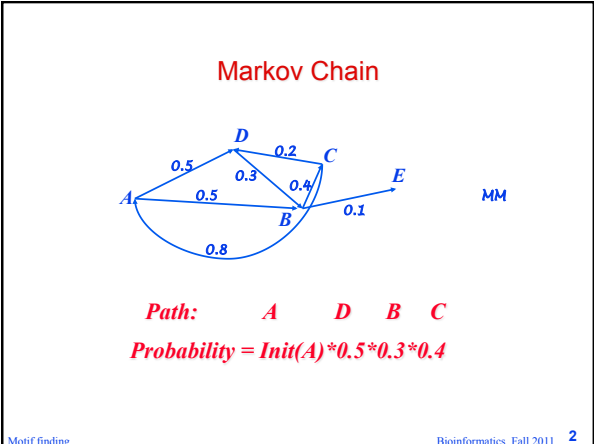


Hidden Markov Model (HMM)



- ## Motivation
- Suppose we have a (Markov Chain) model for motif and one for background, or similarly, models for
 - Fair and loaded dice
 - CpG island and background
 - Given a short sequence, one can compute its probability according to the alternative models and compute log-odds.
 - What if we have a single long sequence consisting of multiple such short sequences?
 - Combine multiple MC models into a composite model with transition probabilities between basic models
 - No longer 1-1 mapping between symbols and states
 - States are hidden, symbols are visible.
- Motif findline Bioinformatics, Fall 2011 3

- ## Questions of interest
- What is the most probable path for a sequence?
 - Decoding problem
 - What is the probability of generating a given sequence?
 - Sum over all possible paths
 - HMM generates a pdf over all strings of a given length
 - How to train an HMM using a set of sequences?
- Motif findline Bioinformatics, Fall 2011 4

Hidden Markov Model:

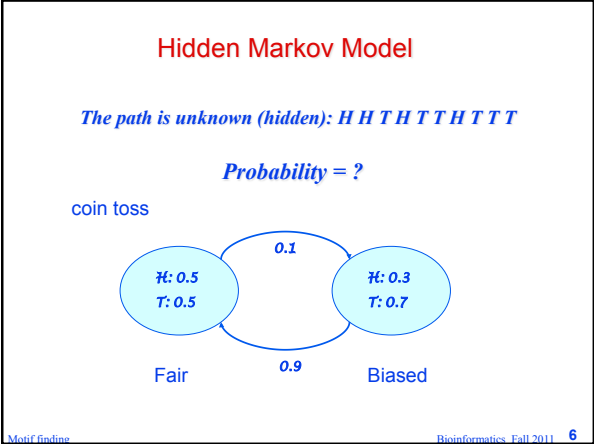
- a composition of finite number of states,
- each state emits symbols, according to symbol-emission probabilities

Starting from an initial state, a sequence of symbols is generated by moving from state to state until an end state is reached.

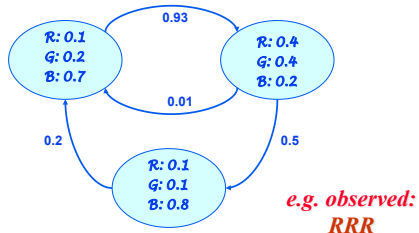
state sequence (hidden): ... ① ① ① ① ① ② ② ② ① ① ...
transitions: ① 0.99 0.99 0.99 0.01 0.9 0.9 0.9 0.1 0.99 ...
symbol sequence (observable): ... A T C A A G G C G A T ...
emissions: 0.4 0.4 0.1 0.4 0.4 0.5 0.5 0.4 0.5 0.4 0.4 ...

(Figures from Eddy, Curr. Opin. Struct. Biol.)

Motif findline Bioinformatics, Fall 2011 5



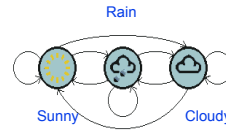
Probability of generating a given sequence



e.g. observed: **RRR**

Probability of a given sequence =
Sum of probability over ALL paths giving that sequence

Markov Chain for weather



		weather today		
		Sun	Cloud	Rain
weather yesterday	Sun	0.5	0.25	0.25
	Cloud	0.375	0.125	0.375
	Rain	0.125	0.625	0.375

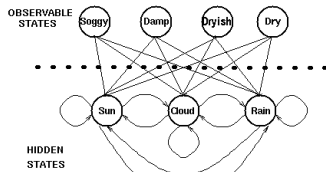
States : Three states - sunny, cloudy, rainy.

State transition matrix : The probability of the weather given the previous day's weather.

	Sun	Cloud	Rain
(1.0	0.0	0.0

Initial Distribution : Defining the probability of the system being in each of the states at time 0.

Predicting weather by observing seaweed



Hidden states : the (TRUE) states of a system that may be described by a Markov process (e.g., the weather).

Observable states : the states of the process that are 'visible' (e.g., dampness of seaweed).

Components Of HMM

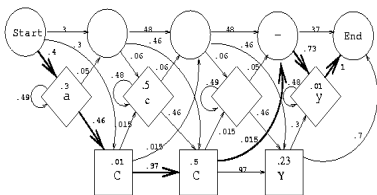
		Seaweed			
		Dry	Dryish	Damp	Soggy
weather	Sun	0.60	0.20	0.15	0.05
	Cloud	0.25	0.25	0.25	0.25
	Rain	0.05	0.10	0.35	0.50

Output matrix : containing the probability of observing a particular observation given that the model is in a particular hidden state.

Initial Distribution : contains the probability of the (hidden) model being in a particular hidden state at time t = 1. Equivalently, add a special begin state "0"

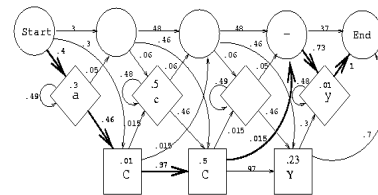
State transition matrix : holding the probability of a hidden state given the previous hidden state.

Alignment/Decoding problem



Given a sequence, what is the optimal state sequence that the HMM would use to generate it?

Scoring Problem



Given an existing HMM and observed sequence, what is the probability that the HMM can generate the sequence?

The most probable path

- $P(x, \pi)$: joint probability of sequence x and path π
 - $a_{0, \pi_1} \left(\prod_{i=1}^{L-1} e_{\pi_i}(x_i) * a_{\pi_i, \pi_{i+1}} \right) e_{\pi_L}(x_L)$
- The most probable path π for x
 - $\operatorname{argmax}_{\pi} P(x, \pi) = \operatorname{argmax}_{\pi} P(\pi|x) P(x) = \operatorname{argmax}_{\pi} P(\pi|x)$
- $P(x)$, the probability of sequence x
 - $\sum_{\pi} P(x, \pi)$

Motif finding

Bioinformatics, Fall 2011 13

Viterbi algorithm for the most probable path

- Use dynamic programming to compute the most probable path for an observed sequence $x = x_1, x_2, \dots, x_L$.
- $f(k, i)$ = probability of the most probable path for prefix x_1, x_2, \dots, x_i that ends at state k .
 - $f(0, 0) = 1, f(k, 0) = 0$ for $k > 0$ 0 is the begin state
 - $f(k, i) = e_k(x_i) * \max_{l \in S} \{f(l, i-1) * a_{l, k}\}$ Maintain backtracking pointers
- The desired value is the largest one in column L .

Motif finding

Bioinformatics, Fall 2011 14

Forward and backward probabilities

- $f(k, i)$ = probability of emitting prefix x_1, x_2, \dots, x_i and reaching state k .
- $F(x) = \sum_k f(k, L) = P(x)$
- $b(k, i)$ = probability of emitting suffix $x_{i+1}, x_{i+2}, \dots, x_L$ beginning from state k .
 - $b(k, L) = 1$
 - $b(k, i) = \sum_l b(l, i+1) * a_{k, l} * e_l(x_{i+1})$
- $B(x) = b(0, 0) = F(x) = P(x)$
- $P(k, i) = f(k, i) * b(k, i)$ = probability of generating x while passing through state k at the i th symbol

Motif finding

Bioinformatics, Fall 2011 15

Posterior decoding

- $P(x)$, the probability of sequence x of length L
 - $\sum_{\pi} P(x, \pi) = \sum_k f(k, L) = b(0, 0)$
- Posterior decoding: $P(\pi_i = k | x)$
 - = $P(x, \pi_i = k) / P(x)$
 - = $P(x_1, \dots, x_i, x_{i+1}, \dots, x_L, \pi_i = k) / P(x)$
 - = $P(x_1, \dots, x_i, \pi_i = k) P(x_{i+1}, \dots, x_L | x_1, \dots, x_i, \pi_i = k) / P(x)$
 - = $P(x_1, \dots, x_i, \pi_i = k) P(x_{i+1}, \dots, x_L | \pi_i = k) / P(x)$
 - = $f(k, i) b(k, i) / P(x)$

Motif finding

Bioinformatics, Fall 2011 16

Parameter estimation for HMM

- Given a set of sequences, assign the state transition and emission probabilities that best characterize the set, i.e., build the best *model*.
- If the state sequences are known then compute
 - $A(k, l)$: the number of transitions from state k to state l
 - $E(k, c)$: the emission probability of symbol c in state k
 - use maximum likelihood estimator
 - add pseudocounts and normalize
- If the state sequences are not known (the common case)
 - Use EM algorithm, called Baum-Welch
 - The missing parameter is the state to which a particular observation belongs.

Motif finding

Bioinformatics, Fall 2011 17

Baum-Welch algorithm

- Repeat
 - For each sequence x ,
 - Compute $f(k, i)$ and $b(k, i)$
 - Add the contribution of sequence x to A and E
 - $A(k, l) += (1/P(x)) \sum_i f(k, i) * a_{k, l} * e_l(x_{i+1}) * b(l, i+1)$
 - $E(k, c) += (1/P(x)) \sum_{x_i=c} f(k, i) * b(k, i)$
 - Normalize and use the new parameters
- Until convergence

Expected number of times the transition (k, l) is taken

Expected number of times symbol c is emitted in state k

Motif finding

Bioinformatics, Fall 2011 18

HMMs in Biology

- Gene finding and prediction
- Profile HMM

Motif findline

Bioinformatics, Fall 2011 19

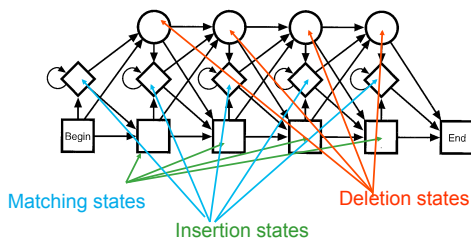
Profile HMM

- Model an alignment of sequences (may include gaps)
- Extension of PSSM to gaps
- Score a given sequence with the model to understand similarity
- Include insert and delete states in each column

Motif findline

Bioinformatics, Fall 2011 20

Profile HMMs

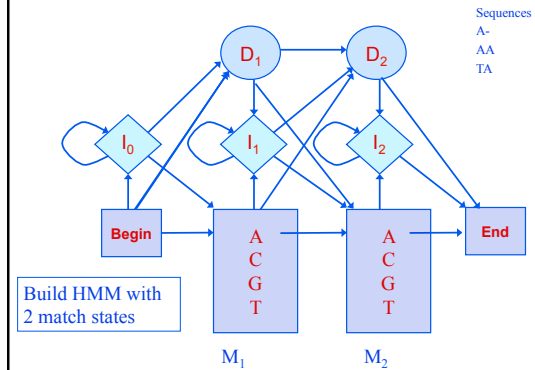


No of matching states = average sequence length in the family
Can align against a profile HMM

Motif findline

Bioinformatics, Fall 2011 21

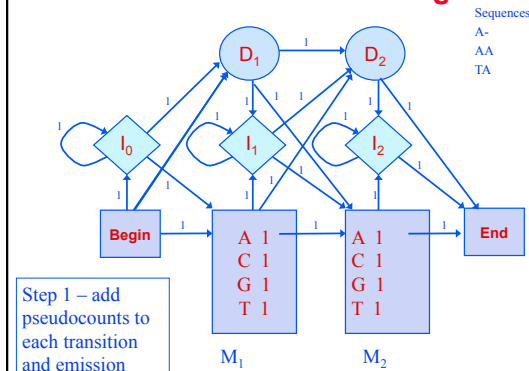
Profile HMM Training



Motif findline

Bioinformatics, Fall 2011 22

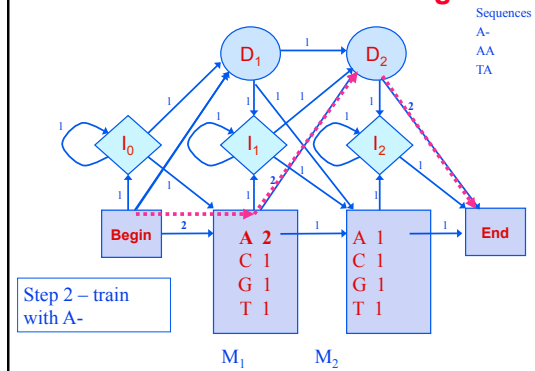
Profile HMM Training



Motif findline

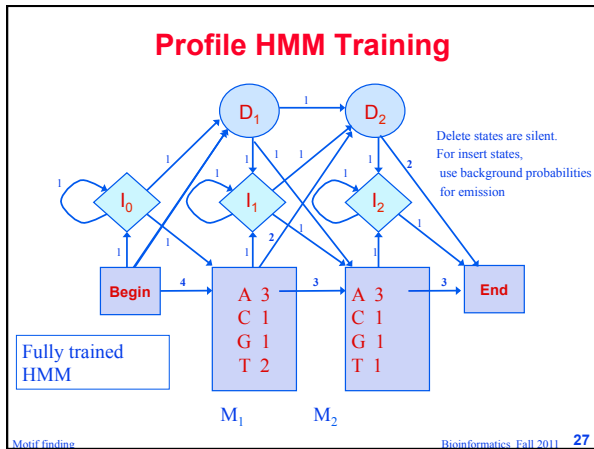
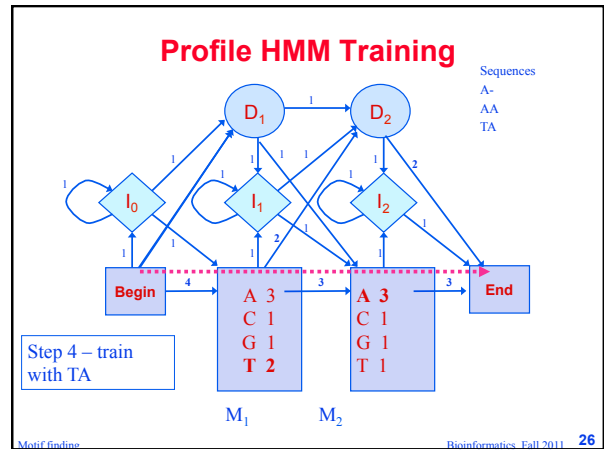
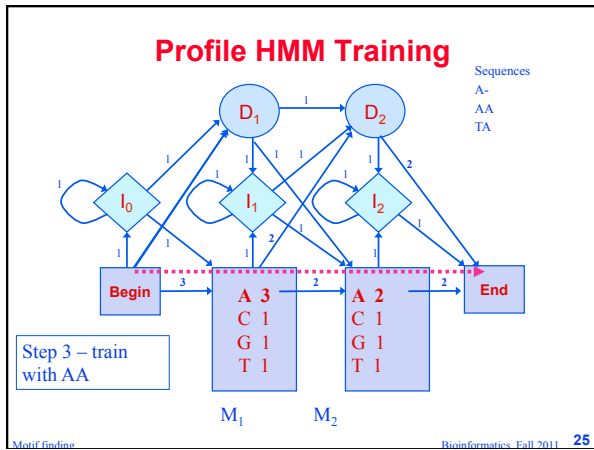
Bioinformatics, Fall 2011 23

Profile HMM Training



Motif findline

Bioinformatics, Fall 2011 24



- ### HMM Programs
- HMMER (“hammer”)
 - profile-HMM package for protein sequences
 - Input is a multiple alignment, and the tool produces a profile-HMM
 - Use HMM to check for membership to a particular family
 - HMMER is freely available for download:
 - <http://hmmer.janelia.org/>
- Motif findline Bioinformatics, Fall 2011 28

- ### HMM Programs
- SAM
 - “Sequence Alignment and Modeling System”
 - profile-HMM package used for protein sequences
 - similar to HMMER in functionality
 - <http://www.cse.ucsc.edu/research/compbio/sam.html>
- Motif findline Bioinformatics, Fall 2011 29

- ### HMM Programs
- Meta-Meme
 - Input is a set of similar protein sequences and a set of motifs (discovered by MEME) as PSSM
 - Output is a motif-based HMM (more sensitive to the motifs)
 - <http://metameme.sdsc.edu/>
- Motif findline Bioinformatics, Fall 2011 30

PROSITE

- A manually created collection of regular expressions (and profiles) associated with different protein families/functions.
- A description of sequence motifs associated with function for elucidating function of new sequences
- <http://us.expasy.org/prosite/>
- [RK]-G-{EDRKHPCG}-[AGSCI]-[FY]-[LIVA]-x-[FYM]
 - []: +
 - {}: not
 - x: don't care
 - A(m): A^m
 - A(m-n): between m and n occurrences of A

Motif finding

Bioinformatics, Fall 2011 31

Pfam: Protein Family Classification

- Large collection of multiple sequence alignments and Hidden Markov Models
- Covers many common protein domains and families
 - 10,140 protein families (release 23.0)
 - Pfam-A and Pfam-B
 - 73% of known protein sequences have a match within Pfam-A
 - Clans are used to group related families
- <http://pfam.janelia.org/>

Motif finding

Bioinformatics, Fall 2011 32

BLOCKS database

- Ungapped multiple alignments of segments of related protein sequences that correspond to the most conserved regions of proteins
- Understand the sequence variability in a particular motif. Can use to create scoring matrices.
- Search a protein sequence against the database.
- <http://blocks.fhcrc.org/>

Motif finding

Bioinformatics, Fall 2011 33

InterPro

- Integrated resource of protein families, domains, and functional sites in a single, comprehensive resource. Combines information from
 - PROSITE
 - Pfam
 - PRINTS
 - ProDom
 - SMART
 - TIGRFAMs, and others
- <http://www.ebi.ac.uk/interpro/>

Motif finding

Bioinformatics, Fall 2011 34