Introduction to Hidden Markov Models (HMMs)

Lecture 3

Reichman University Efi Arazi School of Computer Science

1/2/2024

Algorithms in Computational Biology CS3571 IDC 2023/4

Lecture overview

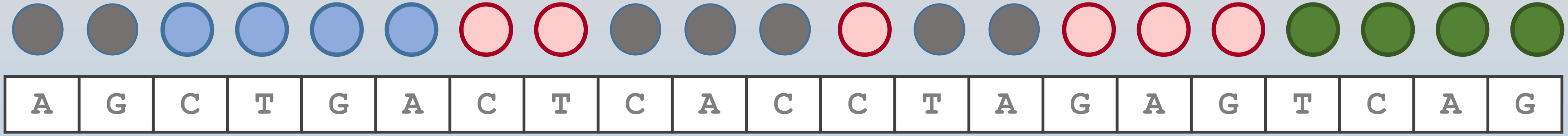
- Sequence annotation using HMMs
- Decoding problems in probabilistic models and HMMs
- Recap of multi-variate probability distributions
- Viterbi algorithm for most probable path

Lecture overview

- Sequence annotation using HMMs ←
- Decoding problems in probabilistic models and HMMs
- Recap of multi-variate probability distributions
- Viterbi algorithm for most probable path

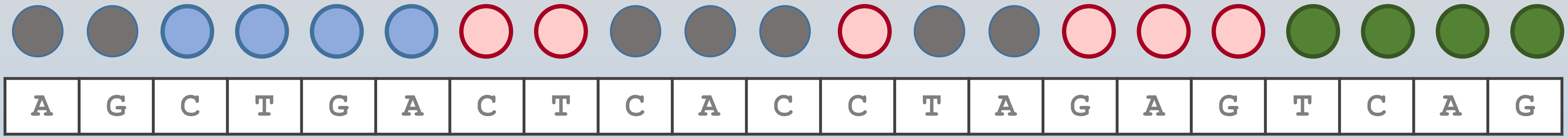
Sequence annotation

Many problems in sequence analysis can be casted as annotation problems, which aim to mark a given sequence of symbols with meaningful labels.



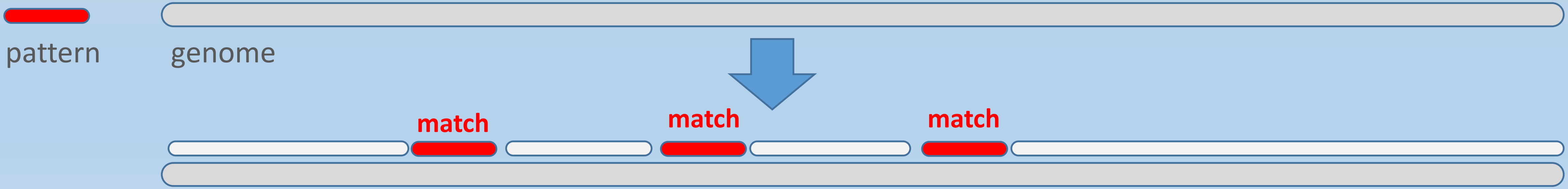
Sequence annotation

Many problems in sequence analysis can be casted as annotation problems, which aim to mark a given sequence of symbols with meaningful labels.



Examples:

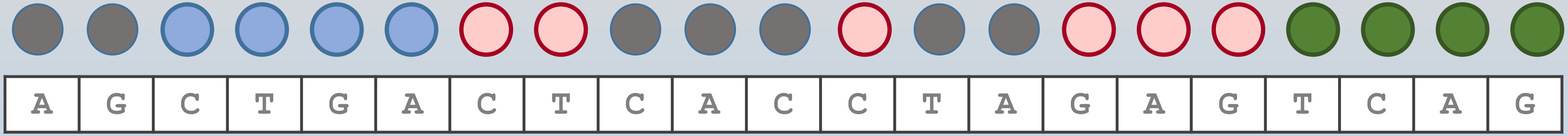
- Finding sequence patterns and local alignment



Sequence annotation

Reichman University
Efi Arazi School of Computer Science

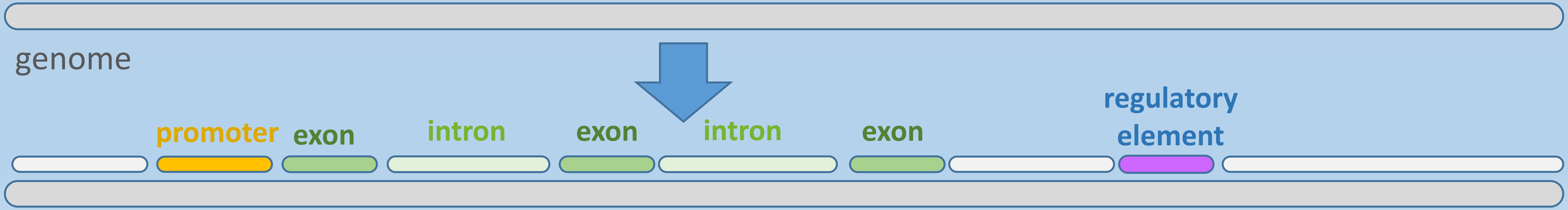
Many problems in sequence analysis can be casted as annotation problems, which aim to mark a given sequence of symbols with meaningful labels.



Examples:

- Finding genomic segments with specific features

Gene model



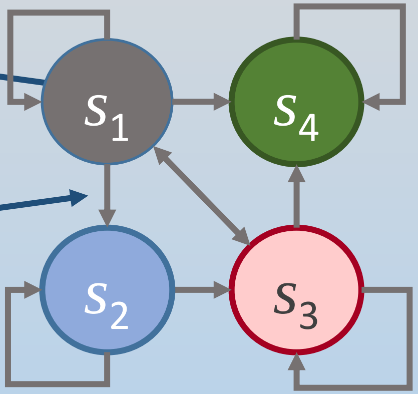
Sequence annotation using hidden Markov models (HMMs)

Annotation problems can be solved efficiently if you consider the labels as hidden (unobserved) states in a stochastic finite state machine (FSM).

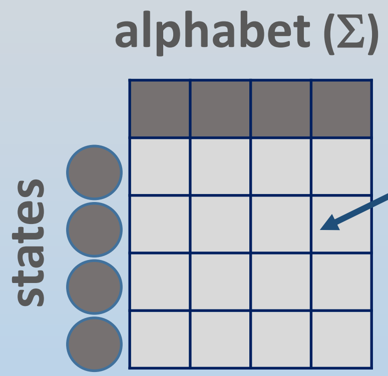
The stochastic FSM:

Hidden labels / states

Directed edges weighted with probabilities (sum of outgoing weights is 1 for each state)

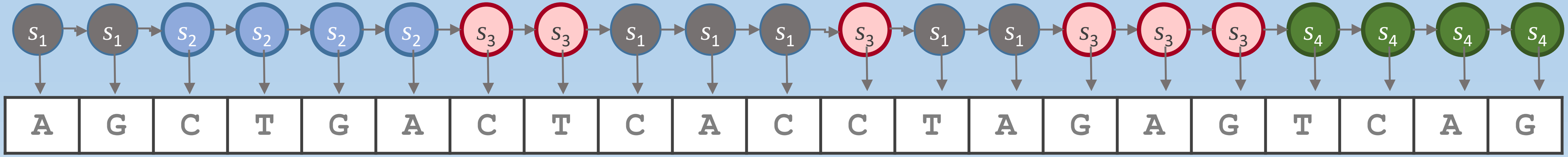


Emission probabilities:



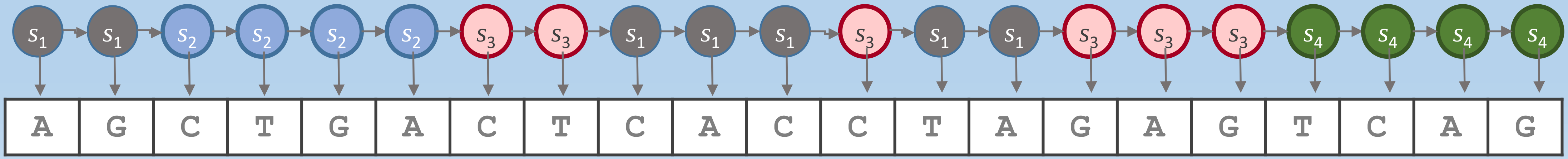
Each state has a different distribution over possible emissions

Annotation as a Markov chain over hidden states:



Hidden Markov models - history

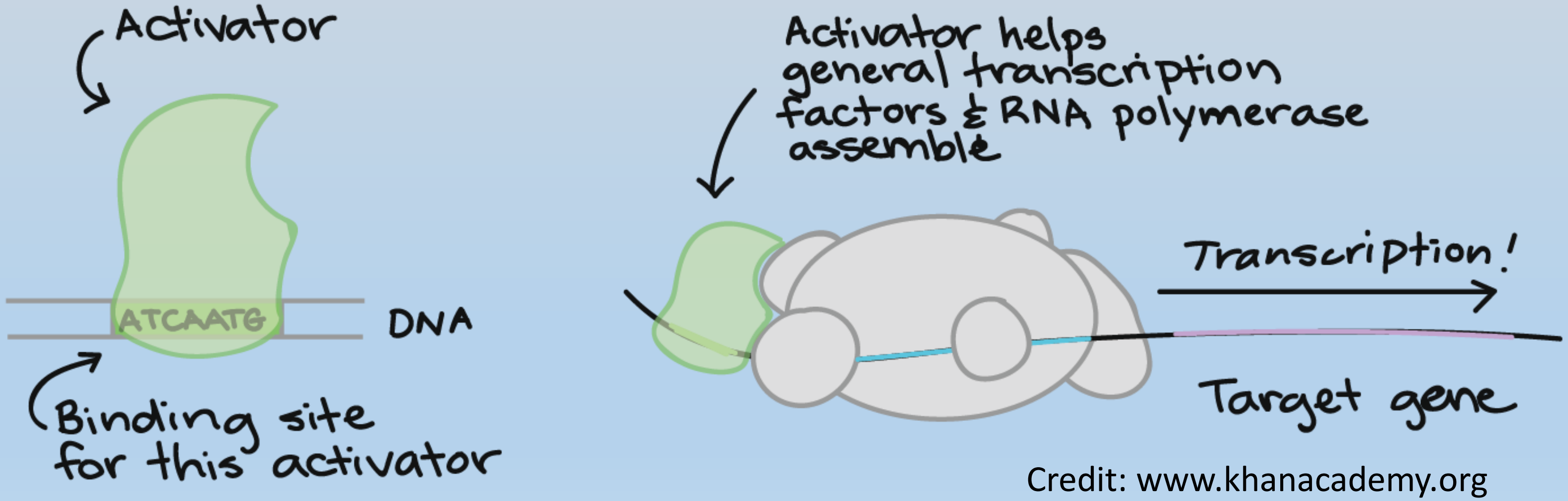
- Work on non-linear filtering and smoothing done in the late 1950’s by Ruslan Stratonovich laid out much of the techniques used in HMM decoding
- HMMs were formally defined and algorithms proposed in the 1960’s, mostly by Leonard Baum
- Viterbi’s algorithm for most probable path proposed in 1967, three years before Needleman-Wunsch’s algorithm for most probable alignment
- Widely used in applications in speech/writing recognition and signal analysis



Examples of HMMs – sequence motifs

What are sequence motifs?

- Transcription factors are proteins that bind to DNA in specific sequence patterns (motifs), and regulate transcription
- The binding motif is typically represented by a probabilistic sequence pattern, e.g.

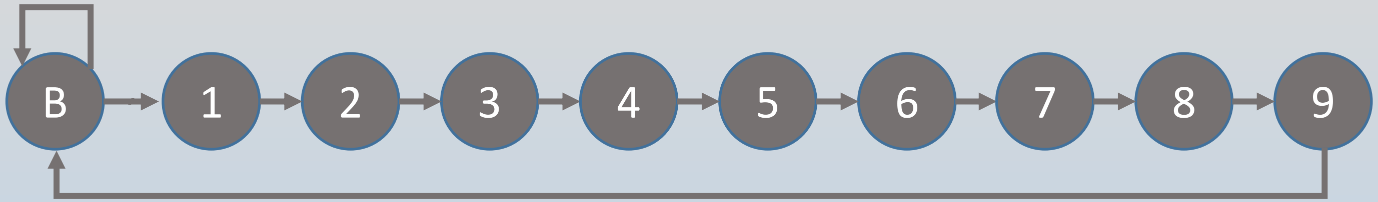


Credit: www.khanacademy.org

Examples of HMMs – sequence motifs

Reichman University Efi Arazi School of Computer Science

Sequence motif HMM

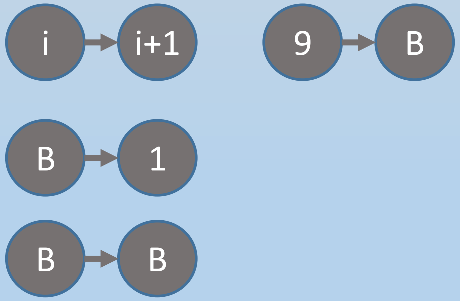


Emission probabilities:

	1	2	3	4	5	6	7	8	9	B
A	**0.33**	0.01	0.01	**0.97**	0.01	0.01	0.01	**0.97**	0.01	q_A
G	**0.33**	0.01	**0.97**	0.01	**0.49**	0.01	0.01	0.01	0.01	q_G
C	**0.33**	0.01	0.01	0.01	**0.49**	0.01	**0.97**	0.01	**0.49**	q_C
T	0.01	**0.97**	0.01	0.01	0.01	**0.97**	0.01	0.01	**0.49**	q_T

q_X – background probability for base X

Transition probabilities:



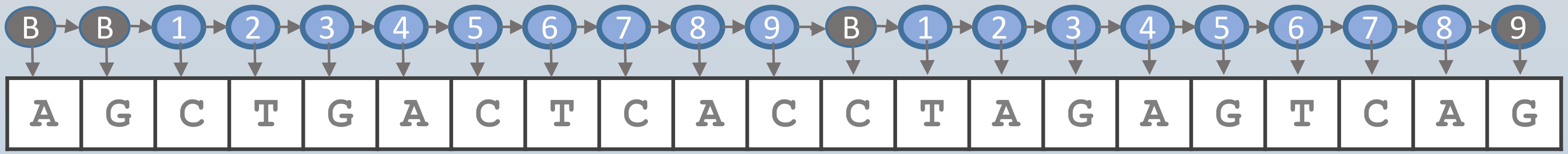
$p = 1$

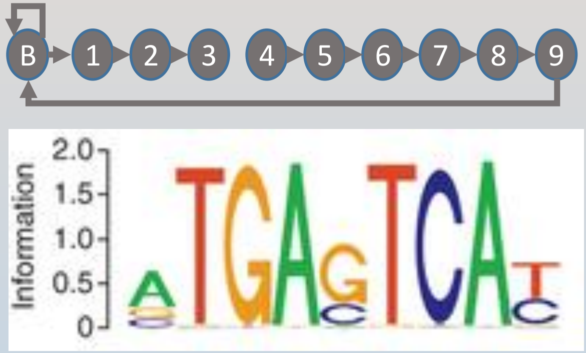
$p = p_{start}$

$p = 1 - p_{start}$

Examples of HMMs – sequence motifs

Detecting sequence motifs using the HMM:





Examples of HMMs – sequence alignment

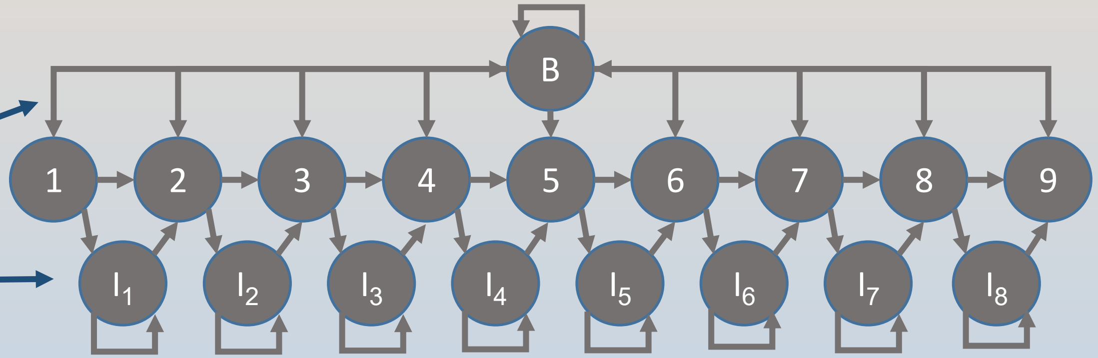
Reichman University
Efi Arazi School of Computer Science

HMM for local alignment with insertions

T = **AGGCCTTGC**

start / end alignment at any point in T

allow insertions (gaps in T)



Emission probabilities:

	1	2	3	4	5	6	7	8	9	B	I_i
A	p_M	p_{MM}	p_{MM}	p_{MM}	p_{MM}	p_{MM}	p_{MM}	p_{MM}	p_{MM}	q_A	q_A
G	p_{MM}	p_M	p_M	p_{MM}	p_{MM}	p_{MM}	p_{MM}	p_M	p_{MM}	q_G	q_G
C	p_{MM}	p_{MM}	p_{MM}	p_M	p_M	p_{MM}	p_{MM}	p_{MM}	p_M	q_C	q_C
T	p_{MM}	p_{MM}	p_{MM}	p_{MM}	p_{MM}	p_M	p_M	p_{MM}	p_{MM}	q_T	q_T

q_X – background probability for base X

p_M – probability of a match p_{MM} – probability of a mismatch $(1-p_M)/3$

Transition probabilities:

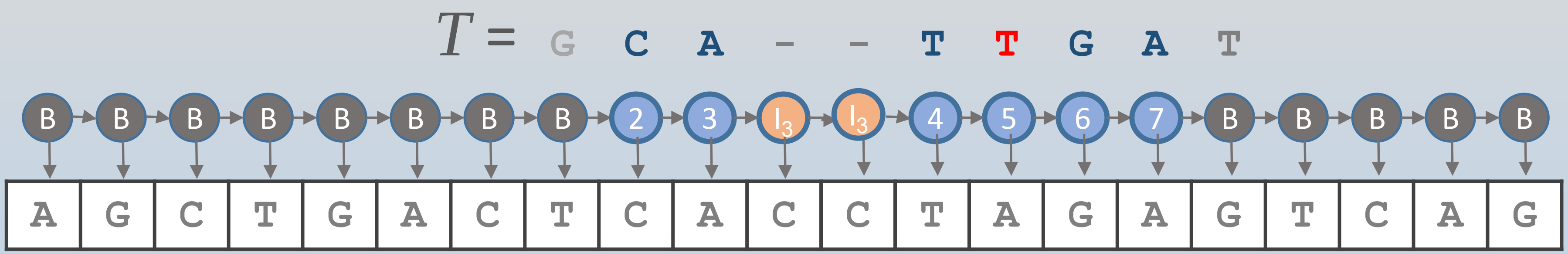
- B → i: $p = p_{start}$
- B → B: $p = 1-9p_{start}$
- I_i → I_i: $p = p_{elong}$
- I_i → i+1: $p = 1-p_{elong}$
- i → I_i: $p = p_{gap}$
- i → B: $p = p_{end}$
- i → i+1: $p = 1-p_{gap}-p_{end}$

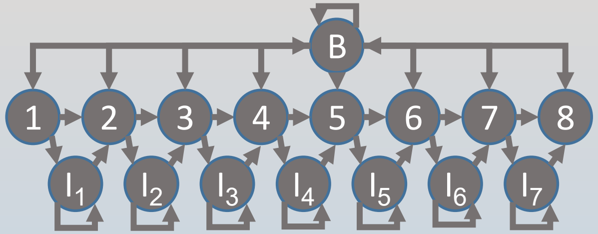
(deletions (gaps in S) can be accommodated by adding an other series of states $D_1 \ldots D_8$)

Examples of HMMs – sequence alignment

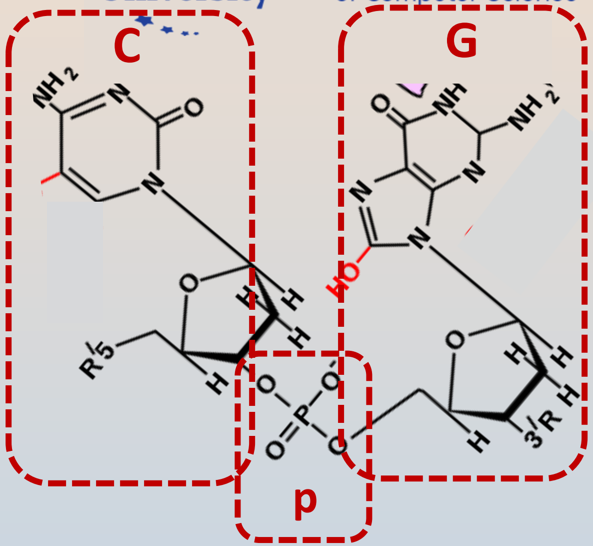
Reichman University Efi Arazi School of Computer Science

Inferring local alignment using the HMM:



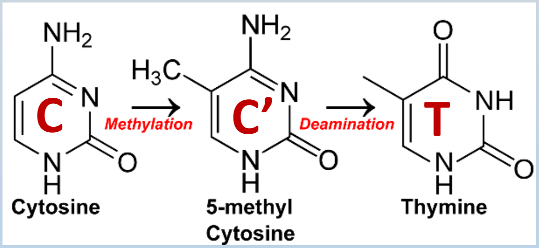


Examples of HMMs – CpG islands



What are CpGs and what is their significance?

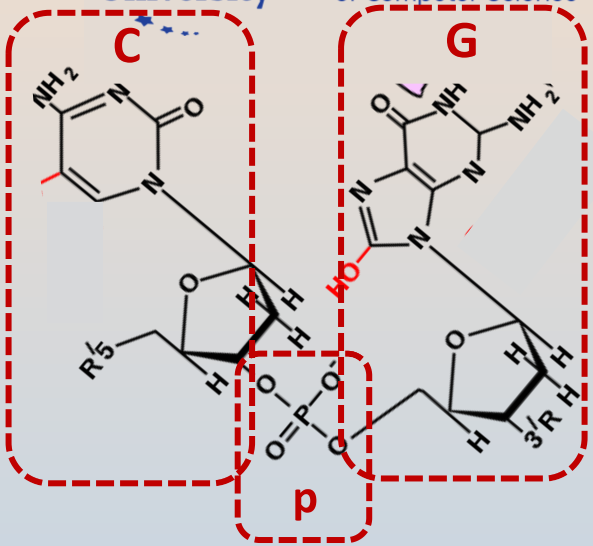
- CpG represents the C-G di-nucleotide chain **5′–C–(phosphate)–G–3′**
  (as opposed to a CG base pair)
- A common mechanism for silencing of genes (prevention of transcription) in vertebrate species is based on methylation of the C in CpGs in a gene body
- Spontaneous de-amination of a methylated C turns it to a T
- A methylated CpG can mutate into:
  - TpG (C→T)
  - CpA (C→T opposite to G, and this leads to G→A)



5′ . . .–C–G–. . .3′
3′ . . .–G–C–. . .5′

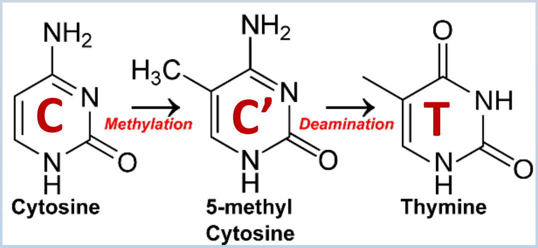
Images from https://en.wikipedia.org/wiki/CpG_site

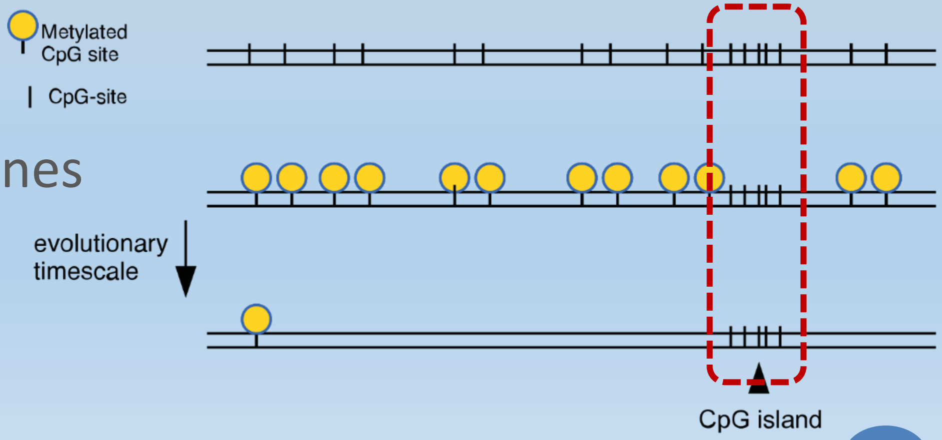
Examples of HMMs – CpG islands



What are CpG islands?

- In evolutionary timescales, CpGs get depleted from the genome.
- Roughly 0.8% of pairs are CpG, which is ~5x smaller than the expected frequency of q(C) x q(G) = 4.4%
- Genomic segments where we see relatively high concentrations of CpGs are called CpG islands.
- CpG islands occur around the start of genes, because these regions do not tend to get methylated.
- Detecting CpG islands played a central role in finding new genes in the human genome.

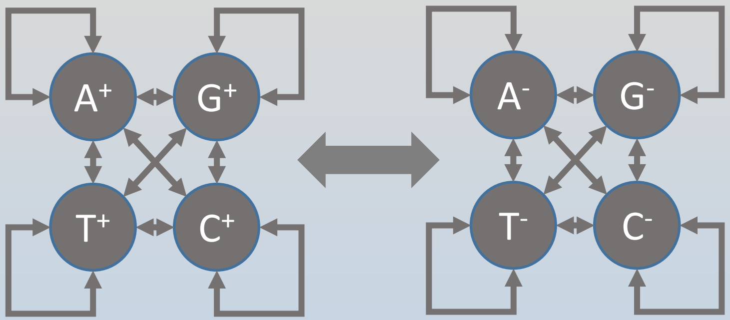




Images from https://en.wikipedia.org/wiki/CpG_site

Examples of HMMs – CpG islands

CpG island HMM:



Emission probabilities:

	A+	G+	C+	T+	A-	G-	C-	T-
A	1	0	0	0	1	0	0	0
G	0	1	0	0	0	1	0	0
C	0	0	1	0	0	0	1	0
T	0	0	0	1	0	0	0	1

(deterministic emissions)

Transition probabilities:

	A+	G+	C+	T+	A-	G-	C-	T-
A+								
G+								
C+		0.27						
T+								
A-								
G-								
C-						0.05		
T-								

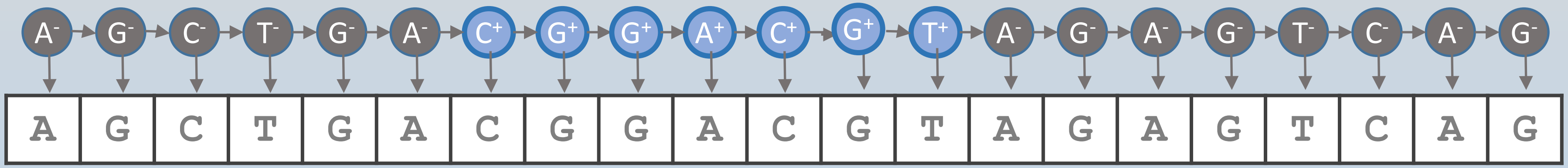
Transition in / out of CpG islands

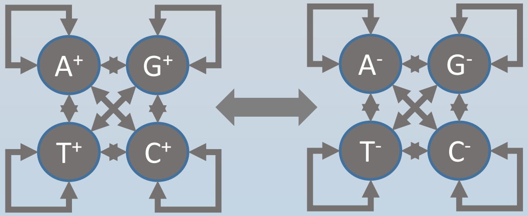
Dinucleotide frequencies in CpG islands

Dinucleotide frequencies outside of CpG islands

Examples of HMMs – CpG islands

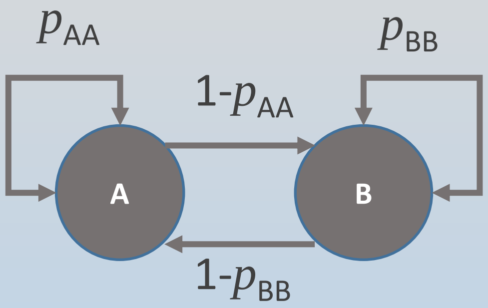
Detecting CpG islands using the HMM:





Examples of HMMs – noisy bit transmission

Simple two-state HMM used in class demonstrations:



Transitions:

	A	B
A	p_{AA}	$1\text{-}p_{AA}$
B	$1\text{-}p_{BB}$	p_{BB}

$p_{AB} = p_{AB} = 0.8$

Emissions:

	0	**1**
A	p_{A0}	$1\text{-}p_{A0}$
B	$1\text{-}p_{B1}$	p_{B1}

$p_{A0} = p_{B1} = 0.9$

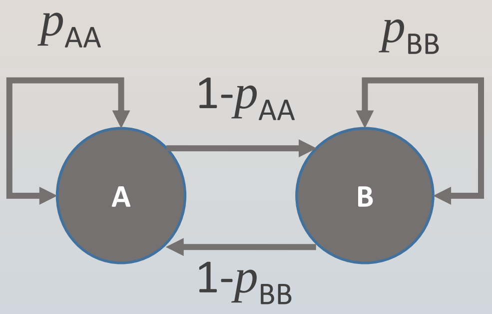
Examples of HMMs – noisy bit transmission

Generating random sequences using the HMM:

```
➔ ./hmm-run G 20   (randomly generate a sequence of length 20)
Generated sequence = 00001011111011000111
Generated path =
 - A - A - A - A - B - B - B - B - B - B - B - B - B - B - A - A - A - B - B - B
   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
   0   0   0   0   1   0   1   1   1   1   1   0   1   1   0   0   0   1   1   1
Probability: 1.68992e-07

➔ ./hmm-run G 20 (re-generate another sequence)
Generated sequence = 00000011000110010001
Generated path =
 - A - A - A - A - A - A - B - B - A - A - A - B - A - A - A - B - A - A - A - B
   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
   0   0   0   0   0   0   1   1   0   0   0   1   1   0   0   1   0   0   0   1
Probability: 5.94111e-09

➔ ./hmm-run G 20 (re-generate another sequence)
Generated sequence = 00011100001110000110
Generated path =
 - A - A - A - B - B - B - A - A - A - B - B - B - B - A - A - A - A - B - B - A
   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
   0   0   0   1   1   1   0   0   0   0   1   1   1   0   0   0   0   1   1   0
Probability: 2.37644e-08
```



Transitions:

	A	B
A	p_{AA}	$1-p_{AA}$
B	$1-p_{BB}$	p_{BB}

$p_{AB} = p_{AB} = 0.8$

Emissions:

	0	1
A	p_{A0}	$1-p_{A0}$
B	$1-p_{B1}$	p_{B1}

$p_{A0} = p_{B1} = 0.9$

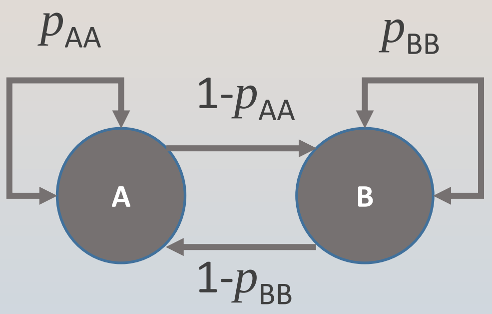
Examples of HMMs – noisy bit transmission

Reichman University
Efi Arazi School of Computer Science

Sequence annotation:

```
➔ ./hmm-run G 20   (randomly generate a sequence of length 20)
Generated sequence = 00001011111011000111
Generated path =
 - A - A - A - A - B - B - B - B - B - B - B - B - B - B - A - A - A - B - B - B
   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
   0   0   0   0   1   0   1   1   1   1   1   0   1   1   0   0   0   1   1   1

Probability: 1.68992e-07
➔ ./hmm-run V 00001011111011000111 (infer the most probable annotation for seq)
Viterbi path =
 - A - A - A - A - A - A - B - B - B - B - B - B - B - B - A - A - A - B - B - B
   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
   0   0   0   0   1   0   1   1   1   1   1   0   1   1   0   0   0   1   1   1
Probability: 1.68992e-07
```



Transitions:

	A	B
A	p_{AA}	1-p_{AA}
B	1-p_{BB}	p_{BB}

$p_{AB} = p_{AB} = 0.8$

Emissions:

	0	1
A	p_{A0}	1-p_{A0}
B	1-p_{B1}	p_{B1}

$p_{A0} = p_{B1} = 0.9$

Lecture overview

- Sequence annotation using HMMs
- Decoding problems in probabilistic models and HMMs ←
- Recap of multi-variate probability distributions
- Viterbi algorithm for most probable path

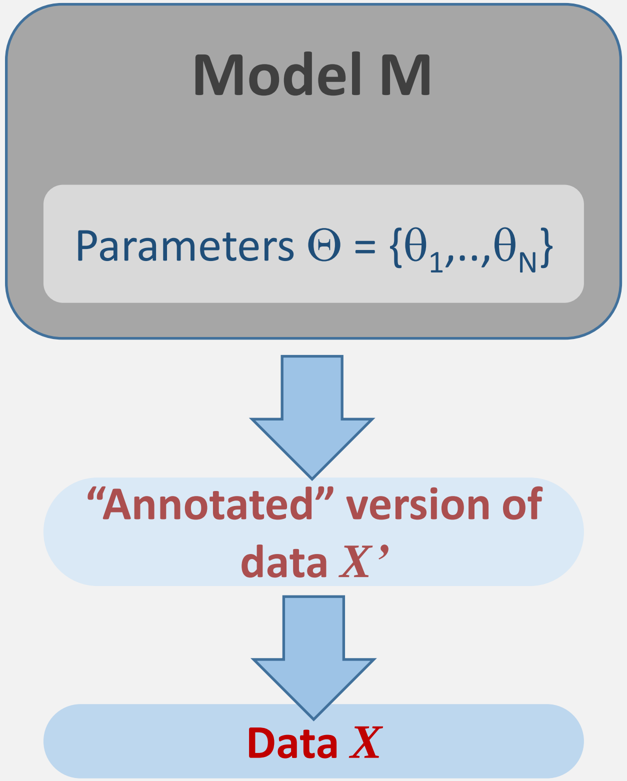
Probabilistic Models

Probabilistic model consist of two main components:

- Structural model assumptions ← Markov model on states (bounded memory)
  (e.g., hidden/observed random variables and dependencies)
- Model parameters ← Transition / emission probabilities

Three main tasks:

- Decoding: given data and complete model with parameter values, infer hidden components of the model (e.g., finding best HMM path explaining a given sequence of observations)
- Model comparison: given data and two (or more) models (with or without model parameters), decide which model fits the data better (which HMM out of a given set provides best explanation of given sequence)
- Parameter inference: given data and a model without parameter values, infer parameter values that maximize model fit (compute transition and emission probabilities that best explain a given set of sequences)

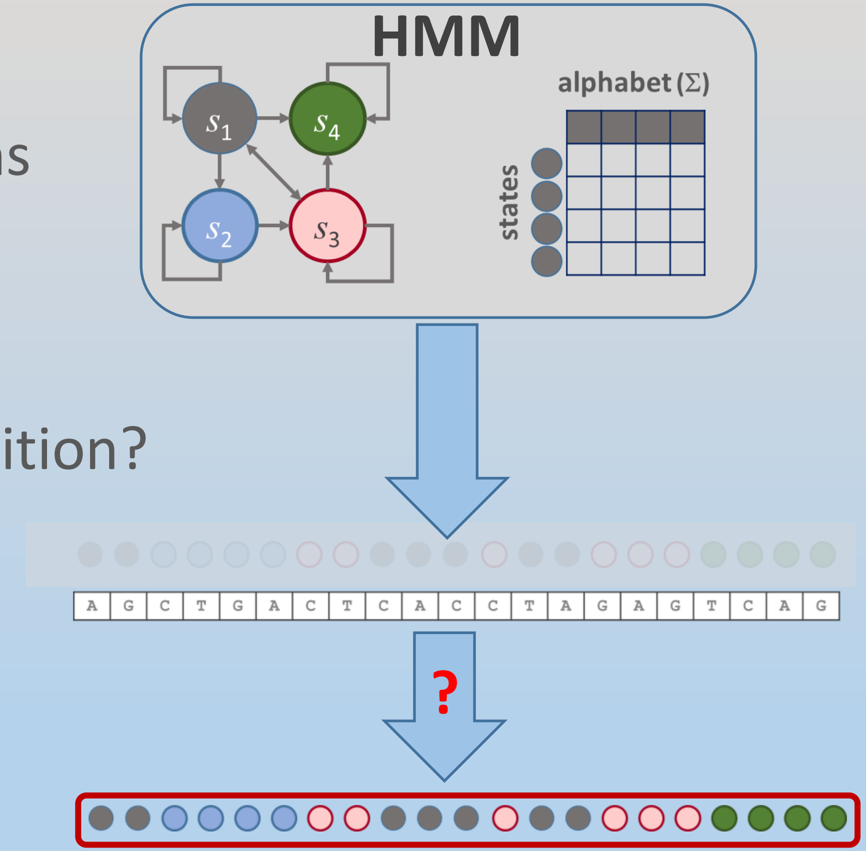


Data likelihood as measure of model fit: $P(X|M,\Theta)$

Decoding HMMs

A given HMM and observed sequence of symbols X imply a probability distribution over annotations S. Decoding questions ask questions about this distribution, such as:

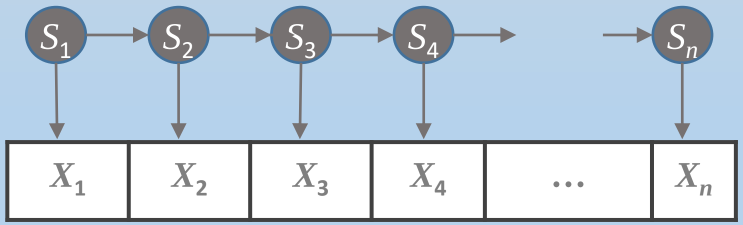
- What is the most probable annotation?
- What is the most probable state assignment at a given position?
- What is the data likelihood under the assumed model?



Distribution of annotations / paths in the HMM

- Treat the observed data $X = X_1 \ldots X_n$ and unobserved (hidden) annotation $S = S_1 \ldots S_n$ as a collection of $2n$ random variables (RVs).
- Answer questions about the joint probability distribution of X and S given the model

$$P(X,S|\text{HMM}) = ?$$



Lecture overview

- Sequence annotation using HMMs
- Decoding problems in probabilistic models and HMMs
- Recap of multi-variate probability distributions ←
- Viterbi algorithm for most probable path

Probability distribution of three random variables

Consider a joint probability distribution over three random variables A, B, and C

Marginal distribution: $P(A=a, B=b) = \sum_c P(A=a, B=b, C=c)$

Conditional distribution: $P(A=a, B=b \mid C=c) = P(A=a, B=b, C=c) / P(C=c)$

Chain law:

$$P(A=a, B=b, C=c) = P(A=a)\,P(B=b \mid A=a)\,P(C=c \mid A=a, B=b)$$
$$= P(B=b)\,P(C=c \mid B=b)\,P(A=a \mid B=b, C=c)$$

Conditional independence:

C and A are conditionally independent given B, iff $P(C=c \mid A=a, B=b) = P(C=c \mid B=b)$

This implies that:

$$P(A=a, B=b, C=c) = P(A=a)\,P(B=b \mid A=a)\,P(C=c \mid A=a, B=b)$$
$$= P(A=a)\,P(B=b \mid A=a)\,P(C=c \mid B=b)$$

Class exercise

Consider the joint probability distribution over A, B, and C specified in the table below

- Are the RVs A and B conditionally independent given C?

Steps:

1. Compute the marginal distributions P(A,C), P(B,C), and P(C).
2. Use these marginal distributions to compute the conditional distributions P(A=a| B=b, C=c) and P(A=a| C=c).
3. Compare the conditional distributions to reach a conclusion.

a	b	c	$p(a,b,c) = P(A=a, B=b, C=c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

Detailed solution will be published after class

Class exercise – solution

Consider the joint probability distribution over A, B, and C specified in the table below

- Are the RVs A and B conditionally independent given C?

Steps:

1. Compute the marginal distributions P(A,C), P(B,C), and P(C).

a	c	$P(A=a,C=c)$
0	0	0.192+0.048 = 0.24
0	1	0.144+0.216 = 0.36
1	0	0.192+0.048 = 0.24
1	1	0.064+0.096 = 0.16

a	b	c	$p(a,b,c) = P(A=a,B=b,C=c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

Class exercise – solution

Reichman University Efi Arazi School of Computer Science

Consider the joint probability distribution over A, B, and C specified in the table below

- Are the RVs A and B conditionally independent given C?

Steps:

1. Compute the marginal distributions P(A,C), P(B,C), and P(C).

b	c	$P(B=b,C=c)$
0	0	0.192+0.192 = 0.384
0	1	0.144+0.064 = 0.208
1	0	0.048+0.048 = 0.096
1	1	0.216+0.096 = 0.312

a	b	c	$p(a,b,c) = P(A=a,B=b,C=c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

Class exercise – solution

Consider the joint probability distribution over A, B, and C specified in the table below

- Are the RVs A and B conditionally independent given C?

Steps:

1. Compute the marginal distributions P(A,C), P(B,C), and P(C).

c	$P(C = c)$
0	0.384+0.096 = 0.48
1	0.208+0.312 = 0.52

b	c	$P(B = b, C = c)$
0	0	0.192+0.192 = 0.384
0	1	0.144+0.064 = 0.208
1	0	0.048+0.048 = 0.096
1	1	0.216+0.096 = 0.312

a	b	c	$p(a, b, c) = P(A = a, B = b, C = c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

Class exercise – solution

Reichman University · Efi Arazi School of Computer Science

Consider the joint probability distribution over A, B, and C specified in the table below

- Are the RVs A and B conditionally independent given C?

Steps:

1. Compute the marginal distributions P(A,C), P(B,C), and P(C).
2. Use these marginal distributions to compute the conditional distributions P(A=a| B=b, C=c) and P(A=a| C=c).

a	b	c	$p(a,b,c) = P(A=a, B=b, C=c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

b	c	$P(B=b, C=c)$
0	0	0.384
0	1	0.208
1	0	0.096
1	1	0.312

a	c	$P(A=a, C=c)$
0	0	0.24
0	1	0.36
1	0	0.24
1	1	0.16

c	$P(C=c)$
0	0.48
1	0.52

Class exercise – solution

Consider the joint probability distribution over A, B, and C specified in the table below

- Are the RVs A and B conditionally independent given C?

Steps:

1. Compute the marginal distributions P(A,C), P(B,C), and P(C).
2. Use these marginal distributions to compute the conditional distributions P(A=a| B=b, C=c) and P(A=a| C=c).
   →P(A=a| C=c) = P(A=a, C=c) / P(C=c)

c	$P(C = c)$
0	0.48
1	0.52

a	c	$P(A = a\|C = c)$
0	0	0.24 / 0.48 = 0.5
0	1	0.36 / 0.52 = 0.6923
1	0	0.24 / 0.48 = 0.5
1	1	0.16 / 0.52 = 0.3077

a	c	$P(A = a, C = c)$
0	0	0.24
0	1	0.36
1	0	0.24
1	1	0.16

Class exercise – solution

Consider the joint probability distribution over A, B, and C specified in the table below

- Are the RVs A and B conditionally independent given C?

Steps:

1. Compute the marginal distributions P(A,C), P(B,C), and P(C).
2. Use these marginal distributions to compute the conditional distributions P(A=a| B=b, C=c) and P(A=a| C=c).
   → P(A=a| B=b, C=c)
   = P(A=a , B=b, C=c) / P(B=b, C=c)

a	b	c	$p(a,b,c) = P(A=a,B=b,C=c)$
0	0	0	0.192
0	0	1	0.144
0	1	0	0.048
0	1	1	0.216
1	0	0	0.192
1	0	1	0.064
1	1	0	0.048
1	1	1	0.096

a	b	c	$P(A=a \mid B=b,C=c)$
0	0	0	0.192 / 0.384 = 0.5
0	0	1	0.144 / 0.208 = 0.6933
0	1	0	0.048 / 0.096 = 0.5
0	1	1	0.216 / 0.312 = 0.6933
1	0	0	0.192 / 0.384 = 0.5
1	0	1	0.064 / 0.208 = 0.3077
1	1	0	0.048 / 0.096 = 0.5
1	1	1	0.096 / 0.312 = 0.3077

b	c	$P(B=b,C=c)$
0	0	0.384
0	1	0.208
1	0	0.096
1	1	0.312

Class exercise – solution

Consider the joint probability distribution over A, B, and C specified in the table below

- Are the RVs A and B conditionally independent given C?

Steps:

1. Compute the marginal distributions P(A,C), P(B,C), and P(C).
2. Use these marginal distributions to compute the conditional distributions P(A=a| B=b, C=c) and P(A=a| C=c).
3. Compare the conditional distributions to reach a conclusion.

a	b	c	$P(A = a \mid B = b, C = c)$
0	0	0	0.5
0	0	1	0.6933
0	1	0	0.5
0	1	1	0.6933
1	0	0	0.5
1	0	1	0.3077
1	1	0	0.5
1	1	1	0.3077

a	c	$P(A = a \mid C = c)$
0	0	0.5
0	1	0.6923
1	0	0.5
1	1	0.3077

Class exercise – solution

Reichman University — Efi Arazi School of Computer Science

Consider the joint probability distribution over A, B, and C specified in the table below

- Are the RVs A and B conditionally independent given C? Yes!

Steps:

1. Compute the marginal distributions P(A,C), P(B,C), and P(C).
2. Use these marginal distributions to compute the conditional distributions P(A=a | B=b, C=c) and P(A=a | C=c).
3. Compare the conditional distributions to reach a conclusion.

a	b	c	$P(A=a \mid B=b, C=c)$	$P(A=a \mid C=c)$
0	0	0	0.5	0.5
0	0	1	0.6933	0.6933
0	1	0	0.5	0.5
0	1	1	0.6933	0.6933
1	0	0	0.5	0.5
1	0	1	0.3077	0.3077
1	1	0	0.5	0.5
1	1	1	0.3077	0.3077

a	c	$P(A=a \mid C=c)$
0	0	0.5
0	1	0.6923
1	0	0.5
1	1	0.3077

Probability distribution of multiple random variables

Same goes with larger sets of random variables $A_1, A_2, \ldots, A_N$:

Marginal distribution: $P(A_1=a_1, \ldots, A_i=a_i) = \sum_{a_{i+1} \ldots a_n} P(A_1=a_1, \ldots, A_i=a_i, A_{i+1}=a_{i+1}, \ldots, A_n=a_n)$

Conditional distribution:

$$P(A_1=a_1, \ldots, A_i=a_i \mid A_{i+1}=a_{i+1}, \ldots, A_n=a_n) = P(A_1=a_1, \ldots, A_i=a_i, A_{i+1}=a_{i+1}, \ldots, A_n=a_n) \;/\; P(A_{i+1}=a_{i+1}, \ldots, A_n=a_n)$$

Chain law: $P(A_1, A_2, \ldots, A_N) = P(A_1) \times P(A_2|A_1) \times P(A_3|A_2 A_1) \times \ldots \times P(A_N|A_{N-1} \ldots A_1)$

Conditional independence:

A_N is conditionally independent of $A_1, A_2, \ldots, A_{i-1}$ given $A_i, \ldots, A_{N-1}$, iff $P(A_N|A_{N-1}, \ldots, A_1) = P(A_N|A_{N-1}, \ldots, A_i)$

- When the value (a_i) is known from context, we will use the shorthand $P(A_i)$ for $P(A_i=a_i)$

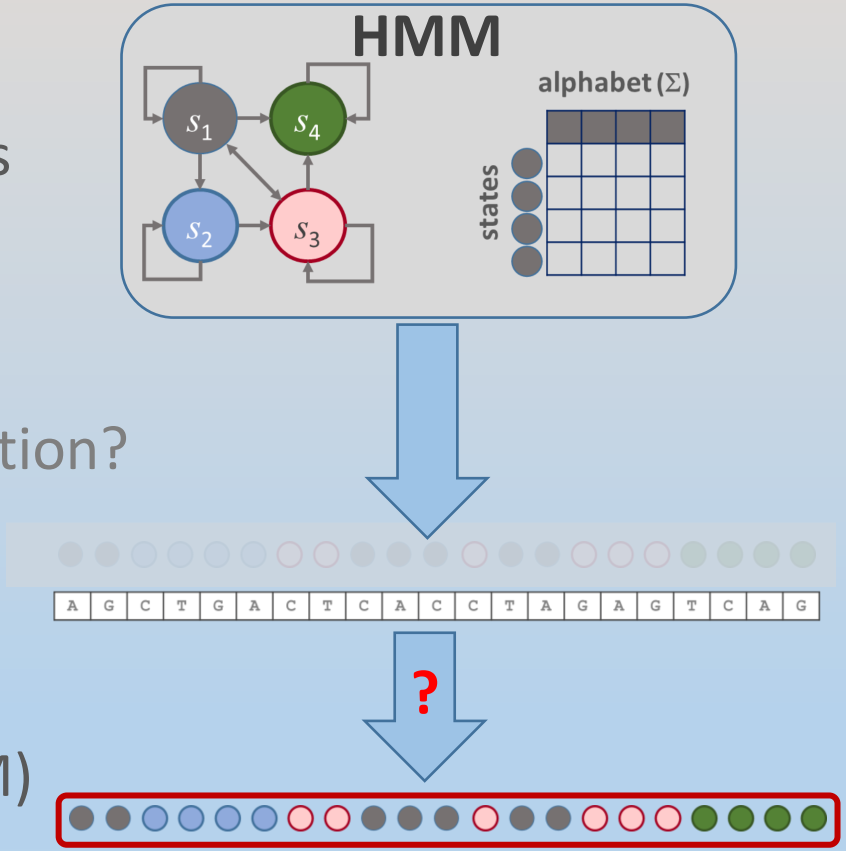
Lecture overview

- Sequence annotation and HMMs
- Examples of sequence HMMs
- Decoding problems in HMMs
- Viterbi algorithm for most probable path ←

Decoding HMMs

A given HMM and observed sequence of symbols X imply a probability distribution over annotations S. Decoding questions ask questions about this distribution, such as:

- What is the most probable annotation? ←
- What is the most probable state assignment at a given position?
- What is the data likelihood under the assumed model?

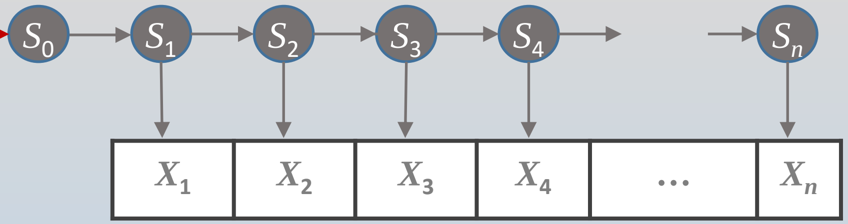


Objective: for a given observed sequence $X = X_1 \ldots X_n$ find the sequence of annotations $S = S_1 \ldots S_n$ that maximizes $P(X,S|HMM)$

Distribution of annotations / paths in the HMM

The joint probability of an annotation path and the observed sequence:

S_0 – an added constant initial state with no emissions



Using the chain law, we get:

$$P(X,S|\text{HMM}) = P(S_1 \mid S_0)\, P(X_1 \mid S_1 S_0)\, P(S_2 \mid X_1 S_1 S_0)\, P(X_2 \mid S_2 X_1 S_1 S_0) \ldots$$

$$= \prod_i P(S_i \mid X_1 \ldots X_{i-1} S_0 \ldots S_{i-1})\, P(X_i \mid X_1 \ldots X_{i-1} S_1 \ldots S_i)$$

$$= \prod_i P(S_i \mid S_{i-1})\, P(X_i \mid S_i)$$

$$= \prod_i t(S_{i-1} \rightarrow S_i)\ \ e(S_i \rightarrow X_i)$$

transition emission

conditional independence (CI) in HMMs:

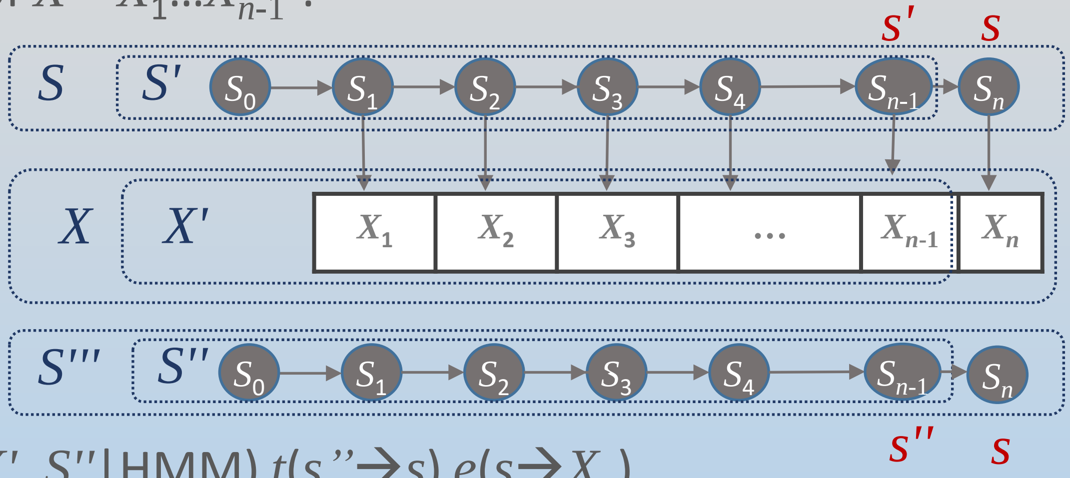
given S_i , X_i is CI of all other variables S_j , X_j $(j \neq i)$ and S_{i+1} is CI of X_j $(j \leq i)$, S_j $(j < i)$.

Decoding – finding most probable annotation

Reichman University
Efi Arazi School of Computer Science

Question: Can we use a greedy algorithm to compute the most probable path? **Answer:** No!!

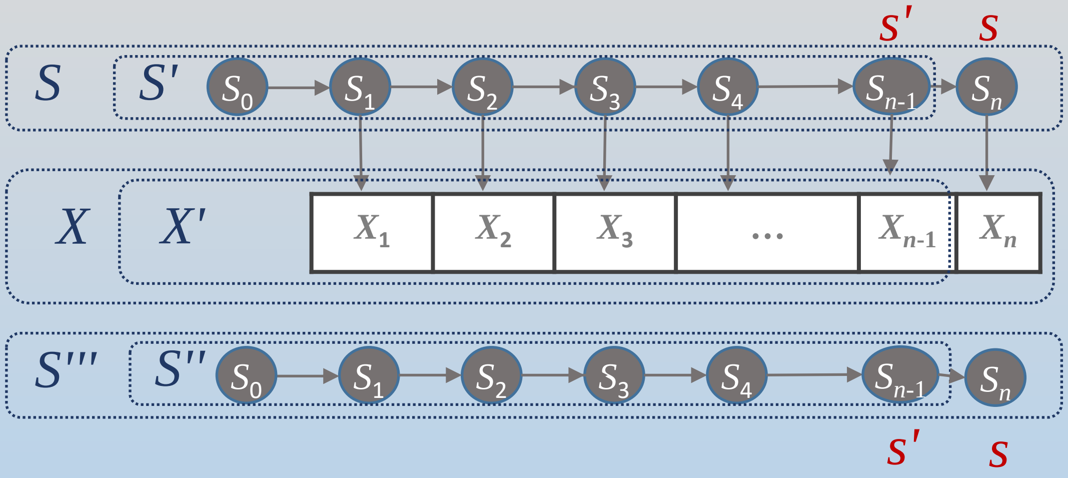
If $S = S_1 \ldots S_n$ is a max-probability annotation of $X = X_1 \ldots X_n$,
is its prefix $S' = S_1 \ldots S_{n-1}$ a max-probability annotation of $X' = X_1 \ldots X_{n-1}$?



- Let s,s' be the last two states in S.
  Then, $P(X,S|HMM) = P(X', S'|HMM)\ t(s' \rightarrow s)\ e(s \rightarrow X_n)$.
  [derive using formula from last slide]
- Now consider an arbitrary path S'' of length n-1,
  and let s'' be the last state in this path.
- The path $S''' = S''s$ has joint probability: $P(X, S'''|HMM) = P(X', S''|HMM)\ t(s'' \rightarrow s)\ e(s \rightarrow X_n)$.
- This implies: $P(X,S|HMM) / P(X, S'''|HMM) = (P(X',S'|HMM) / P(X',S''|HMM)) \times (t(s' \rightarrow s) / t(s'' \rightarrow s))$
- So, it's possible to have $P(X',S'|HMM) < P(X',S''|HMM)$ if $t(s' \rightarrow s) > t(s'' \rightarrow s)$
- Note, however, that S' has the highest joint probability with X' among all paths that end with s'.

Decoding – finding most probable annotation

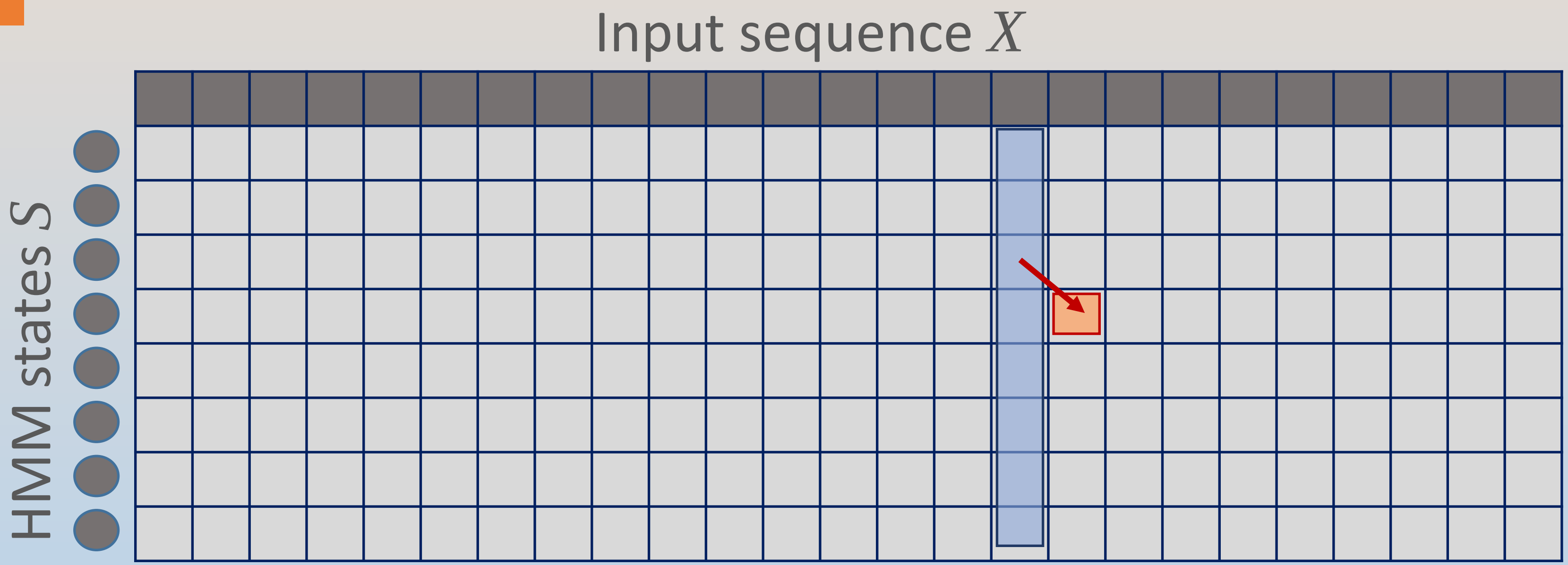
Claim: If $S = S_1 \ldots S_n$ is a max-probability annotation of $X = X_1 \ldots X_n$ among annotations that end with $S_n = s$, then its prefix $S' = S_1 \ldots S_{n-1}$ a max-probability annotation of $X' = X_1 \ldots X_{n-1}$ among annotations that end with $S_{n-1} = s'$.



Proof:

- Let s, s' be the last two states in S.
  Then, $P(X,S|HMM) = P(X',S'|HMM)\ t(s' \rightarrow s)\ e(s \rightarrow X_n)$.
  [derive using formula from last slide]
- Now consider an arbitrary path S'' of length n-1 that ends with state s'.
- The path $S''' = S''s$ has joint probability: $P(X,S'''|HMM) = P(X',S''|HMM)\ t(s' \rightarrow s)\ e(s \rightarrow X_n)$.
- This implies: $P(X,S|HMM) / P(X,S'''|HMM) = (P(X',S'|HMM) / P(X',S''|HMM)) \times (t(s' \rightarrow s) / t(s' \rightarrow s))$
- Since we know that $P(X,S|HMM) / P(X,S'''|HMM) > 1$,
  then we must have $P(X',S'|HMM) / P(X',S''|HMM) > 1$

Outline for Viterbi's algorithm



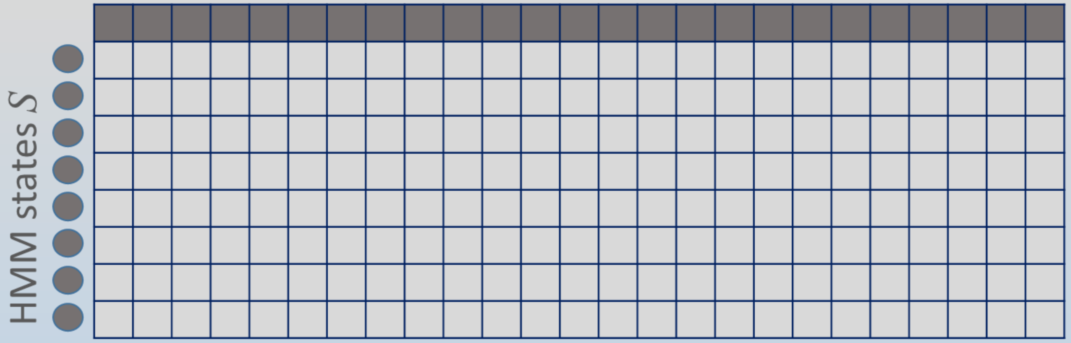
- Fill a DP matrix V, where V[i,j] will hold the max-probability annotation $S_1 \ldots S_i$ of the prefix $X_1 \ldots X_i$ among those that end with state $S_i = s_j$ (the i'th state in the path equals s_j)
- Update formula: (implied by claim from previous slide)

  $V[i,j] = \max_l \{V[i-1,l] \times t(s_l \rightarrow s_j) \times e(s_j \rightarrow X_i)\} = \max_l \{V[i-1,l] \times t(s_l \rightarrow s_j)\} \times e(s_j \rightarrow X_i)$
- Update pointers are used to reconstruct the path

Viterbi's algorithm

Input: an HMM with k states $s_1 \ldots s_k$ and a sequence of observed symbols $X_1 \ldots X_n$

Objective: Compute a matrix V s.t. V[i,j] holds the max-probability annotation $S_1 \ldots S_i$ of the prefix $X_1 \ldots X_i$ among those that end with state $S_i = s_j$ (the i'th state in the path equals s_j)



Initialization: V[0,0] = 1 and V[0, j] = 0 for j=1..k (s_0 is the added initial state)

Update: for each i=1..n and j=1..k compute V[i,j] as follows:

$$V[i,j] = \max_l \{V[i-1,l] \times t(s_l \rightarrow s_j)\} \times e(s_j \rightarrow X_i)$$ + keep pointer to cell [i-1,l] used in update

Output: reconstruct max-probability annotation by finding cell V[n,j] with highest value and tracing back from it using the update pointers all the way to V[0,0].

Viterbi's algorithm

Correctness: implied by the claim on slide 41

Complexity:

Space – $O(kn)$ the algorithm keeps kxn values V[i,j] and pointers

Time – $O(k^2n)$ calculation of V[i,j] requires finding the maximum among k possible values.

Comments:

- Hirschberg's technique can be used to reduce space complexity to $O(k)$ with time complexity $O(k^2n\log(n))$ (typically $k<n$)
- Time complexity eventually depends on the number of non-zero transitions, which can be less than k^2 (as is in the case of the alignment and motif detection HMMs)

Noisy bit transmission

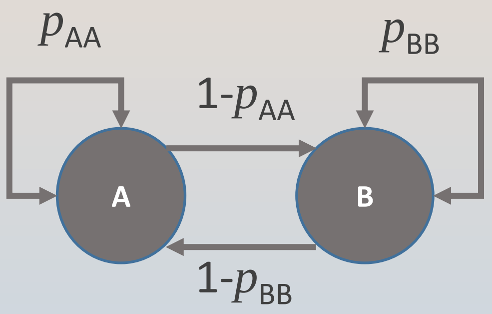
Execution of Viterbi's algorithm on simple HMM:

```
➔ ./hmm-run G 20   (randomly generate a sequence of length 20)
Generated sequence = 00001011111011000111
Generated path =
 - A - A - A - A - B - B - B - B - B - B - B - B - B - B - A - A - A - B - B - B
   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
   0   0   0   0   1   0   1   1   1   1   1   0   1   1   0   0   0   1   1   1

Probability: 1.68992e-07
➔ ./hmm-run V 00001011111011000111 (infer the most probable annotation for seq)
Viterbi path =
 - A - A - A - A - A - A - B - B - B - B - B - B - B - B - A - A - A - B - B - B
   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
   0   0   0   0   1   0   1   1   1   1   1   0   1   1   0   0   0   1   1   1
Probability: 1.68992e-07
Viterbi matrix:

   0.45    0.32    0.23    0.17    0.013  0.0097 0.00077 6.2e-05 2.5e-05 1.8e-05 1.3e-05
   0.05   0.009  0.0065  0.0047     0.03  0.0024  0.0017  0.0013  0.0009 0.00065 0.00047

 8.4e-05 6.7e-06 5.4e-07 3.5e-06 2.5e-06 1.8e-06 1.4e-07 1.2e-08 4.7e-09
 3.7e-05 2.7e-05 1.9e-05 1.6e-06 1.2e-07   5e-08 3.3e-07 2.3e-07 1.7e-07
```



Transitions:

	A	B
A	p_{AA}	$1-p_{AA}$
B	$1-p_{BB}$	p_{BB}

$p_{AB} = p_{AB} = 0.8$

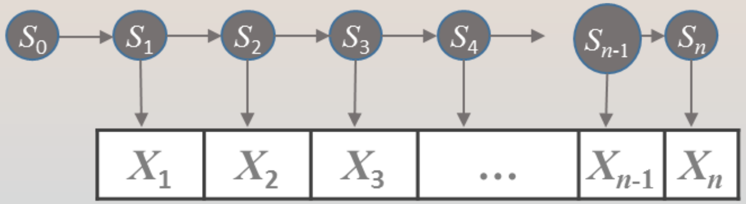
Emissions:

	0	1
A	p_{A0}	$1-p_{A0}$
B	$1-p_{B1}$	p_{B1}

$p_{A0} = p_{B1} = 0.9$

HMM introduction summary

Reichman University Efi Arazi School of Computer Science



- HMMs are general and useful models for defining "hidden" annotations for sequential data.
- HMMs are based on a Markov model of the hidden states + emissions that connect observations with the states.
- The Markov model enables efficient decoding algorithms
- Next week we'll discuss more decoding algorithms and start thinking about how the probabilities of the model can be inferred.