Exercises 2: Hidden Markov Models

Protein Structure

Proteins = large molecules that are among the most important components in cells of living organisms.

They are build from 20 smaller molecules, the amino acids:

A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y

Protein Structure

A protein can be defined by the linear sequence of its amino acids.

E.g.: >gi|10946489|gb|AAG24916.1|AF308694_1 ELAC2 [Gorilla gorilla]

MWALCSLLRSAAGRTMSQGRTISQAPARRERPRKDPLRHLRTREKRGPSGCSGGPNTVY
LQVVAAGSRDSGAALYVFSEFNRYLFNCGEGVQRLMQEHKLKVVRLDNIFLTRMHWSNV
GGLSGMILTLKETGLPKCVLSGPPQLEKYLEAIKIFSGPLKGIELAVRPHSAPEYEDETMTVY
QIPIHSEQRRGRHQPWQSPERPLSRLSPERSSDSESNENEPHLPHGVSQRRGVRDSSLV
VAFICKLHLKRGNFLVLKAKEMGLPVGTAAIAPIIAAVKDGKSITHEGREILAEELCTPPDPG
AAFVVVECPDESFIQPICENATFQRYQGKADAPVALVVHMAPESVLVDSRYQQWMERFG
PDTQHLVLNENCASVHNLRSHKIQTQLNLIHPDIFPLLTSFPCKKEGPTLSVPMVQGECLLK
YQLRPRREWQRDAIITCNPEEFIVEALQLPNFQQSVQEYRRSVQDVPAPAEKRSQYPEIIFL
GTGSAIPMKIRNVSATLVNISPDTSLLLDCGEGTFGQLCRHYGDQVDRVLGTLAAVFVSH
LHADHHTGLLNILLQREQALASLGKPLHPLLVVAPSQLKAWLQQYHNQCQEVLHHISMIP
AKCLQEGAEISSPAVERLISSLLRTCDLEEFQTCLVRHCKHAFGCALVHTSGWKVVYSGD
TMPCEALVRMGKDATLLIHEATLEDGLEEEAVEKTHSTTSQAISVGMRMNAEFIMLNHFS
QRYAKVPLFSPNFNEKVGVAFDHMKVCFGDFPTMPKLIPPLKALFAGDIEEMEERREKREL
RQVRAALLSGELAGGLEDGEPQQKRAHTEEPQAKKVRAQ

➡ the Primary structure

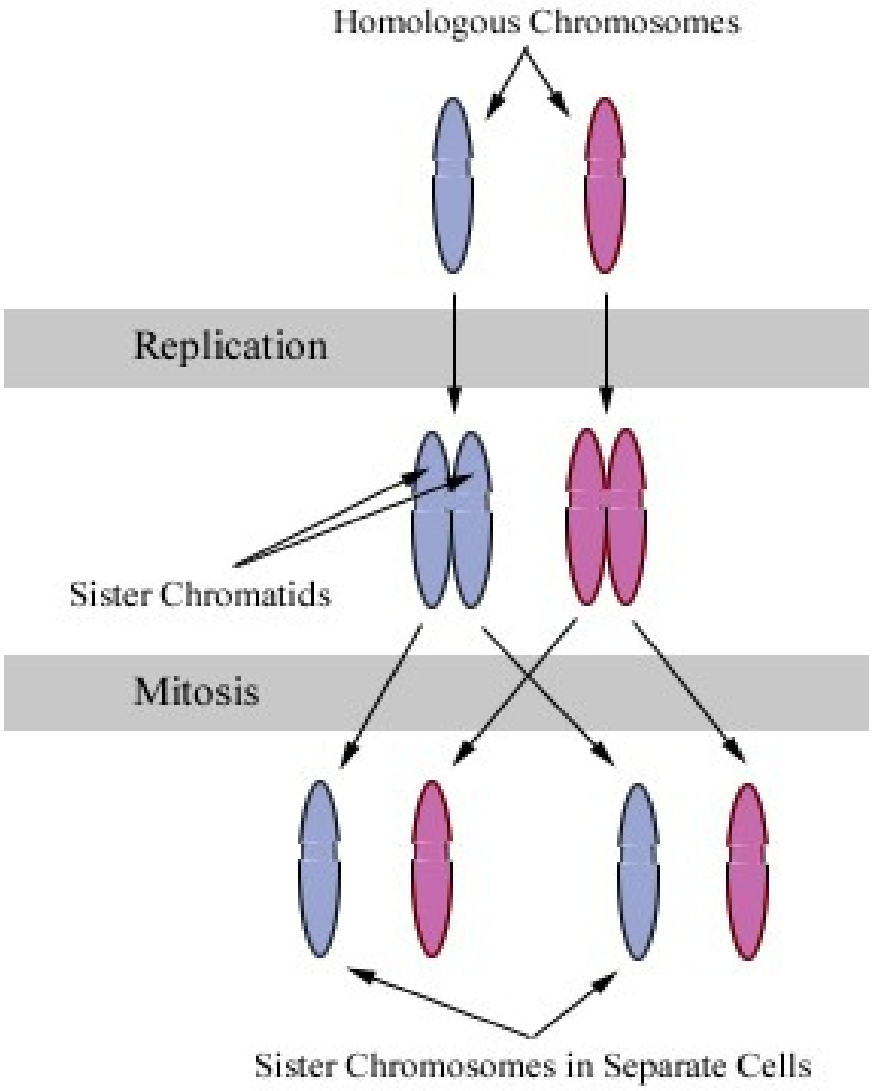
Protein Structure

Mitosis or cell duplication:

In a diploid eukaryotic cell, there are two versions of each chromosome, one from the mother and another from the father. The two corresponding chromosomes are called homologous chromosomes.

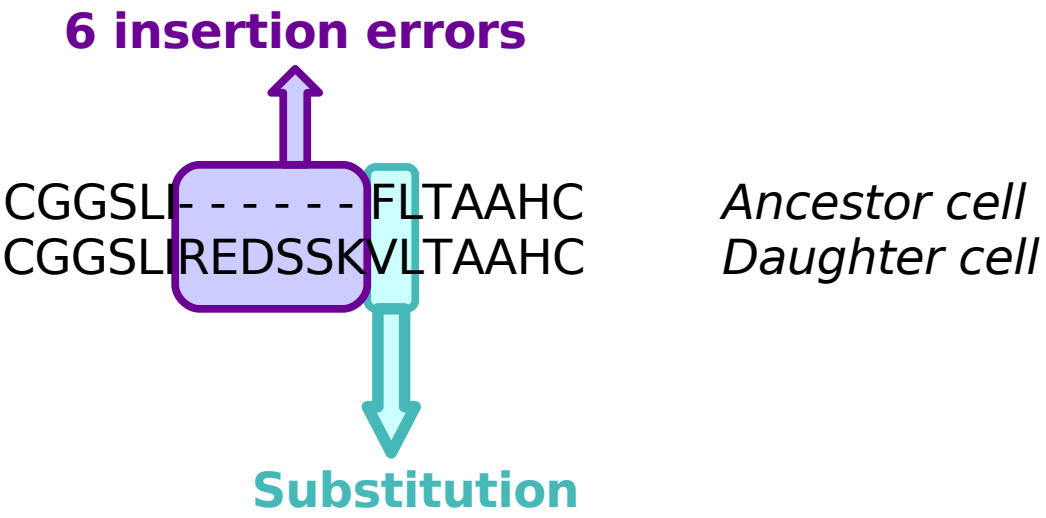
When DNA is replicated, each chromosome will make an identical copy of itself. The copies are called sister chromatids. After separation, however, each sister chromatid is considered a full-fledged chromosome by itself. The two copies of the original chromosome are then called sister chromosomes.

Mitosis allocates one copy, and only one copy, of each sister chromosome to a daughter cell. Consider the diagram above, which traces the distribution of chromosomes during mitosis. The blue and red chromosomes are homologous chromosomes. After DNA replication during S phase, each homologous chromosome contains two sister chromatids. After mitosis, the sister chromatids become sister chromosomes and part ways, going to separate daughter cells. Homologous chromosomes are therefore kept together, resulting in the complete transfer of the parent's genome.



Protein Motifs

Possible errors in the copying process:



Proteins of a common ancestor are not exactly alike, but share many similarities

Statistical Profiles

Goal: To build a statistical model of a protein family based on the structural similarities

Family Members

1 2 3 4 5

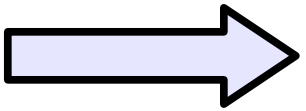
C C G T L

C G H S V

G C G S L

C G G T L

C C G S S



Position	1	2	3	4	5
Prob(C)	0.8	0.6			
Prob(G)	0.2	0.4	0.8		
Prob(H)			0.2		
Prob(S)				0.6	0.2
Prob(T)				0.4	
Prob(L)					0.6
Prob(V)					0.2

Probability of a sequence:
E.g.: CGGSV

Prob = 0.8*0.4*0.8*0.6*0.2=0.031

Or a score in log-scale:

Score = ln(0.8)+ln(0.4)+ln(0.8)+ln(0.6)+ln(0.2)= - 3.48

Statistical profiles

Shortcomings of this simple model...

- Members of protein families have different lengths
- Score penalties for insertions, deletions, substitutions?
- …

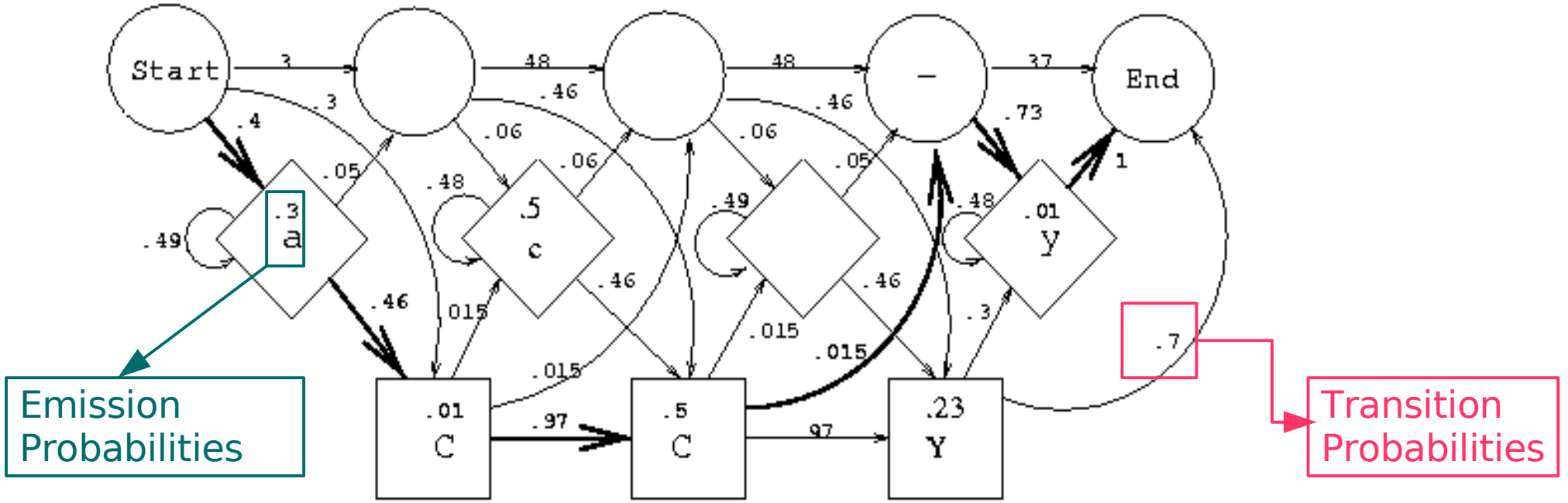
⇒ Hidden Markov Models

Hidden Markov Models

A more dynamical statistical profile, build on a training set of related proteins.

A HMM generates a protein sequence as it progresses through the different states

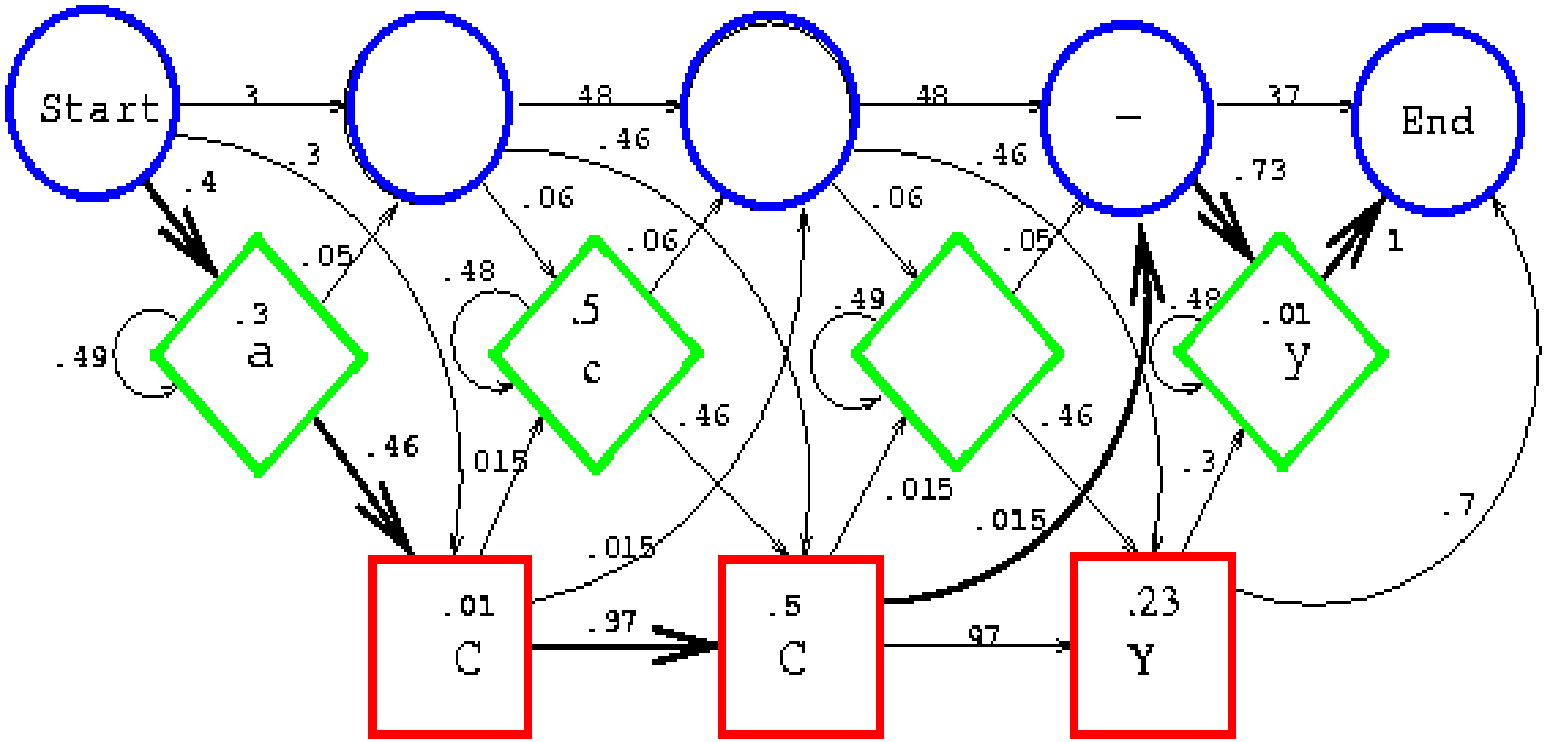
E.g. The topology for a HMM for the protein ACCY



Hidden Markov Models

There are three types of states drawn:

Delete states:
model the deletions



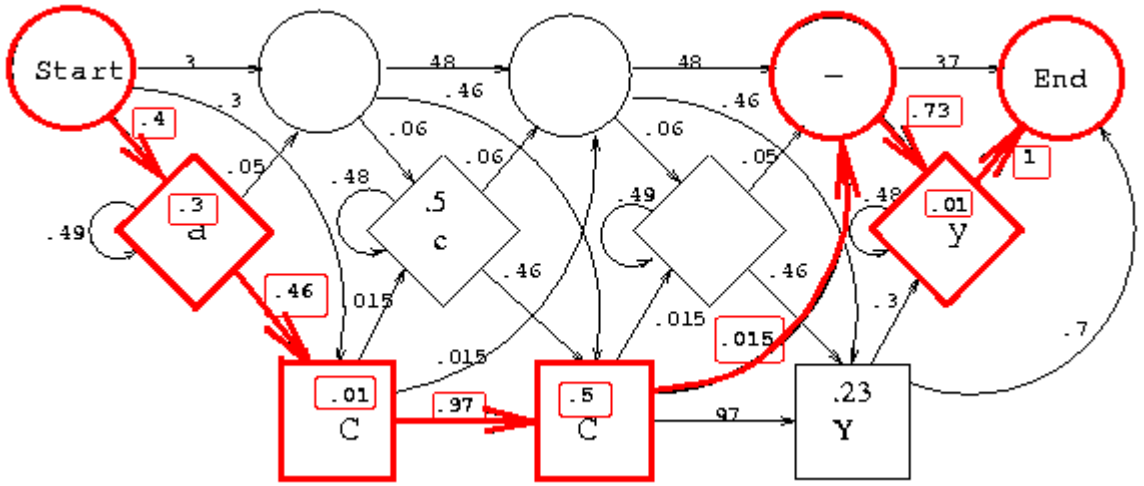
Insert states:
amino acids resulting from insertion - insertions of any length are allowed due to the self-loops

Match states:
Those amino acids are the same as those in the common ancestor or, they are the result of substitution

Scoring a sequence with a HMM

Any sequence can be represented as a path through the model.

Compute the probability for a sequence by multiplying the emission and transition probabilities



The Prob for ACCY:
0.4*0.3*0.46*0.01*0.97*0.5*
0.015*0.73*0.01*1
$= 2.93 \times 10^{-8}$

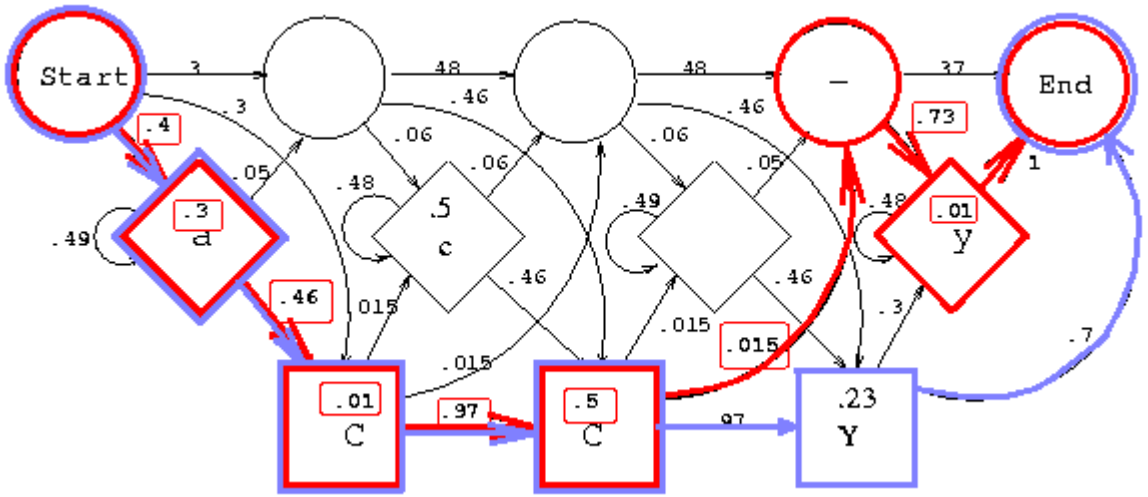
Or

the Score of ACCY:
ln(0.4)+ln(0.3)+ln(0.46)+ln(0.
01)+ln(0.97)+ln(0.5)+ln(0.01
5)+ln(0.73)+ln(0.01)+ln(1)
= -17.35

Scoring a sequence with a HMM

Any sequence can be represented as a path through the model.

Compute the probability for a sequence by multiplying the emission and transition probabilities



The *Prob* for ACCY:
0.4*0.3*0.46*0.01*0.97*0.5*
0.015*0.73*0.01*1
= 2.93 x 10^{-8}

Or

the *Score* of ACCY:
ln(0.4)+ln(0.3)+ln(0.46)+ln(0
.01)+ln(0.97)+ln(0.5)+ln(0.0
15)+ln(0.73)+ln(0.01)+ln(1)
= - 17.35

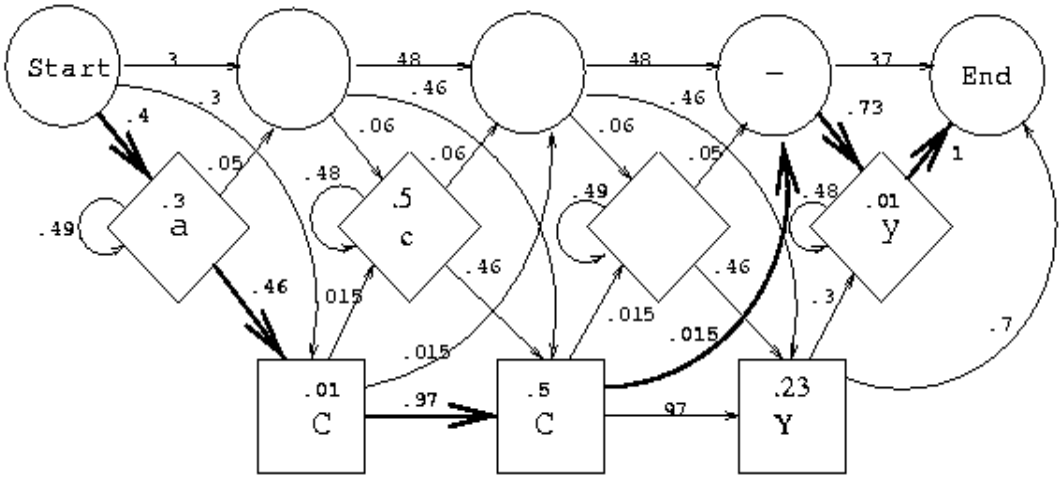
The score is easy to compute for a single, given path. But many path can generate the same sequence...

The correct probability is therefore the sum of the probabilities over all possible state paths

Scoring a sequence with a HMM

Viterbi algorithm:

to compute the most likely path

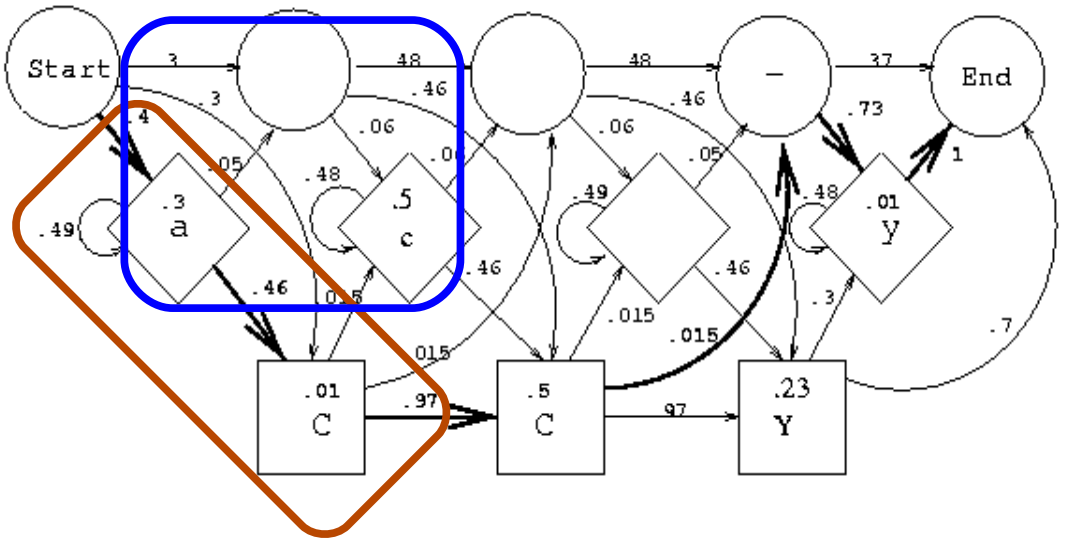


1. Compute the prob of A generated by status I0

	I0	I1	M1	I2	M2	I3	M3
A	.12						
	.4*.3						
C							
C							
Y							

Scoring a sequence with a HMM

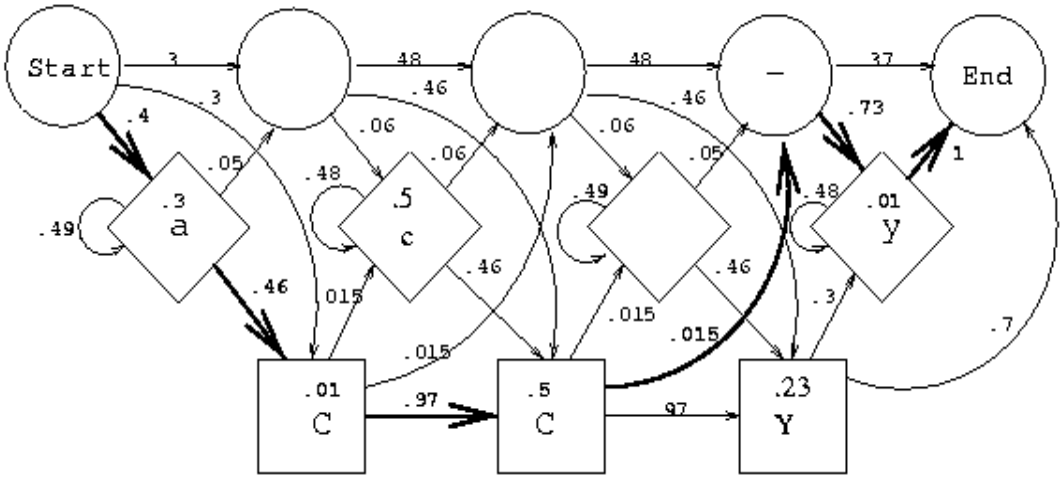
1. Compute the prob of A generated by status I0
2. Compute the prob that C is inserted in status I1 or in status M1



	I0	I1	M1	I2	M2	I3	M3
A	.12						
	.4*.3						
C		0.0015	0.005				
		0.05*0.06*0.5	0.46*0.01				
C							
Y							

Scoring a sequence with a HMM

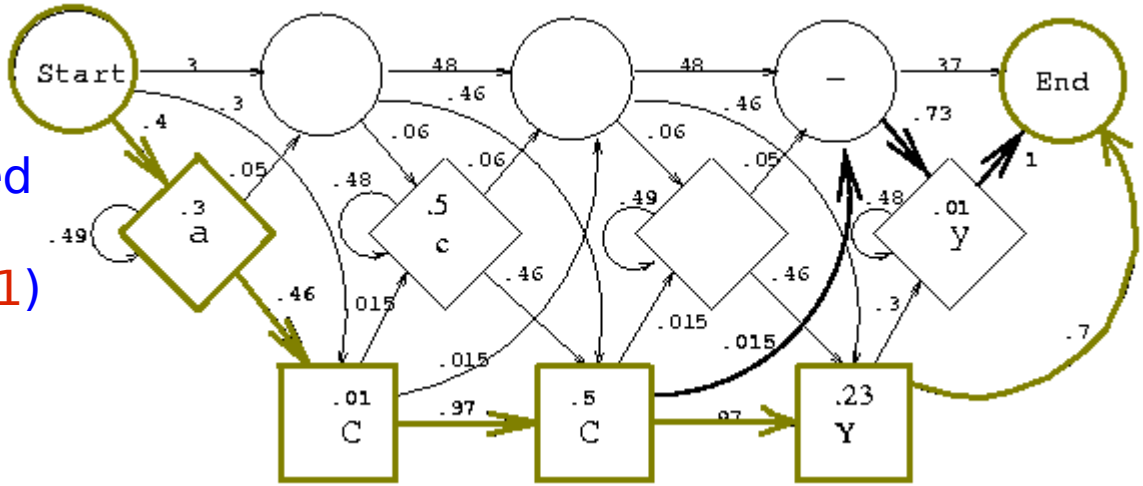
1. Compute the prob of A generated by status I0
2. Compute the prob that C is inserted in status I1 or in status M1
3. Compute the maximum max(I1, M1) and set a pointer back from the max to state I0



	I0	I1	M1	I2	M2	I3	M3
A	.12						
	.4*.3						
C		0.0015	0.005				
		0.05*0.06*0.5	0.46*0.01				
C							
Y							

Scoring a sequence with a HMM

1. Compute the prob of A generated by status I0
2. Compute the prob that C is inserted in status I1 or in status M1
3. Compute the maximum max(I1, M1) and set a pointer back from the max to state I0
4. Repeat steps 2 – 4 until matrix is filled

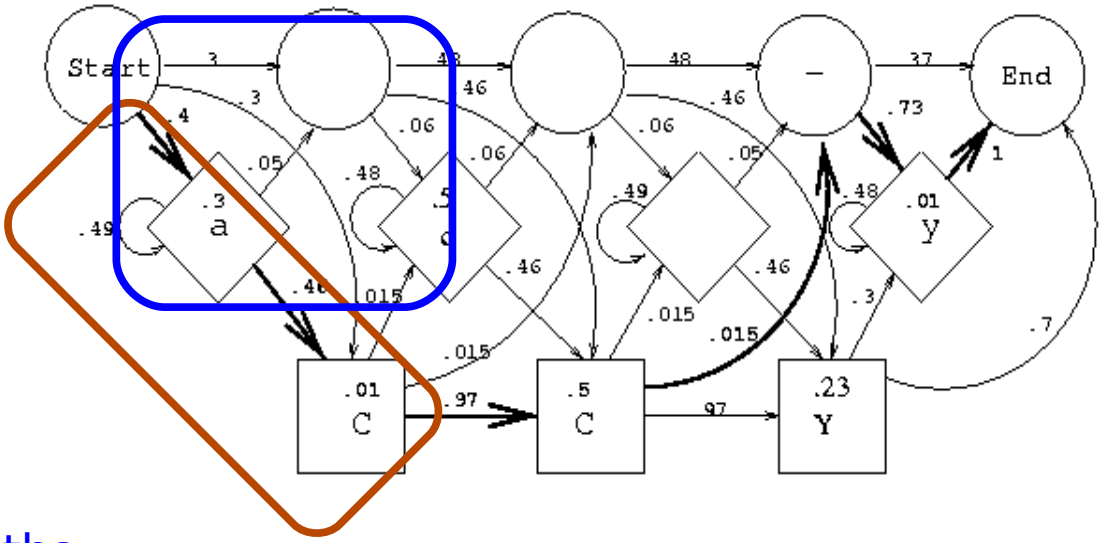


	I0	I1	M1	I2	M2	I3	M3
A	.12						
	.4*.3						
C		0.0015	0.005				
		0.05*0.06*0.5	0.46*0.01				
C				0	0.49		
					0.97*0.5		
Y						0.0001	0.22
						0.015*0.73*0.01	0.97*0.23

Scoring a sequence with a HMM

Forward algorithm:

to calculate the sum over all paths



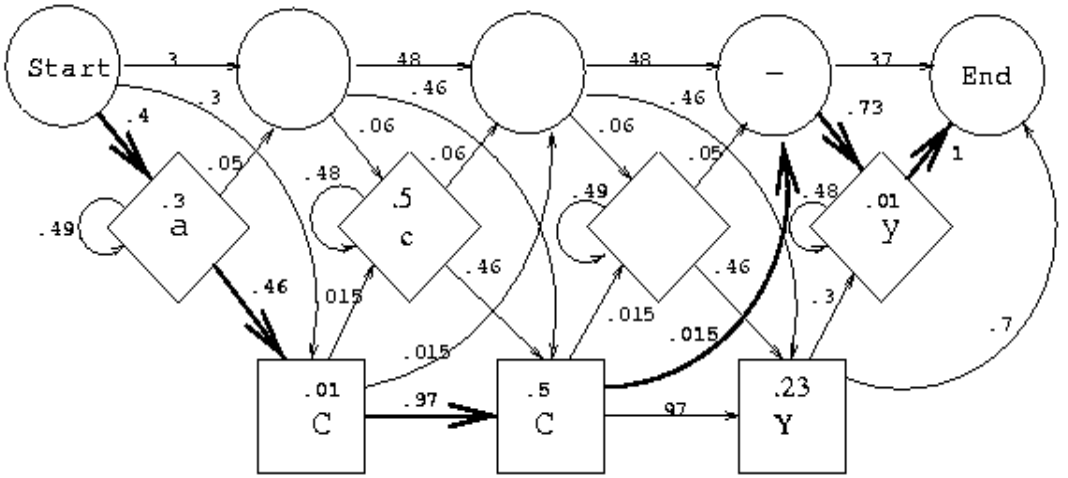
First step is the same as for the Viterbi algorithm

2. Compute the sum of the different paths to the next state

	I0	I1	M1	I2	M2	I3	M3	End
A	.12							
	.4*.3							
C		0.00018	0.000552					
		0.12*0.05*0.06*0.5	.12*0.46*0.01					
C								
Y								

Scoring a sequence with a HMM

1. Compute the prob of A generated by status I0
2. Compute the sum of the different paths to the next state

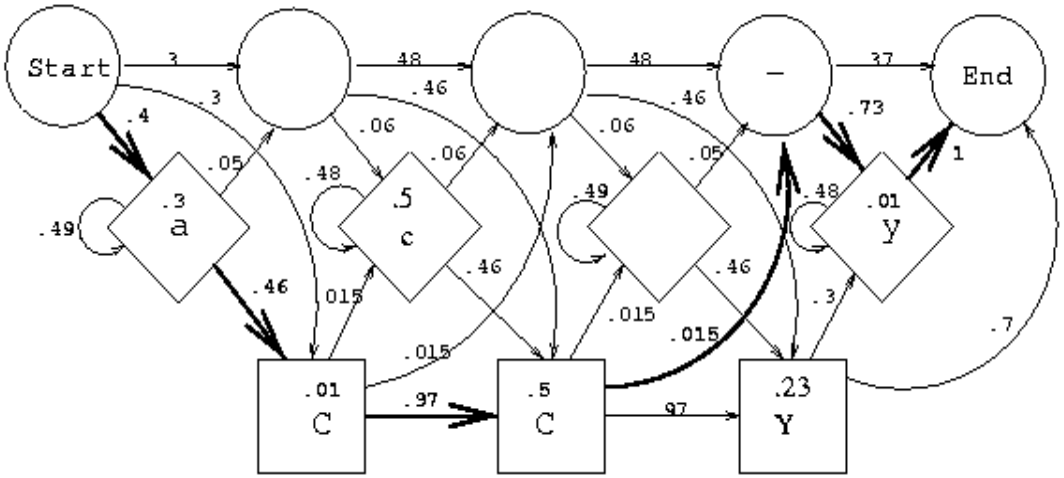


Score = (0.000552*0.97+0.00018*0.46)*0.5

	I0	I1	M1	I2	M2	I3	M3	End
A	.12	0	0	0	0	0	0	
	.4*.3							
C	0	0.00018	0.000552	0	0	0	0	
C	0	0	0	0	0.000309	0	0	
Y	0	0	0	0	0	0	0	

Scoring a sequence with a HMM

1. Compute the prob of A generated by status I0
2. Compute the prob that C is inserted in status I1 or in status M1
3. Compute the sum of the different paths to the next state
4. Repeat steps 2- 4



	I0	I1	M1	I2	M2	I3	M3	End
A	.12	0	0	0	0	0	0	0
	.4*.3							
C	0	0.00018	0.000552	0	0	0	0	0
C	0	0	0	0	0.000309	0	0	0
					(.005*.97+.0015*.46)*.5			
Y	0	0	0	0	0	$3.38\ 10^{-8}$	$6.89\ 10^{-5}$	
						0.000309*0.015*0.73*0.01	0.006*0.23*0.97	

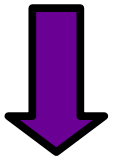
$4.83\ 10^{-5}$

$3.38\ 10^{-8}*1+6.89\ 10^{-5}*.7$

Probability of the sequence

Building a HMM

Set of training sequences



Create an HMM?

– Emission distribution at each state?

– All transition probabilities?

Building a HMM

- If state paths for all training sequences are known:

  use their expected values:

$$\frac{\textit{number transitions / emissions}}{\textit{total number transition / emission}}$$

- If state paths are unknown:

  Find model parameters that maximize the probability of all sequences

  Iterative procedure where parameters are re-estimated for each training set by computing a score against the previous set of model

→ Viterbi Algorithm or

Baum – Welch Algorithm

Building a HMM

Viterbi algorithm

1. Starts with a guess for an initial model
2. Computes for each training sequence the best path
3. Computes a new set of emission/transition probabilities as in the case where the paths are known
4. Updated model replaces initial model and steps are repeated until model convergence

Building a HMM

Baum-Welch algorithm

1. Starts with a guess for an initial model
2. Calculates for each training sequence the score over all possible paths (~Forward algorithm)
3. Computes a new set of emission/transition probabilities as in the case where the paths are known
4. Updated model replaces initial model and steps are repeated until model convergence

Hidden Markov Models

Remarks:

- Only a local maximum is found
  - One can start with different initial models. Compute the probability of the model and select the one with the highest probability
  - To add noise/random data at each iteration. Less and less noise at each iteration step
- Overfitting...

Overfitting and regularisation

Overfitting

C	G	S	L	L	N	A	N	-	-	T	V	L	T
C	G	S	L	I	D	N	K	-	G	W	I	L	T
C	G	S	L	I	R	Q	G	-	-	W	V	M	T
C	G	S	L	I	R	E	D	S	S	F	V	L	T

Only sequences starting with `C' will be recognized.
Probability of all other amino acids is 0.

Pseudocounts will solve this...

Overfitting and regularisation

Pseudocounts

C	G	S	L	L	N	A	N	-	-	T	V	L	T
C	G	S	L	I	D	N	K	-	G	W	I	L	T
C	G	S	L	I	R	Q	G	-	-	W	V	M	T
C	G	S	L	I	R	E	D	S	S	F	V	L	T

Add a fake count for all amino acids.

E.g. a pseudocount of 1 gives for the first column:

Observed counts in column 1

$$\text{Prob(A)} = \frac{0+1}{4+20} = \frac{1}{24}$$

$$\text{Prob(C)} = \frac{4+1}{4+20} = \frac{5}{24}$$

Pseudocounts over all amino acids

Applications of HMMs

Classifying sequences in a data base:

- Build an HMM with training set of known members of the class
- Compute scores of all sequences in the db
- Choose a threshold to separate the class members from the other sequences.

But: scores are length-dependent

One possible solution:

- Score sequence as:

$$score(s) = \log \frac{P(s|M)}{P(s|R)} = \log \left(\frac{\text{Prob sequence generated by HMM}}{\text{Prob sequence in random model}} \right)$$

Applications of HMMs

Creating multiple alignments

- Find the Viterbi algorithm to find most likely path through HMM for each sequence
- Each match state corresponds to a column in the multiple alignment
- A delete state is indicated with –