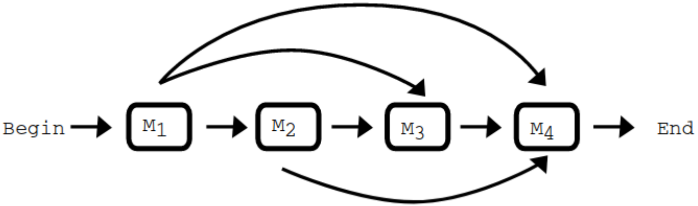
Building HMM-Profile

- Add insertions
- Insertions are portions of sequences that do not match anything in the model.



Building HMM-Profile

- Add Deletions



Building HMM-Profile

- Add Deletions

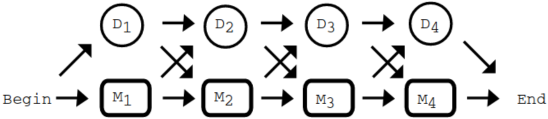


Figure 5: Deletions

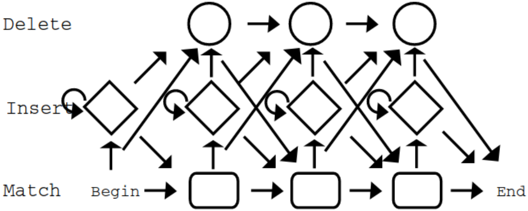


Figure 6: The complete HMM formation

Building HMM-Profile-Summary

Profile-HMM

Given a multiple sequence alignment of a particular domain family, one uses statistical methods to build a specific HMM for that domain family.
The probabilities that are required are estimated from the frequencies in the alignment, together with other data.
This HMM can then be used to test other sequences whether they match this domain family or not.
HMMs can be set up so that insertions, deletions and substitutions can be handled in sensible ways, and their probabilities estimated properly.

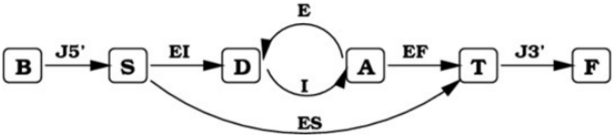
Explore: HMMER, `https://www.ebi.ac.uk/Tools/hmmer/`

Approach

- Use clustalw for MSA
- Use HMMER for profile analysis

Gene Discovery using HMM

- Find the coding and non-coding regions of an unlabelled string of nucleotides
- A gene sequence may contain regions like introns and exons, separated by splice sites, acceptors[1] and donors[2].

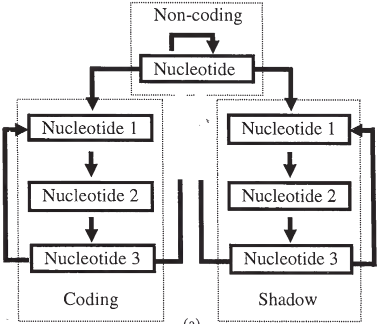


[1] splicing site at the beginning of an intron

[2] splicing site at the end of an intron, intron 3' right end

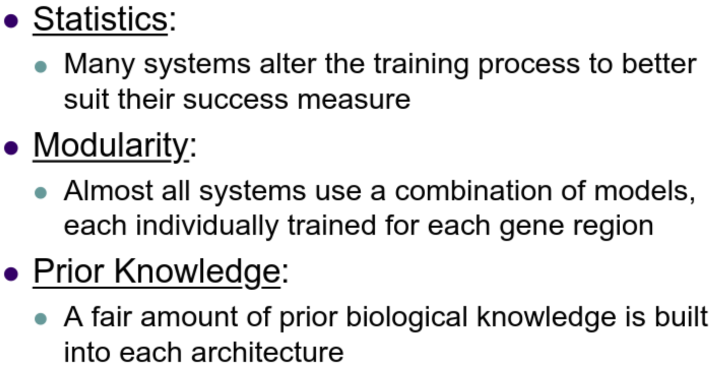
Gene Discovery using HMM

The frequencies of occurrence of mononucleotides, dinucleotides, trinucleotides, as well as higher oligonucleotides at each position on the DNA sequence are not expected to be homogenous, but vary according to whether the position is at the beginning, middle or end of the codon, i.e., it depends on the phase, or reading frame. This feature is used in calculating the probabilities of occurrence of oligonucleotides of various lengths



(a)

HMM-Advantages



- Statistics:
  - Many systems alter the training process to better suit their success measure
- Modularity:
  - Almost all systems use a combination of models, each individually trained for each gene region
- Prior Knowledge:
  - A fair amount of prior biological knowledge is built into each architecture

HMM-Disadvantages

- Markov assumption
- Local maxima may not be the global maxima
- High chance of overfitting
- Slow learning