Bioinformatics and Biostatistics BB2440: Biostatistics Lecture 7: Markov Chains & Hidden Markov Models

Timo Koski

TK

17.09.2014

KTH Matematik

Lecture 7: Purpose

The purpose of this lecture is to make a transition from the biostatistics part of the course to
Marketa Zvelebil & Jeremy Baum: Understanding bioinformatics. 2008, Garland Science.
(to be referred to as ZB) in the sense that the lecture deals with basics of the Markov statistical models underlying some of chapter 6 and other sections of ZB.

KTH Matematik

Lecture 7: Contents

- an introductory example of a Markov chain: McCabe's library
- Markov chain concepts: transition probability, Markov property, state diagram, trellis
- Markov chains and genomic sequence analysis
- Hidden Markov Model (HMM) and sequence analysis
  1. HMM for sequence heterogeneity
  2. HMM for Protein families
- Algorithms for Hidden Markov Models (are named)

KTH Matematik

A generic introduction, neither biostatistics nor bioinformatics involved here

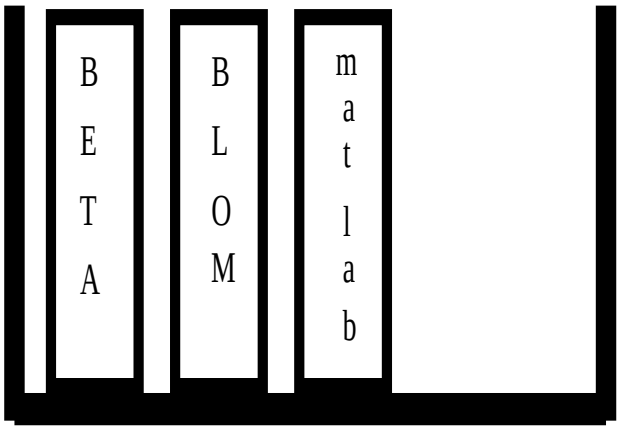
Learning outcomes:

- state
- state diagram/graph
- trellis
- transition
- transition probability
- Markov property

KTH Matematik

Introduction: Linnea's library

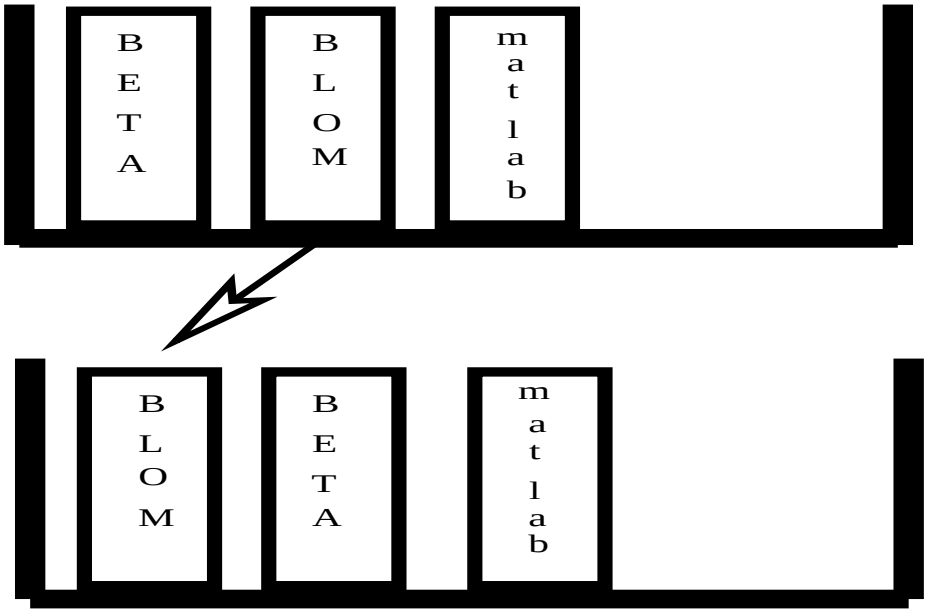
Linnea has a set of three books on a bookshelf. These are (1) L. Råde & B. Westergren: **BETA**, (2) G. Blom: Probability och Statistics, (3) F. Gustafsson & N. Bergman: MATLAB^{R} for Engineers.



KTH Matematik

Linnea's library

Every time Linnea has consulted one of these books, she will insert the book back on the shelf as the first one from the left. The figure depicts the change in the order of the books after Linnea has sought advice and inspiration from the book by G. Blom and put it back to the shelf.



KTH VETENSKAP OCH KONST
KTH Matematik

Linnea's library

Linnea never takes two or three books from the shelf at a time and neither does she introduce new books on the shelf or lets anyone else tamper with the valuable books. Let us assume that the popularities (or the relative frequencies) for Linnea to pick each and every of the three books can be described by the distribution $p_i > 0, i = 1, 2, 3$, respectively, $p_1 + p_2 + p_3 = 1$. In addition we assume that Linnea picks up the books independently of each other.

KTH Matematik

Linnea's library: a sequence of transitions

Suppose that the books are initially ordered as $M\beta B$. Linnea picks up then books in the sequence

$$\beta \quad \beta \quad B \quad M \quad B \quad \beta$$

and returns them. The order of three books in the shelf changes as follows:

$$M\beta B \mapsto \beta MB \mapsto \beta MB \mapsto B\beta M \mapsto MB\beta \mapsto BM\beta \mapsto \beta MB$$

This sequence of orders (read from the left) of the books depicts a sequence of transitions between randomly chosen triplets.

KTH Matematik

Linnea's library: states and state space

We shall refer to the triplets of books as *states*. We call

$$\mathcal{S} = \{\beta BM, \beta MB, B\beta M, BM\beta, M\beta B, MB\beta\}$$

the *state space*.

KTH Matematik

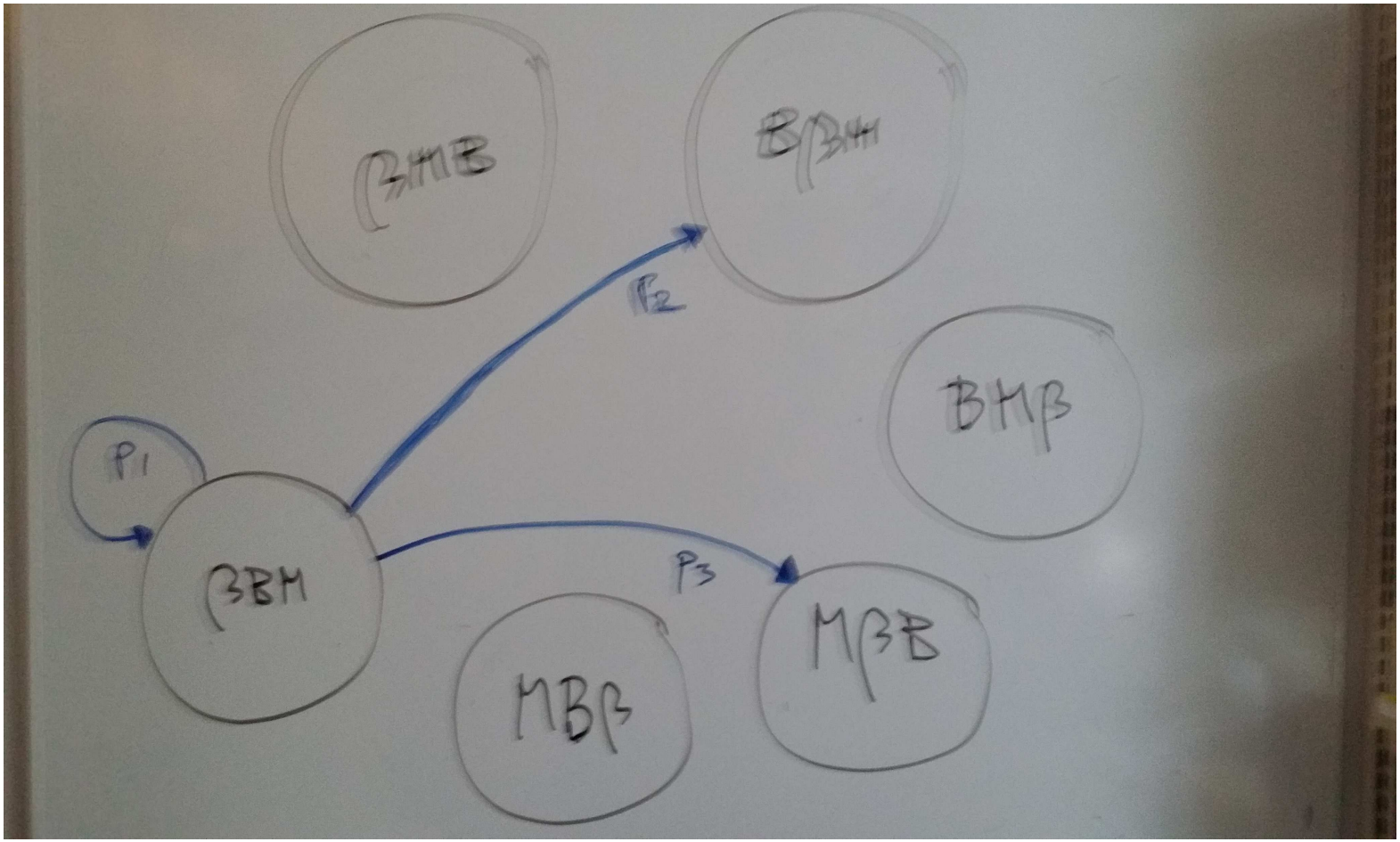
Linnea's library: state graph

The states are visualized as nodes in the state graph.

KTH Matematik

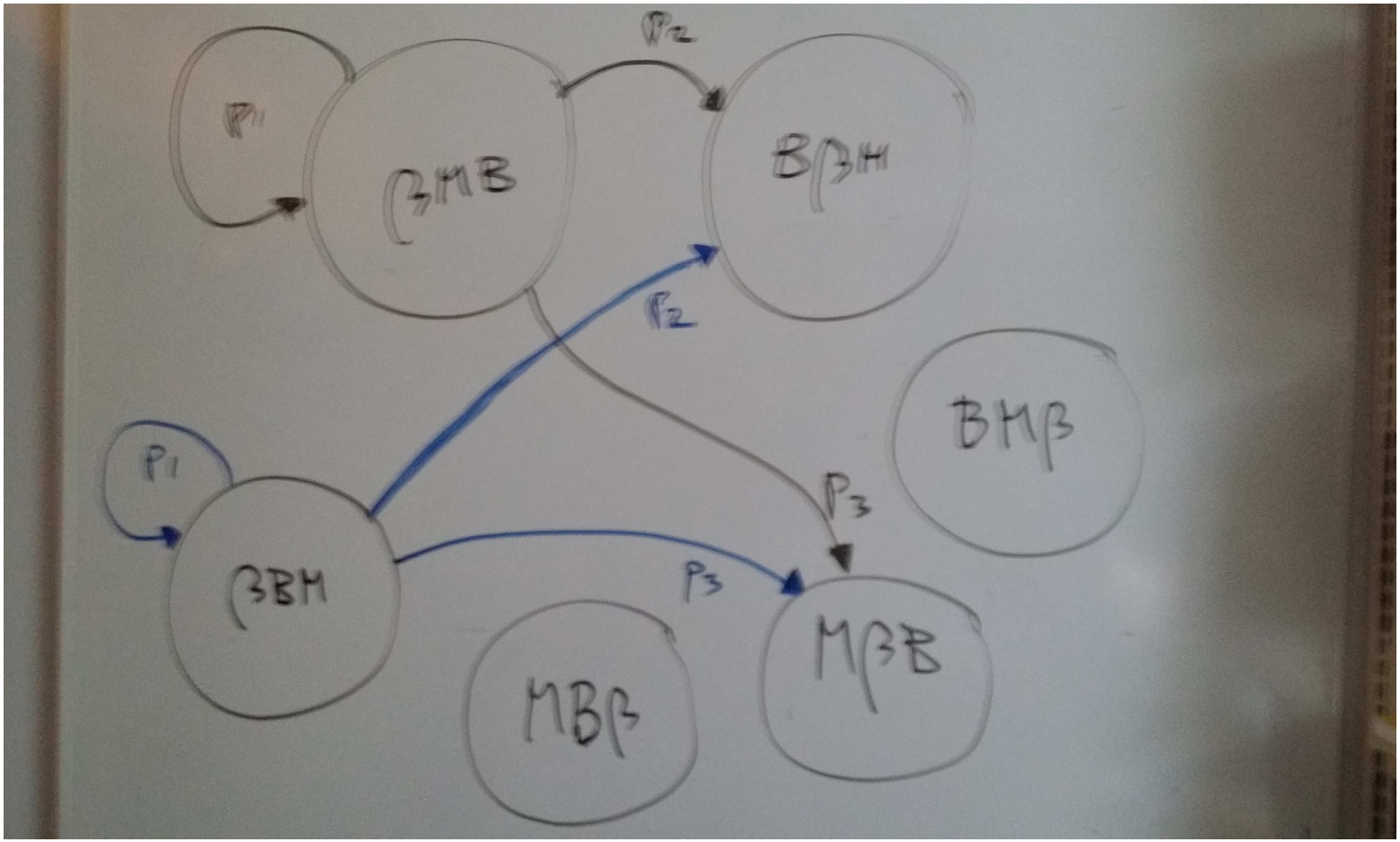
Linnea's library: state graph & transitions

The possible transitions from βBM are depicted by an arrow with the corresponding probability



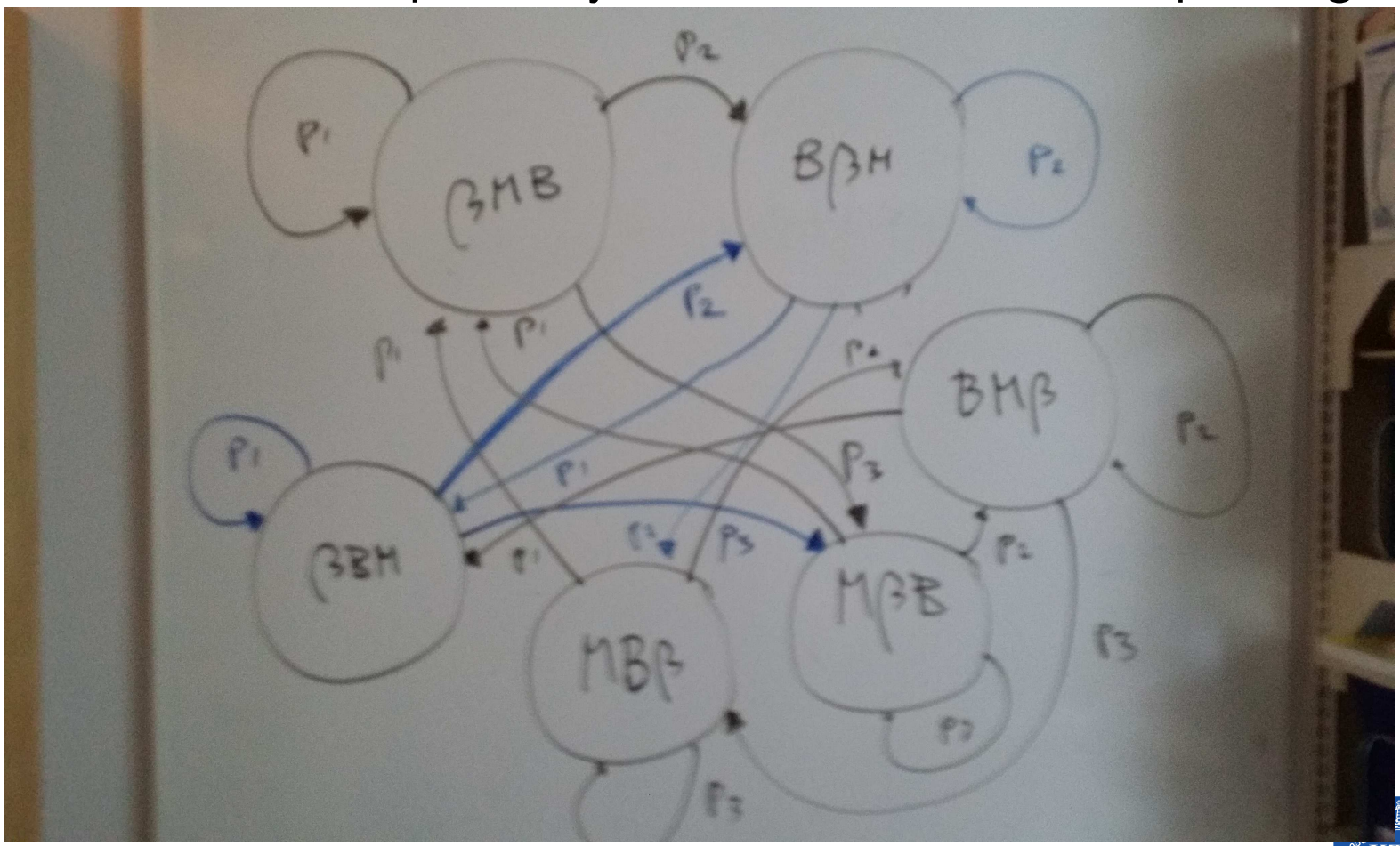
Linnea's library: state graph & transitions

The possible transitions from βBM and from βMB are depicted by an arrow with the corresponding probability



Linnea's library: state graph & transitions

All possible transitions are depicted by an arrow with the corresponding probability



Library: a Matrix

$p_1 = \Pr(\beta),\ p_2 = \Pr(B),\ p_3 = \Pr(M)$

	β BM	β MB	B β M	BM β	M β B	M B β
β BM	p_1	0	p_2	0	p_3	0
β MB	0	p_1	p_2	0	p_3	0
B β M	p_1	0	p_2	0	0	p_3
BM β	p_1	0	0	p_2	0	p_3
M β B	0	p_1	0	p_2	p_3	0
M B β	0	p_1	0	p_2	0	p_3

.

KTH VETENSKAP OCH KONST
KTH Matematik

Linnea's library: Transition probability Matrix

To be understood as follows: take a state in the left column. Say, e.g., β MB. In the row corresponding to β MB you find the **conditional probabilities** of transition to any other state.

	β BM	β MB	B β M	BM β	M β B	M B β
β BM	p_1	0	p_2	0	p_3	0
β MB	0	p_1	p_2	0	p_3	0
B β M	p_1	0	p_2	0	0	p_3
BM β	p_1	0	0	p_2	0	p_3
M β B	0	p_1	0	p_2	p_3	0
M B β	0	p_1	0	p_2	0	p_3

.

KTH VETENSKAP OCH KONST
KTH Matematik

Transition probability

The conditional probabilities of the next state given the current state are called transition probabilities.

KTH Matematik

Linnea's library: Transition probability Matrix

The sum of **transition probabilities** on every row is $= 1$.

	β BM	β MB	B β M	BM β	M β B	M B β
β BM	p_1	0	p_2	0	p_3	0
β MB	0	p_1	p_2	0	p_3	0
B β M	p_1	0	p_2	0	0	p_3
BM β	p_1	0	0	p_2	0	p_3
M β B	0	p_1	0	p_2	p_3	0
M B β	0	p_1	0	p_2	0	p_3

.

KTH VETENSKAP OCH KONST
KTH Matematik

Trellis in gardening

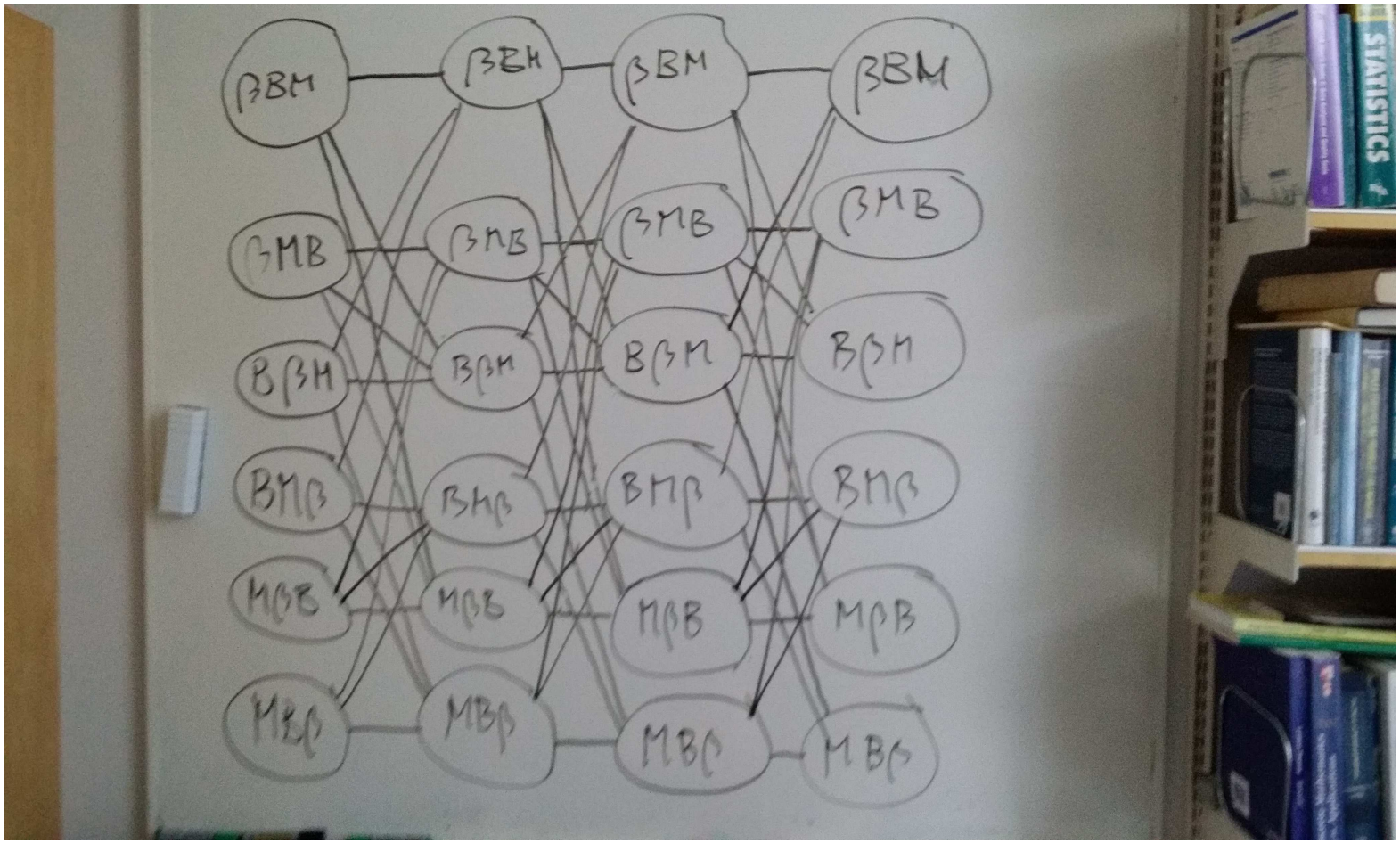
A trellis is a structure, usually made from an open framework or lattice of interwoven or intersecting pieces of wood, bamboo or metal that is usually made to support and display climbing plants.

Trellis in mathematics & engineering

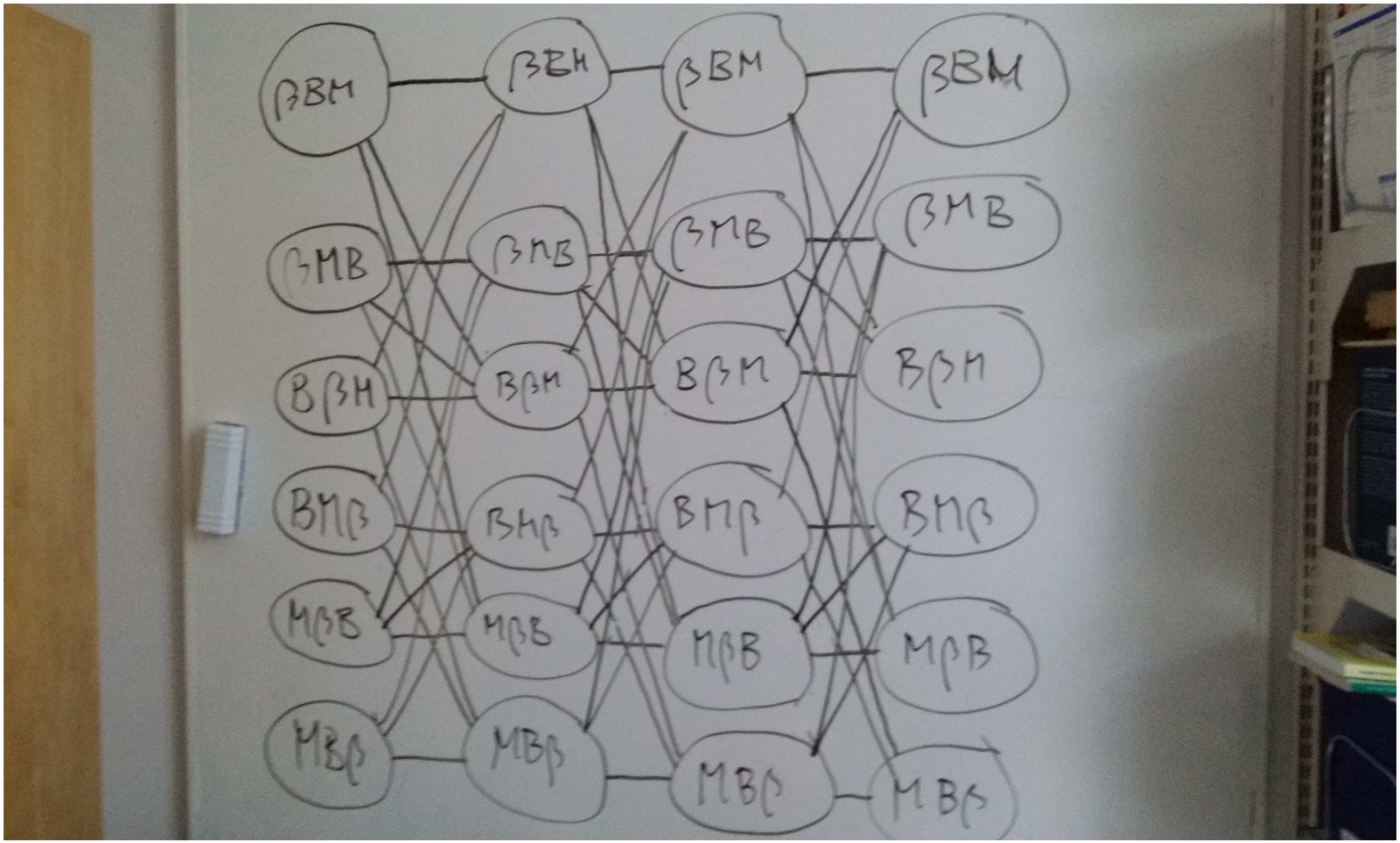
A trellis is a graph whose nodes are ordered into vertical slices (time), and with each node at each time connected to at least one node at an earlier and at least one node at a later time.

KTH Matematik

Trellis: shows all possible paths of state transitions



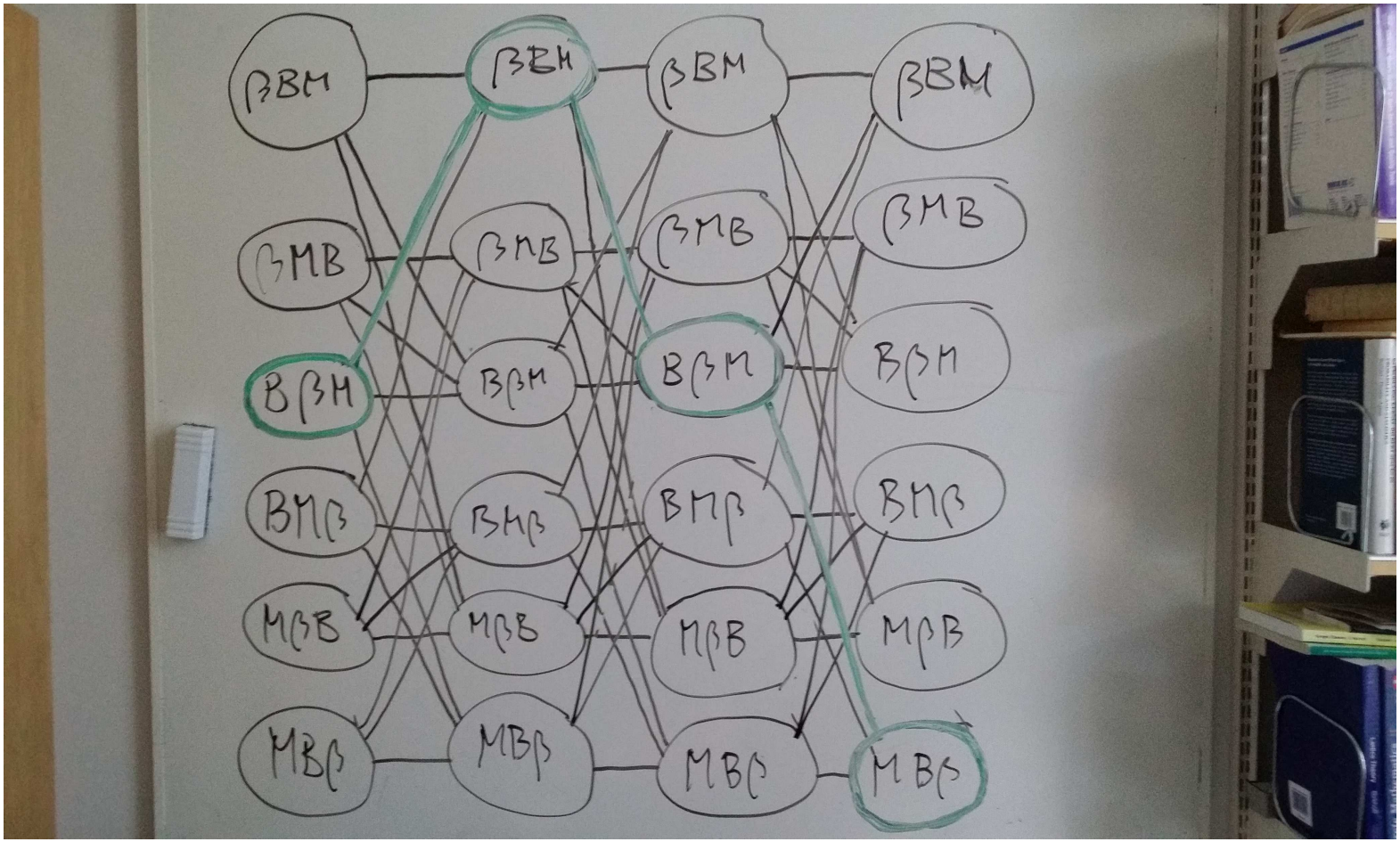
Trellis: shows all possible paths of state transitions



KTH Matematik

A sequence of states visualized in the trellis

Start with $B\beta M$, pick up books in the sequence $\beta \quad B \quad M$. The sequence of states is the green path in the trellis:



A bit more on trellises later in this lecture

There will be more about trellises of Markov chains, when we mention the problem of the most probable path of a Hidden Markov model.

KTH Matematik

The probability of a sequence of states

Start with $M\beta B$, pick up books in the sequence $\beta \quad \beta \quad B \quad M \quad B \quad \beta$
Then the corresponding sequence of states is

$$M\beta B \mapsto \beta MB \mapsto \beta MB \mapsto B\beta M \mapsto MB\beta \mapsto BM\beta \mapsto \beta BM$$

The probability of the sequence is

$$\Pr\left(M\beta B \mapsto \beta MB \mapsto \beta MB \mapsto B\beta M \mapsto MB\beta \mapsto BM\beta \mapsto \beta BM\right)$$

$$= p_1 p_1 p_2 p_3 p_2 p_1$$

We multiply the transition probabilities as Linnea picks up the books independently.

KTH Matematik

Linnea's library: random variables

Let us set

$X_n \stackrel{\text{def}}{=}$ the state after Linnea has returned a book for the nth time

The values of X_n are in $\mathcal{S} = \{\beta BM, \beta MB, B\beta M, BM\beta, M\beta B, MB\beta\}$.

- X_n is a random variable whose values are not numbers !
- the random variables $X_0, X_1, \ldots, X_n, \ldots$ are NOT independent.

KTH VETENSKAP OCH KONST
KTH Matematik

Linnea's library: random variables

X_n = the state after Linnea has returned a book for the nth time and then we have, e.g,

$$Pr\left(X_{n+1} = MB\beta \mid X_n = B\beta M\right) = p_3$$

e.t.c..

Linnea's library: random variables

$$\Pr\left(M\beta B \mapsto \beta MB \mapsto \beta MB \mapsto B\beta M \mapsto MB\beta \mapsto BM\beta \mapsto \beta MB\right)$$

$$= p_1 p_1 p_2 p_3 p_2 p_1$$

$$= Pr\left(X_1 = \beta MB \mid X_0 = M\beta B\right) Pr\left(X_2 = \beta MB \mid X_1 = \beta MB\right) Pr\left(X_3 = B\beta M \mid X_2 = \beta MB\right)$$

$$\cdot Pr\left(X_4 = MB\beta \mid X_3 = B\beta M\right)$$

$$\cdot Pr\left(X_5 = BM\beta \mid X_4 = MB\beta\right) Pr\left(X_6 = \beta BM \mid X_5 = BM\beta\right)$$

KTH Matematik

Linnea's library: random variables

$$Pr(X_0 = M\beta B, X_1 = \beta MB, X_2 = \beta MB, X_3 = B\beta M, X_4 = MB\beta, X_5 = BM\beta, X_6 = \beta MB) =$$

$$\begin{aligned} &Pr\left(X_1 = \beta MB \mid X_0 = M\beta B\right) Pr\left(X_2 = \beta MB \mid X_1 = \beta MB\right) \\ \cdot &Pr\left(X_3 = B\beta M \mid X_2 = \beta MB\right) Pr\left(X_4 = MB\beta \mid X_3 = B\beta M\right) \\ \cdot &Pr\left(X_5 = BM\beta \mid X_4 = MB\beta\right) Pr\left(X_6 = \beta BM \mid X_5 = BM\beta\right) \end{aligned}$$

We should multiply this by a start probability $Pr(X_0 = M\beta B)$, but we take this for simplicity as $= 1$.

KTH VETENSKAP OCH KONST
KTH Matematik

Markov property (1)

Consider the following:

$$\underbrace{X_0 X_1 \ldots X_{n-1}}_{\text{Past}} \underbrace{X_n}_{\text{Present}} \underbrace{X_{n+1}}_{\text{Future}}$$

Or, to be more concrete

$$\underbrace{X_0 = M\beta B, X_1 = \beta MB, X_2 = \beta MB, X_3 = B\beta M, X_4 = MB\beta}_{\text{Past}}$$

$$\underbrace{X_5 = BM\beta}_{\text{Present}}$$

$$\underbrace{X_6 = \beta BM}_{\textit{Future}}$$

Then we want to find

$$Pr\left(X_6 = \beta BM \mid \text{Present}, \text{Past}\right).$$

This can be seen as a prediction of the future given present and past.

Markov property (2)

By definition of conditional probability

$$Pr\left(X_6 = \beta BM \mid \text{Present}, \text{Past}\right) = \frac{Pr(X_6 = \beta BM, \text{Present}, \text{Past})}{Pr(\text{Present}, \text{Past})}$$

From the above this is equal to

$$= \frac{p_1 p_1 p_2 p_3 p_2 p_1}{p_1 p_1 p_2 p_3 p_2} = p_1 = Pr\left(X_6 = \beta BM \mid X_5 = BM\beta\right)$$

Note that the zero transition probabilities can be handled by canceling zeros in the numerator and denominator.

KTH Matematik

Markov property (3)

We found that

$$Pr\left(X_6 = \beta MB \mid \text{Present}, \text{Past}\right) = Pr\left(X_6 = \beta MB \mid X_5 = BM\beta\right)$$

$$= Pr\left(X_6 = \beta MB \mid \text{Present}\right)$$

Because Linnea picks up a new book independently of each other, the sequence $X_1, \ldots, X_n, \ldots$ *lacks memory* (of the past). This is what is known as the **Markov property**: for finding the probability of any the future state it is enough to know the present state, the past plays no role.

KTH VETENSKAP OCH KONST
KTH Matematik

The frog pond, where the frog jumps from leaf to leaf and loses memory (suffers amnesia) at each jump

KTH Matematik

Real name = McCabe's library

The preceding library stuff is a special case of a known model for self-organization of linear lists of data records and is called **McCabe's library**. It is of interest in computer science as **dynamic file management** and **cache management**.

KTH Matematik

End of introduction

KTH Matematik

Markov chains

Think of any state space $\mathcal{S}$ with a finite number of discrete states. Let $X_0, X_1, \ldots, X_{n-1}, X_n, X_{n+1}, \ldots$ be a sequence of r.v.'s with values in $\mathcal{S}$ and such that if

$$\underbrace{X_0 X_1 \ldots X_{n-1}}_{\text{Past}} \underbrace{X_n}_{\text{Present}} \underbrace{X_{n+1}}_{\text{Future}}$$

then **Markov property** holds.

$$Pr\left(\text{Future} \mid \text{Present}, \text{Past}\right) = Pr\left(\text{Future} \mid \text{Present}\right).$$

Then we call $X_0, X_1, \ldots, X_{n-1}, X_n, X_{n+1}, \ldots$ a **Markov chain**.

KTH VETENSKAP OCH KONST
KTH Matematik

Markov chains

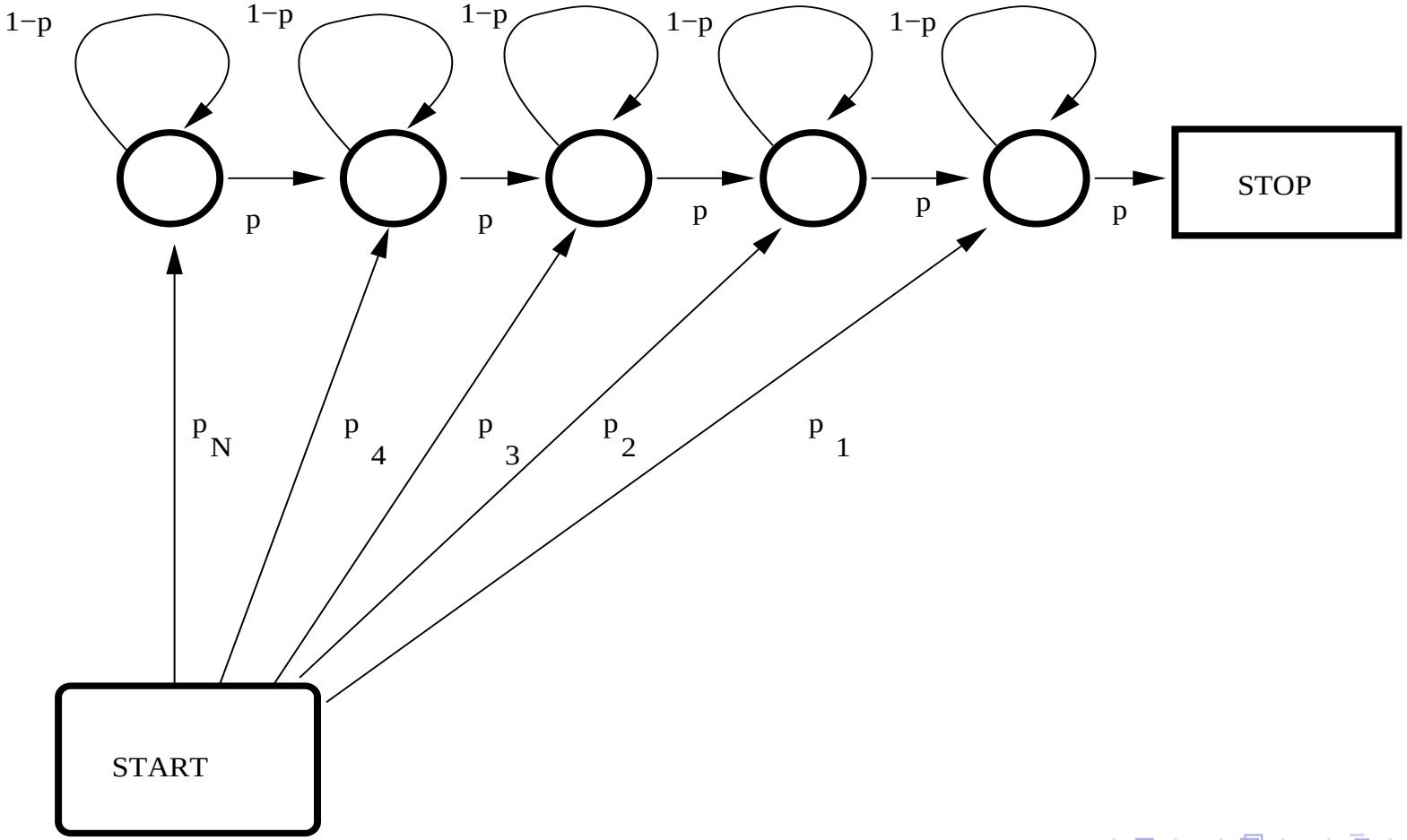
Learning outcomes:

- You can invent Markov chains for your purposes by giving the state graph !
- In bioinformatics the state graph is often called the **architecture** of the model.

KTH Matematik

Example: A terminating Markov chain (1)

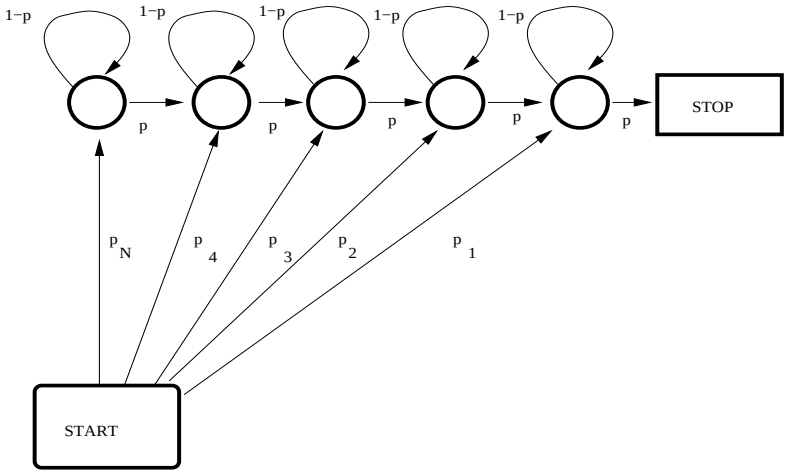
This is a terminating Markov chain. All states are transient (i.e. if the chain leaves a state, it will not return to it), except one (=STOP) which is absorbing, i.e., where the chain terminates. The length of the sequence at termination is random.



A terminating Markov chain (2)

This architecture is used for length modeling in gene prediction (with a hidden Markov model).

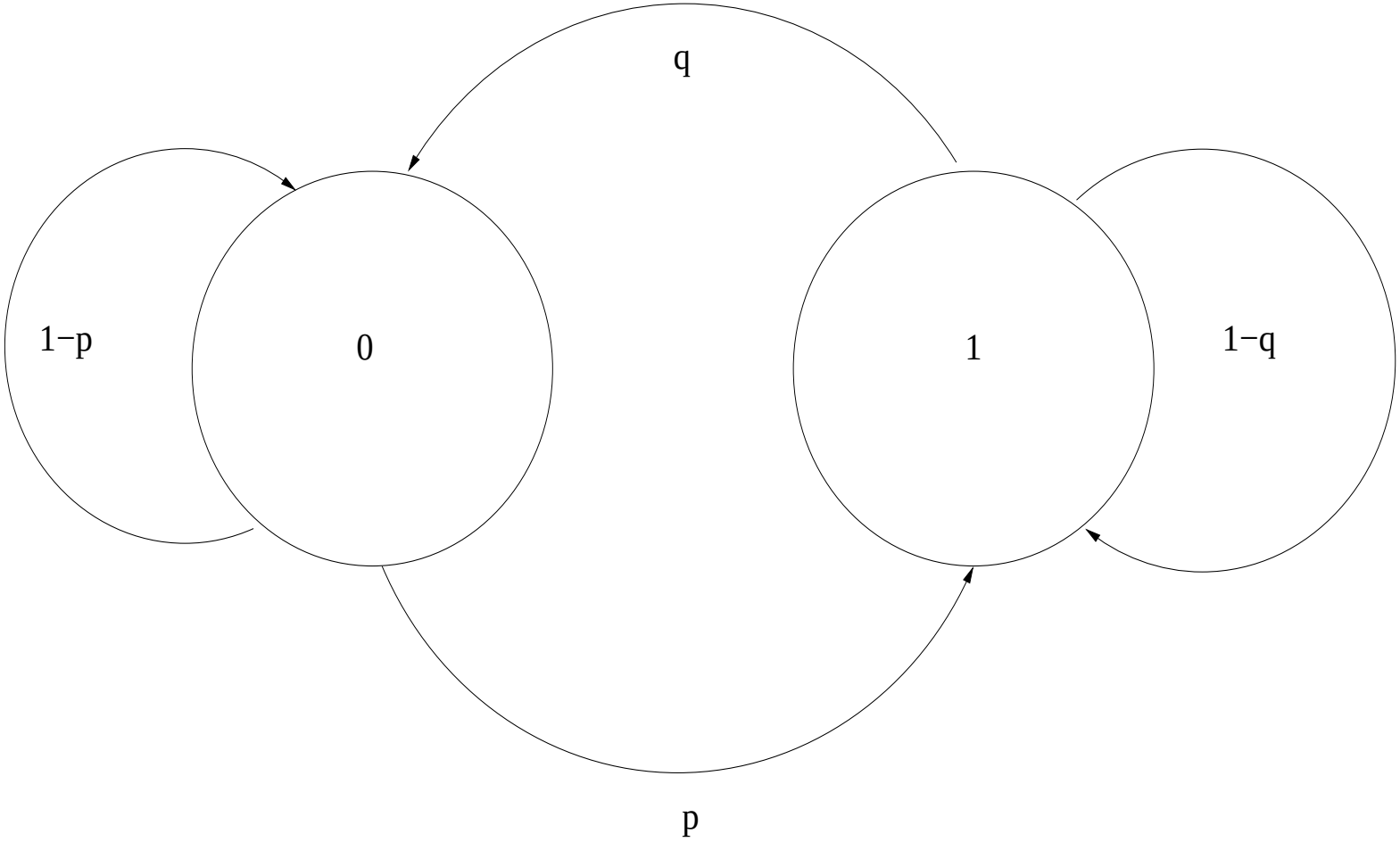
Munch, Kasper and Krogh, Anders: Automatic generation of gene finders for eukaryotic species, BMC bioinformatics, 2006, vol. 7,nr. 1, pp. 263,



KTH VETENSKAP OCH KONST
KTH Matematik

A binary Markov chain

Consider a binary M.c., i.e., each X_n is either 0 or 1.



A binary Markov chain

This Markov chain has the transition matrix

$$\begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},$$

where $0 \leq p \leq 1$ and $0 \leq q \leq 1$, i.e.,

$$p = Pr(X_1 = 1 \mid X_0 = 0), q = Pr(X_1 = 0 \mid X_0 = 1)$$

As in the case of Linnea's library, these do not depend on n,

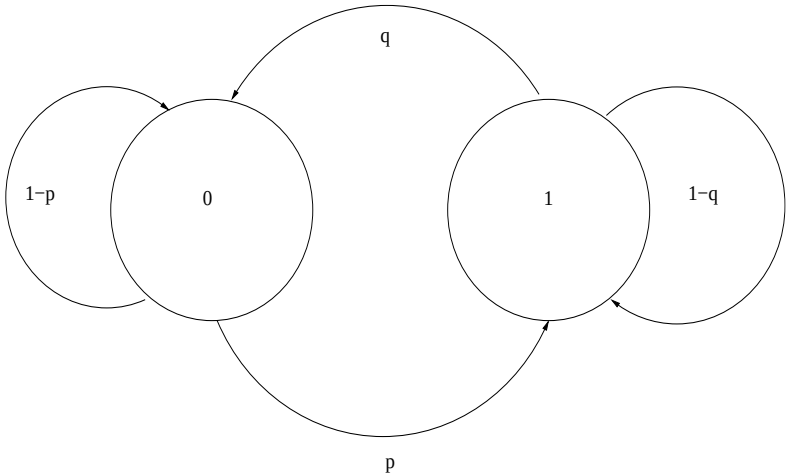
$$p = Pr(X_{n+1} = 1 \mid X_n = 0), q = Pr(X_{n+1} = 0 \mid X_n = 1)$$

The chain is called (time)homogenous.

KTH Matematik

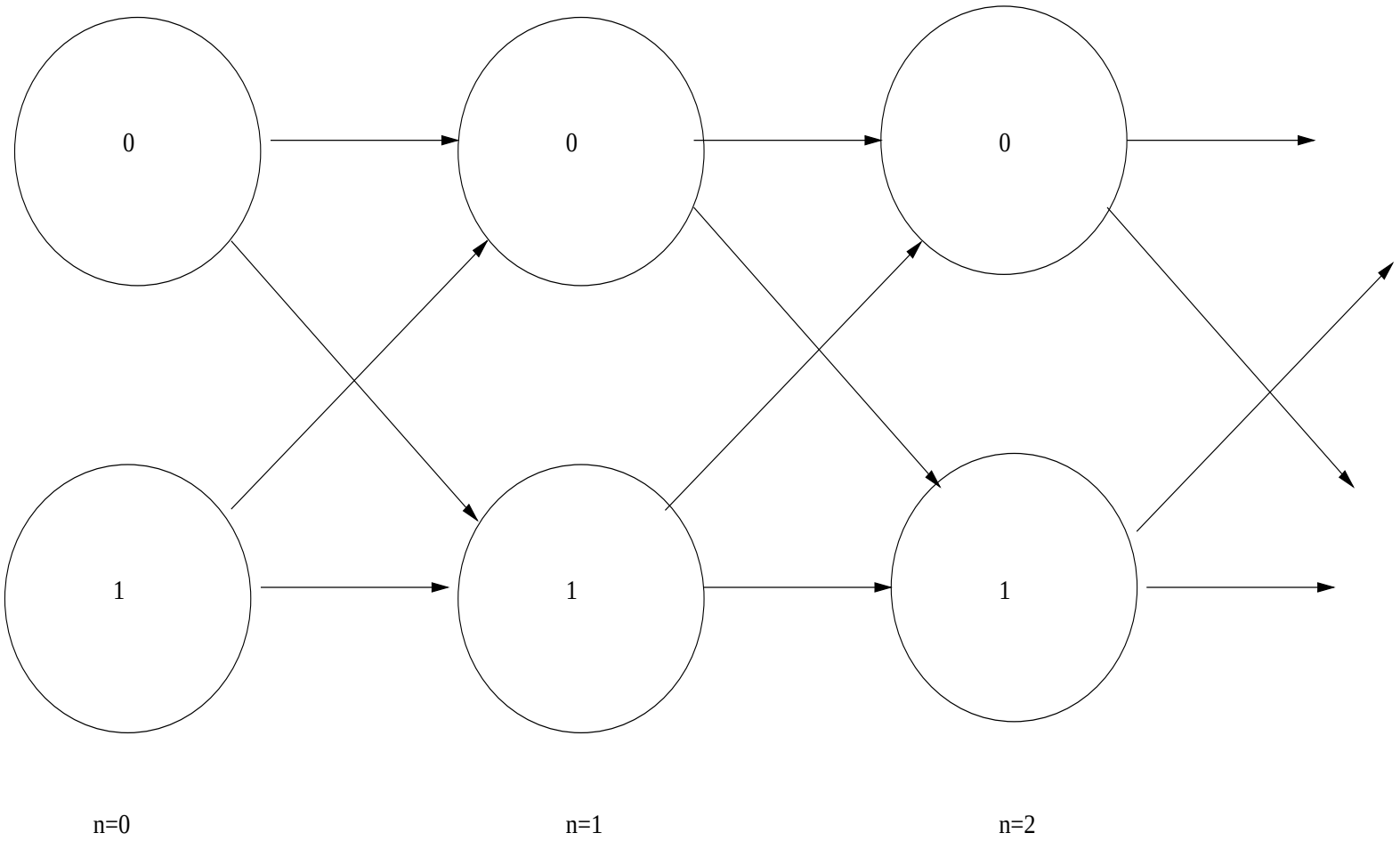
A binary Markov chain

This Markov chain has no state of termination and generates a sequence of bits.



Trellis

The *trellis* of the binary M.c..



Markov chains & genomic sequences

Learning outcomes:

- genomic sequence as a linear sequence
- Markov chain for DNA sequences
- transition probability
- probability of a sequence
- estimation of the transition probabilities

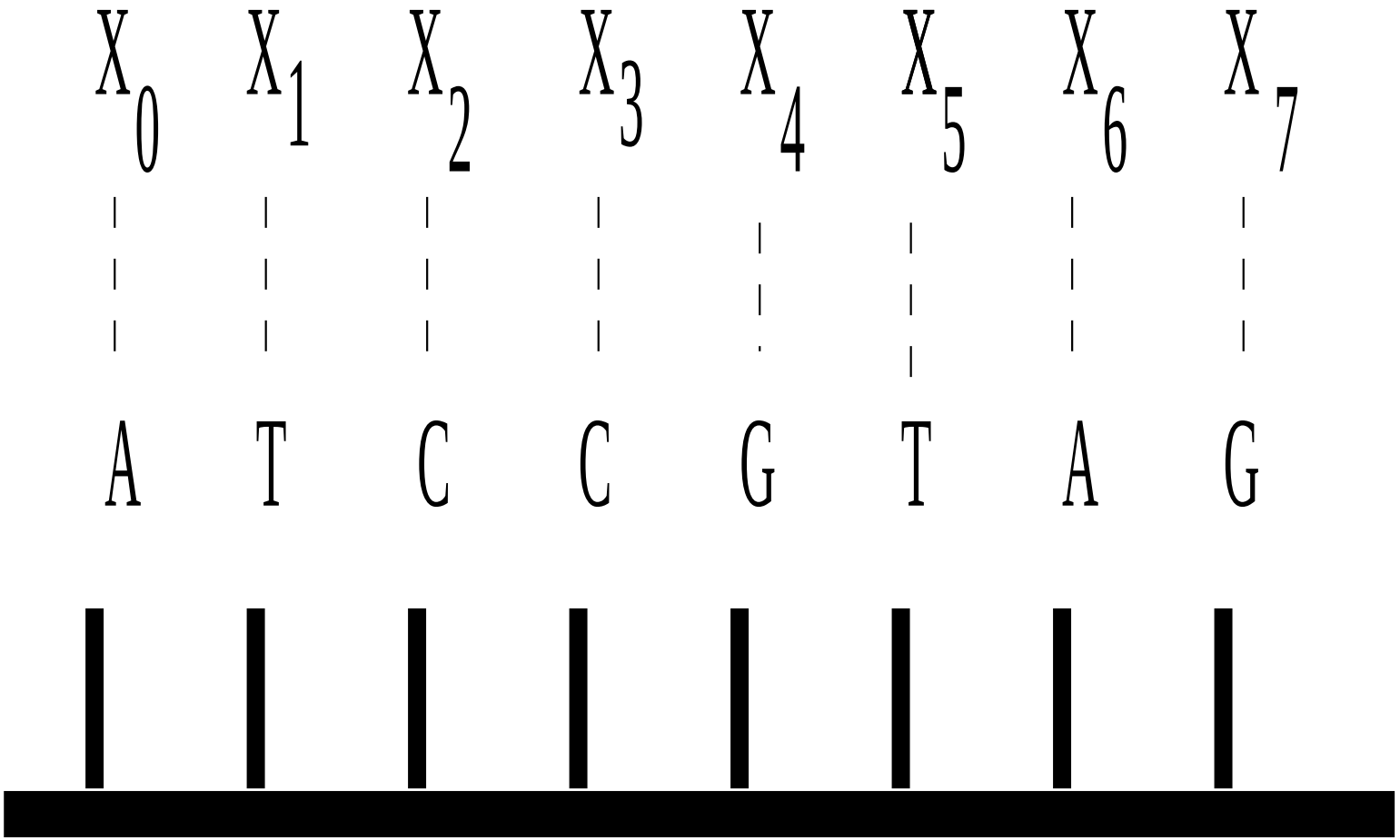
KTH Matematik

Markov chains and DNA

KTH Matematik

Markov chains for DNA sequences

We disregard for the current purposes the double helix structure and think of a linear sequence of nucleotides read from left to right. Here $X_1 = T$ says that the nucleotide at site 1 is $= T$. We shall now regard this as an outcome of a random variable.



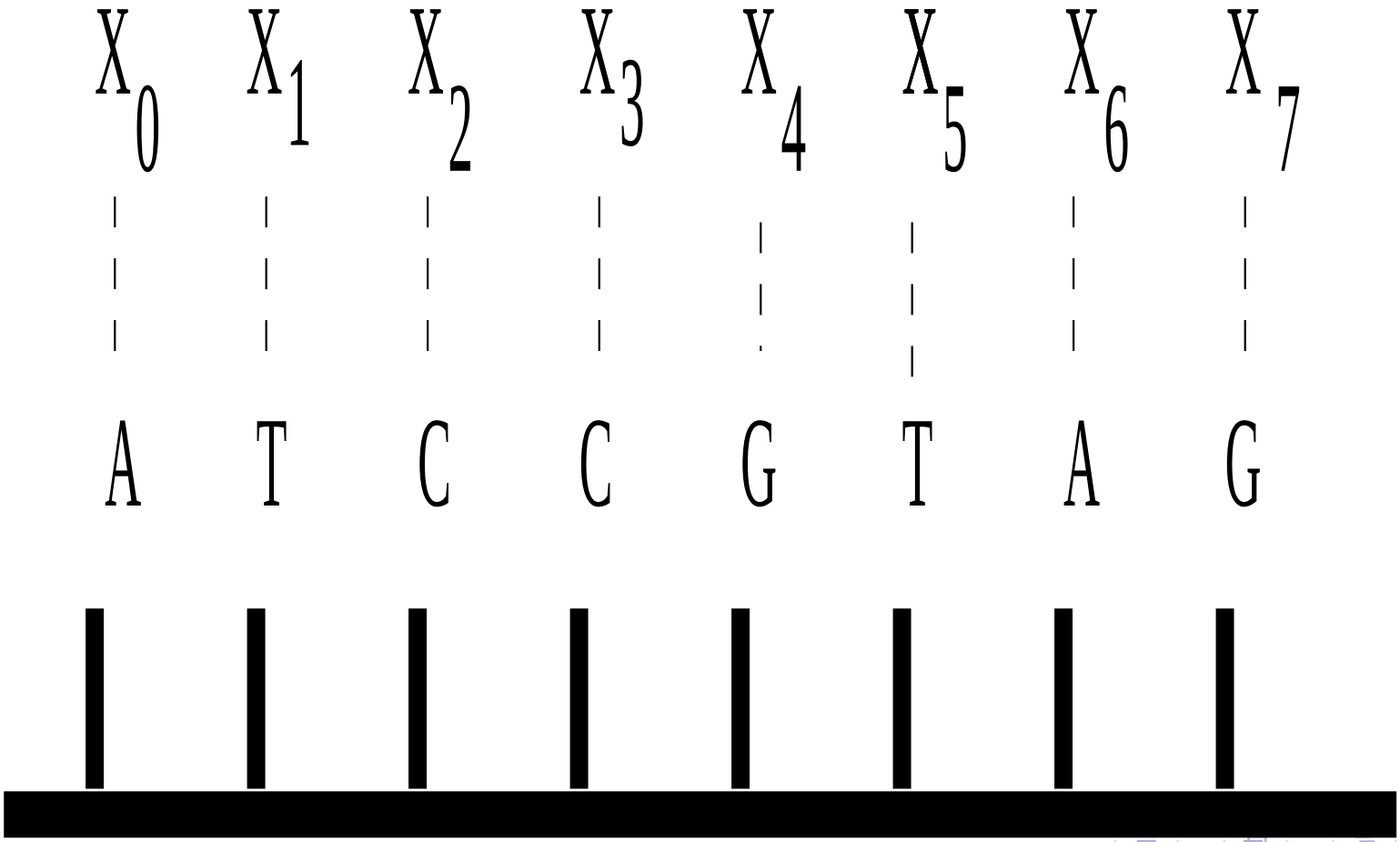
KTH Matematik

Markov chains for DNA sequences

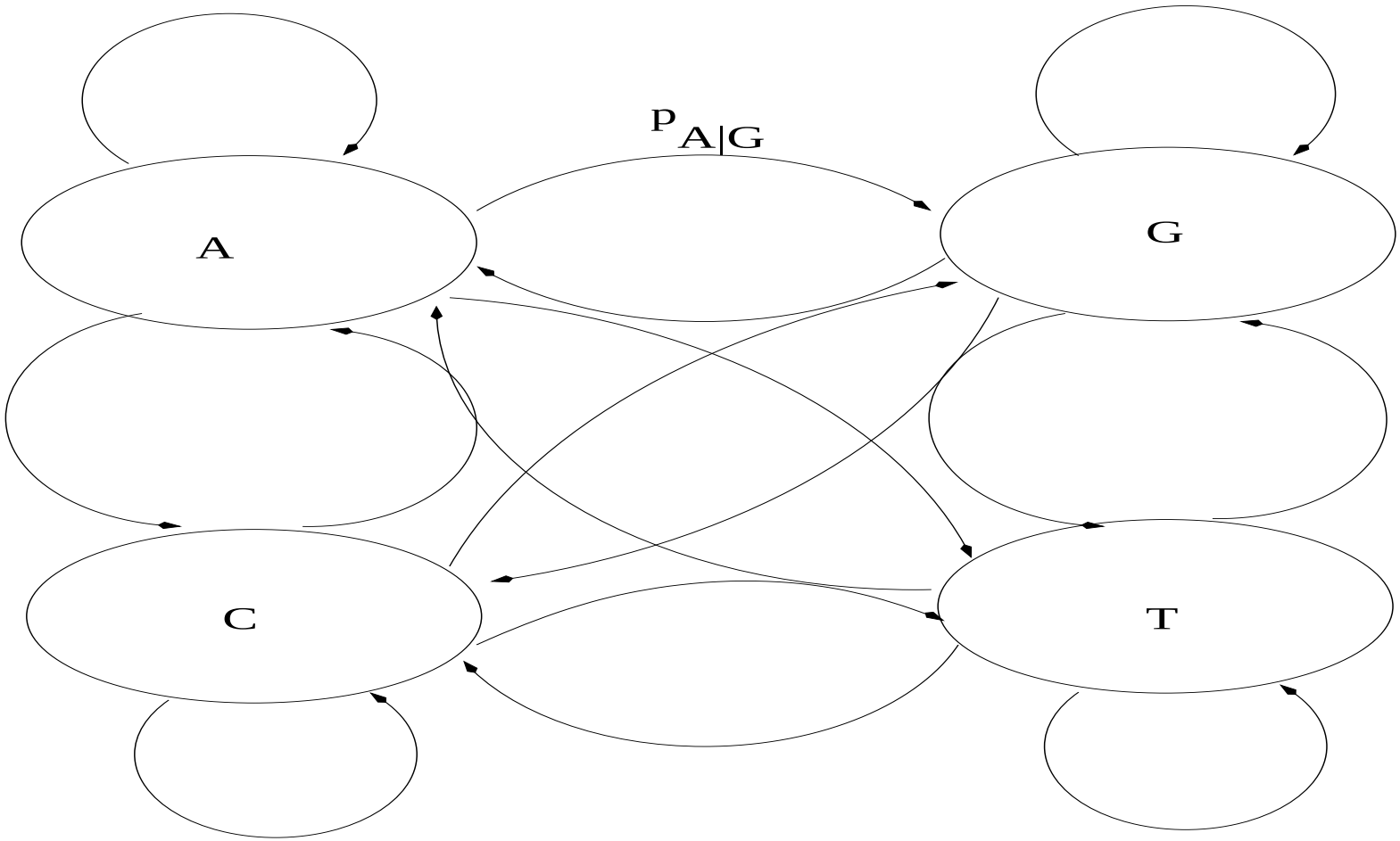
Now we set, e.g.,

$$Pr\left(X_{n+1} = A \mid X_n = T\right) = p_{\mathrm{T|A}}$$

(does not depend on n)



Transition graph



Markov Model: transition probabilities

	A	C	G	T
A	$p_{A\|A}$	$p_{A\|C}$	$p_{A\|G}$	$p_{A\|T}$
C	$p_{C\|A}$	$p_{C\|C}$	$p_{C\|G}$	$p_{A\|T}$
G	$p_{G\|A}$	$p_{G\|C}$	$p_{A\|G}$	$p_{G\|T}$
T	$p_{T\|A}$	$p_{T\|C}$	$p_{T\|G}$	$p_{T\|T}$

We have $p_{A|A} + p_{A|C} + p_{A|G} + p_{A|T} = 1$ and the same holds for all other rows.

KTH VETENSKAP OCH KONST
KTH Matematik

Markov Model: probability of a sequence

The probability of, e.g., the sequence $ATCGAT$ is now

$$Pr\left(X_0 = A, X_1 = T, X_2 = C, X_3 = G, X_4 = A, X_5 = T\right) =$$

$$Pr(A)p_{\mathrm{A}|\mathrm{T}}p_{\mathrm{T}|\mathrm{C}}p_{\mathrm{C}|\mathrm{G}}p_{\mathrm{G}|\mathrm{A}}p_{\mathrm{A}|\mathrm{T}}$$

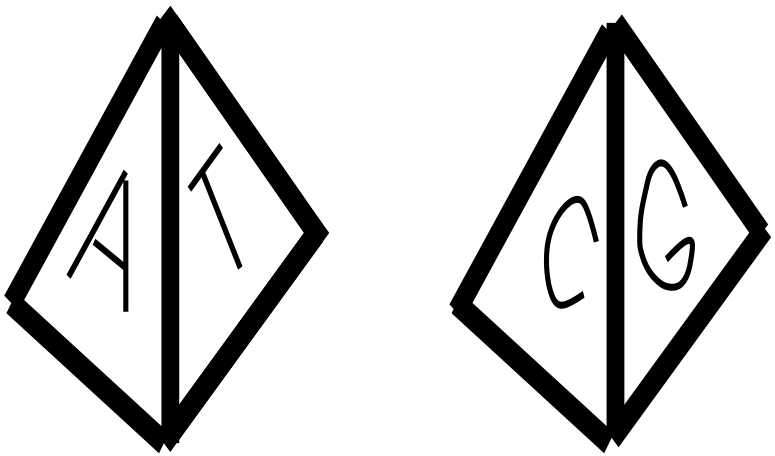
KTH Matematik

Recall: DNA dice model

We have earlier seen the DNA dice with independent tosses. There

$$Pr\left(X_0 = A, X_1 = T, X_2 = C, X_3 = G, X_4 = A, X_5 = T\right)$$

$$= Pr(A)Pr(T)Pr(C)Pr(G)Pr(A)Pr(T)$$



KTH Matematik

Markov Model: probability of a sequence

The Markov property means now that,

$$Pr\left(X_5 = T \mid X_0 = A, X_1 = T, X_2 = C, X_3 = G, X_4 = A\right) = p_{\mathrm{A|T}}$$

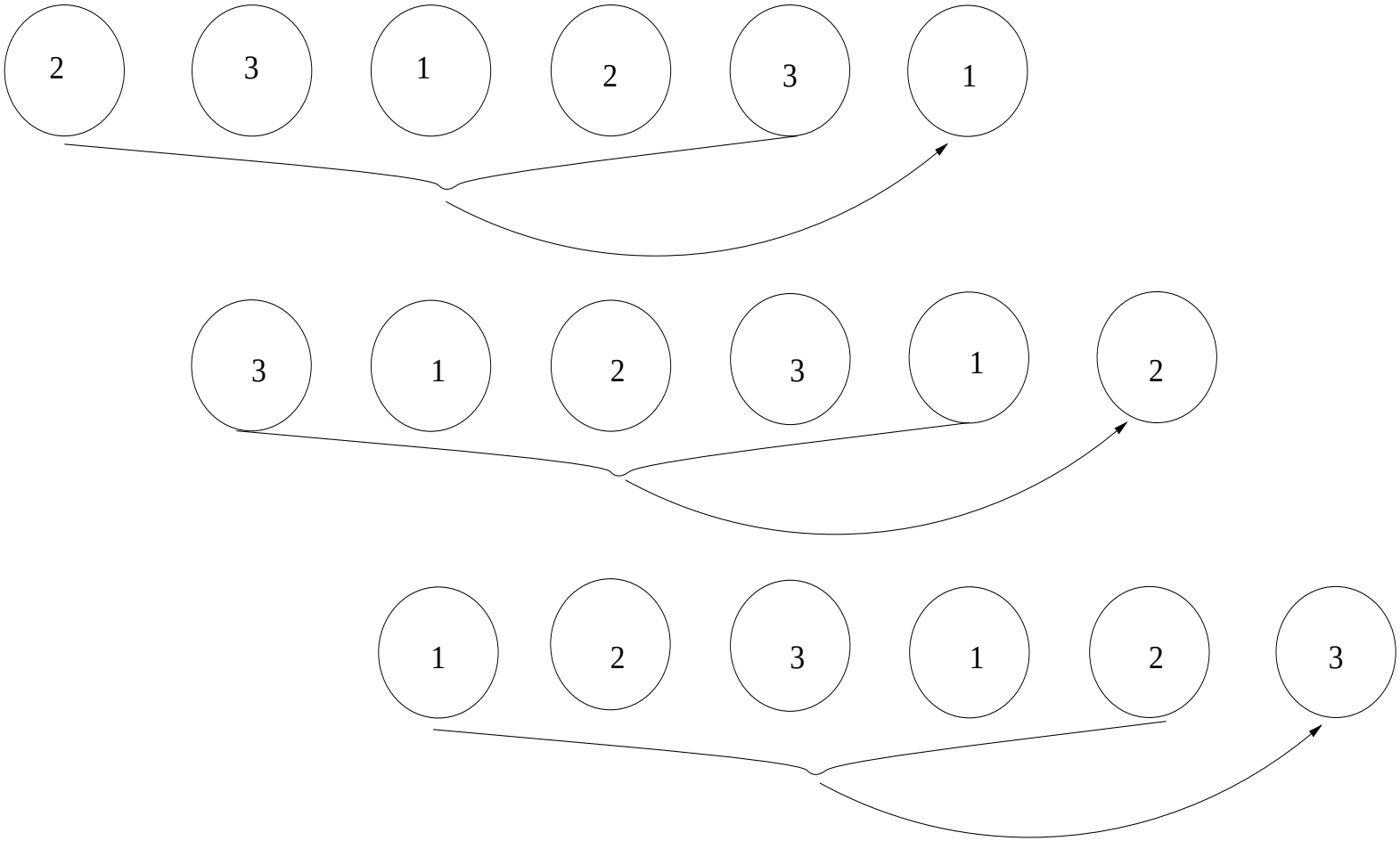
The **third order Markov property**

$$Pr\left(X_5 = T \mid X_0 = A, X_1 = T, X_2 = C, X_3 = G, X_4 = A\right)$$

$$= Pr\left(X_5 = T \mid X_2 = C, X_3 = G, X_4 = A\right)$$

(a memory of length $= 3$) has often turned out to be more useful. Fifth order memory is also found. These require larger transition matrices, probabilities for all transitions of triplets or quintets of nucleotides to one nucleotide.

Markov chain of fifth order (c.f. p. 370 in ZB)



Frame Dependent Markov Chains

A coding region is read as successive non-overlapping codons, which are instances of 3-symbol *words*. Since several different codons can code for the amino acid, the bases may have different importance depending on their position with respect to the codon partition.

KTH Matematik

Therefore a Markov chain of order k with three transition probability matrices, P_1, P_2, P_3 is considered. Here P_m has for $m = 1, 2, 3$ the entries

$$P\left(X_{3t+m} = j_{3t} | X_{3t-k+m} = j_{3t-k+m}, \ldots, X_{3t-1+m} = j_{3t-1+m}\right).$$

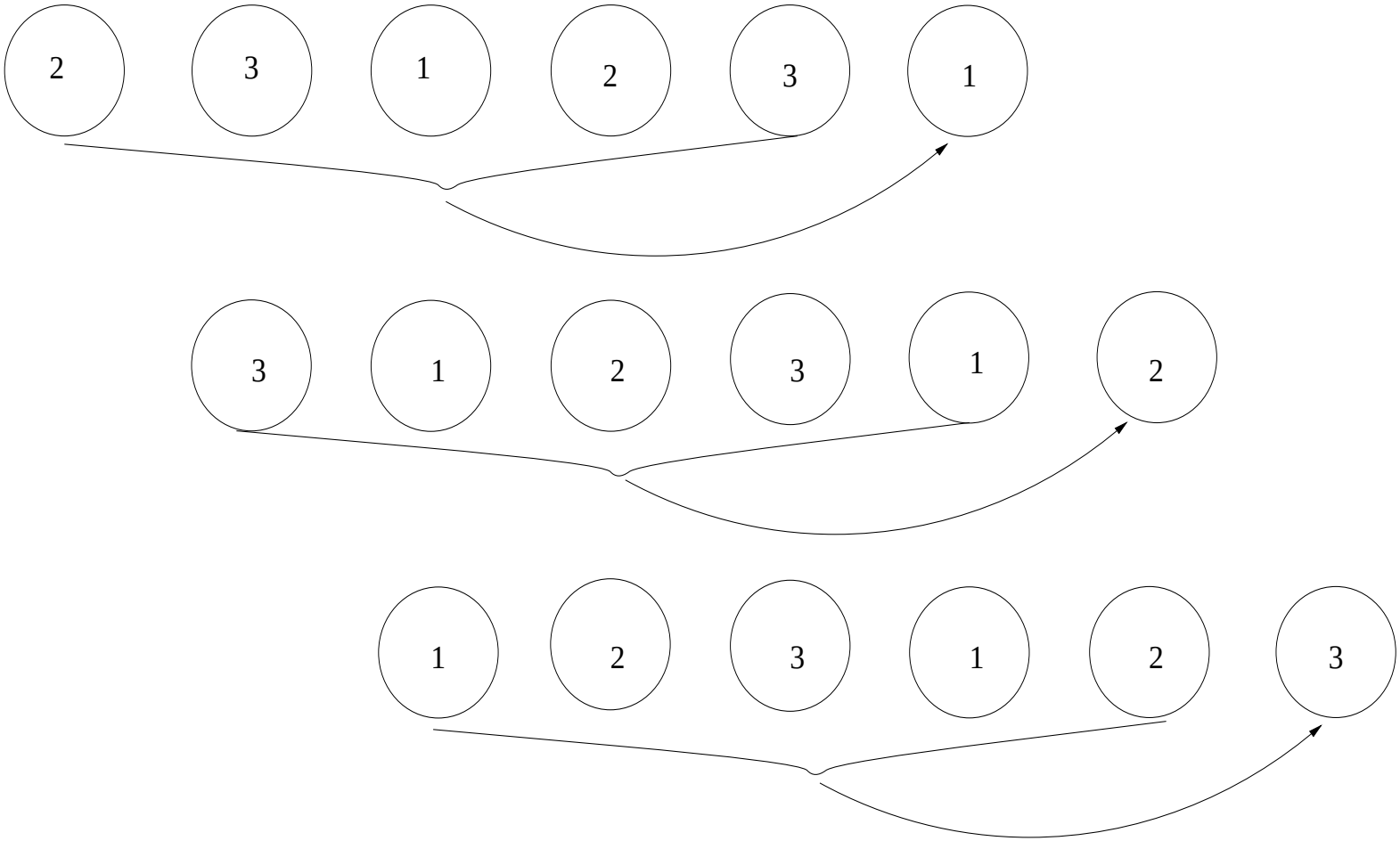
KTH Matematik

Frame Dependent Markov Chains GeneMark

The index m represents the position of the symbol inside the codon. This is illustrated for $k = 5$ in the figure below. The circles represent consecutive DNA bases, the numbers indicate the codon position. The three arrows carry with them the different transition probability matrices for each m. For each m the next base is generated by a distribution conditioned on the five previous bases and depending on the codon position m.

KTH Matematik

Frame Dependent Markov Chains GeneMark



Frame Dependent Markov Chains GeneMark

The kth order 3-phase Mc model is augmented by a first order Mc for non-coding regions and is also learned together with a kth order 3-phase Mc model for the other strand of the DNA sequence. Hence the task of learning this type of model from data requires the estimation of seven transition matrices with orders that are equal to $k = 5$ for coding regions of prokaryotic DNA.

KTH VETENSKAP OCH KONST
KTH Matematik

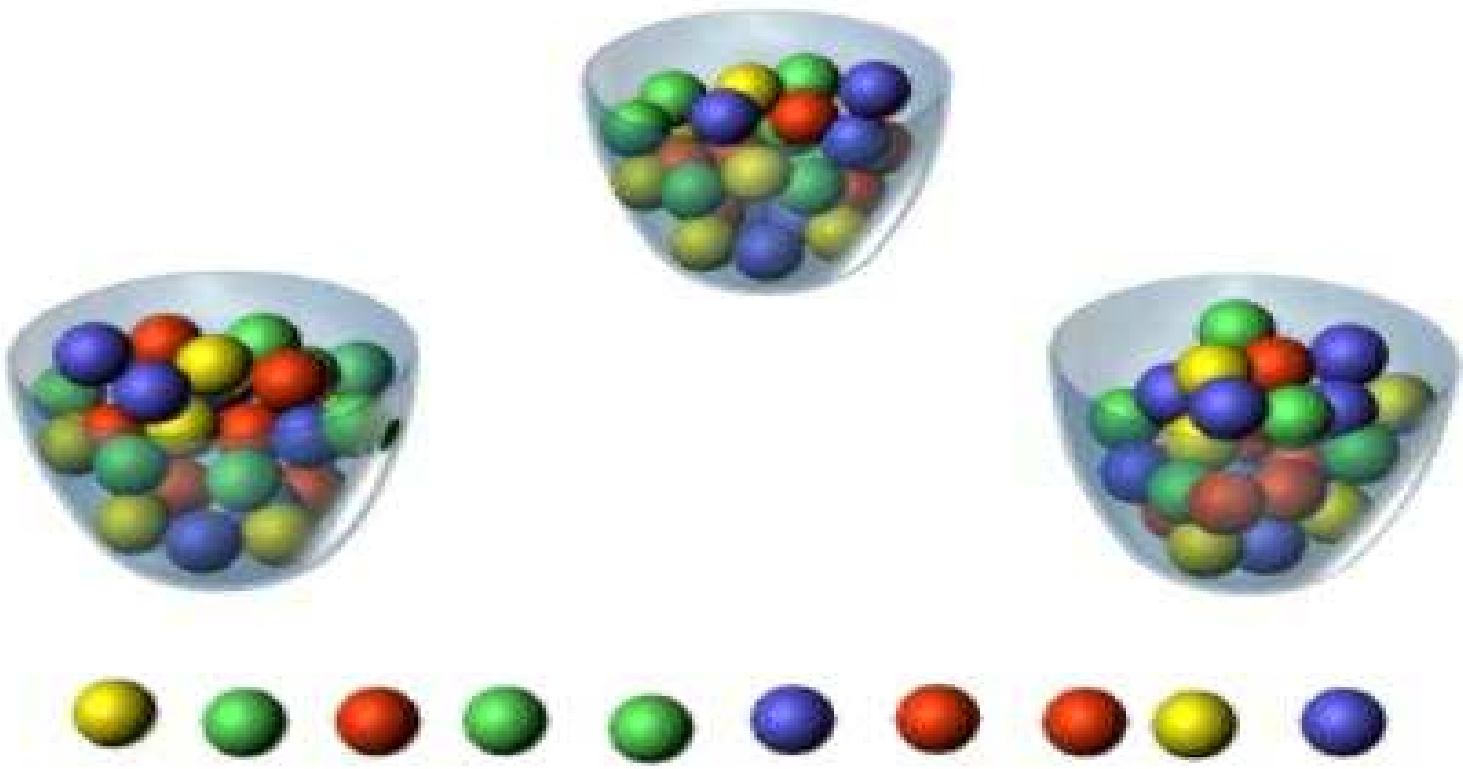
Frame Dependent Markov Chains GeneMark

The probability of an observed fragment is calculated and the **posterior probability** Pr(Model | Data) of each hypothesis about the coding region is computed. The relevant algorithms and their implementation in software are known as GeneMark[1]

[1] ZB pp. 368−370

KTH Matematik

HMM = Hidden Markov Model



KTH Matematik

HMM

Learning outcomes:

- hidden state, emission
- You can invent hidden Markov chains for your purposes by designing the architecture ! .

KTH Matematik

Contents

- First we discuss an examples of HMM: Modelling DNA heterogenity
- Then we give a general definition of HMM
- Gene finding
- Modelling protein families: Profile HMM

KTH Matematik

The topic: GC -content

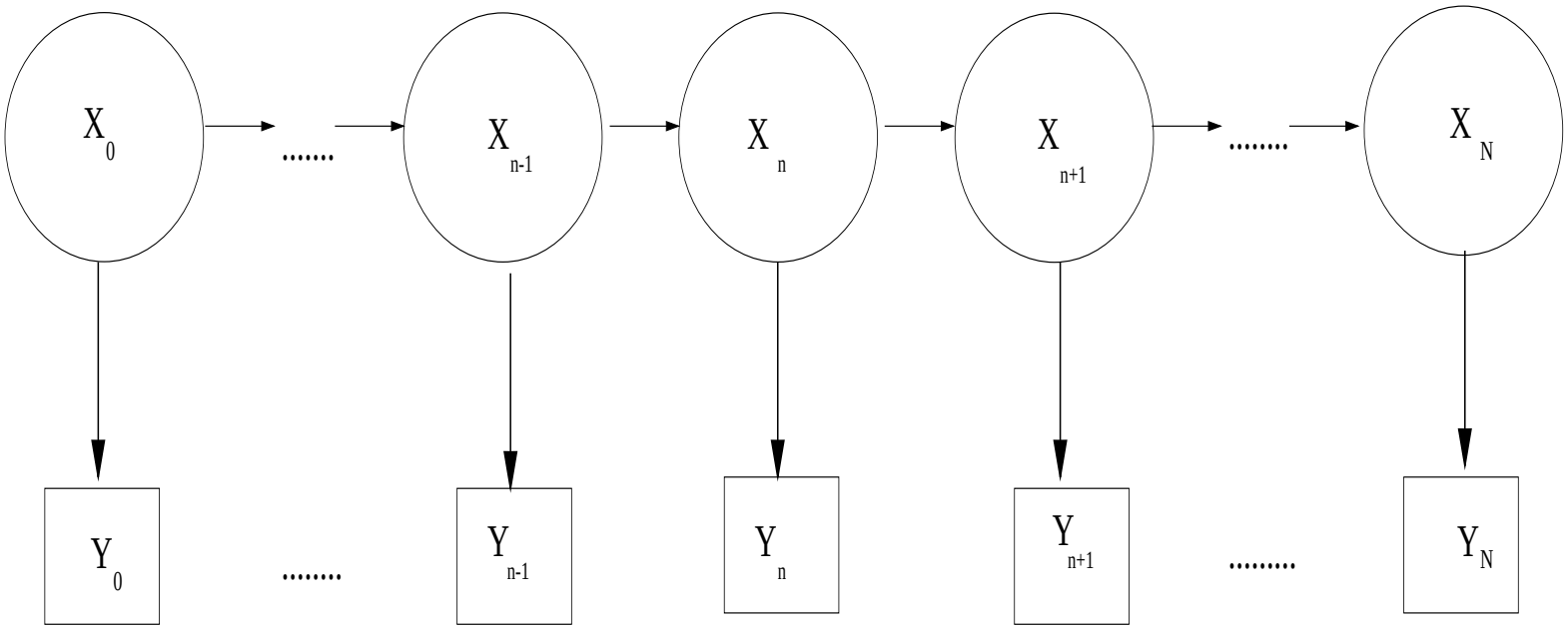
In molecular biology and genetics, **GC-content** (or guanine-cytosine content) is the percentage of nitrogenous bases on a DNA molecule that are either guanine or cytosine. This may refer to a specific fragment of DNA or RNA, or that of the whole genome. There is vast variation in GC-content, both amongst species and within a given species' own genome. Genes are often characterised by having a higher GC-content in contrast to the background GC-content for the entire genome[2].
We shall now try to propose a simple statistical model for this by means of hidden Markov models (HMM)

[2] ZB pp. 238 - 239, 386, 387

KTH Matematik

HHM for GC - content

- $X_n; n = 0, 1, 2, \ldots$ is a Markov chain with with the (binary) state space $S = \{0, 1\}$
- an emitted process $Y_n; n = 0, 1, 2, \ldots$ with values $\mathcal{O} = \{A, T, G, C\}$,

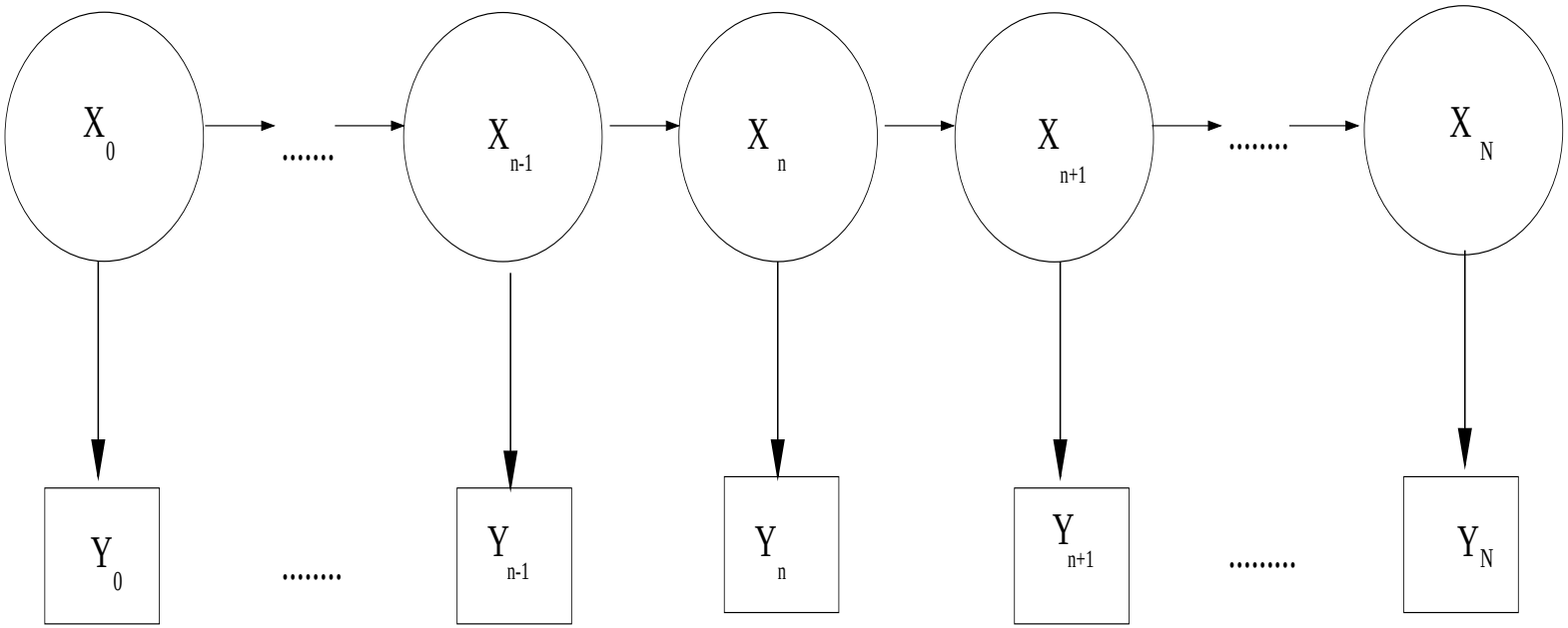


KTH Matematik

HHM for GC - content

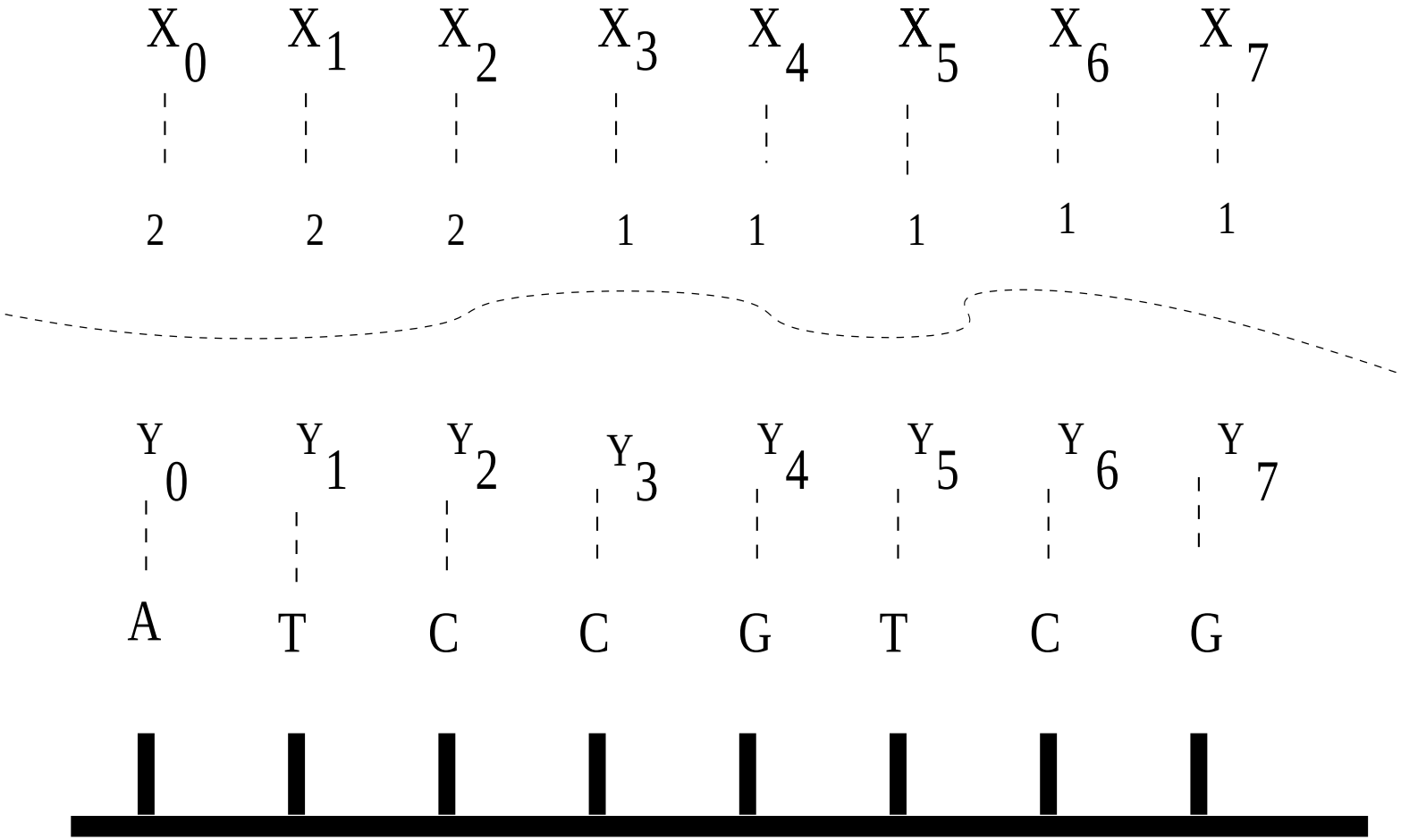
The probabilities of the emissions are conditional probabilites that depend on the state of X_n; $n = 0, 1, 2, \ldots$

$$Pr(Y_n \mid X_n)$$

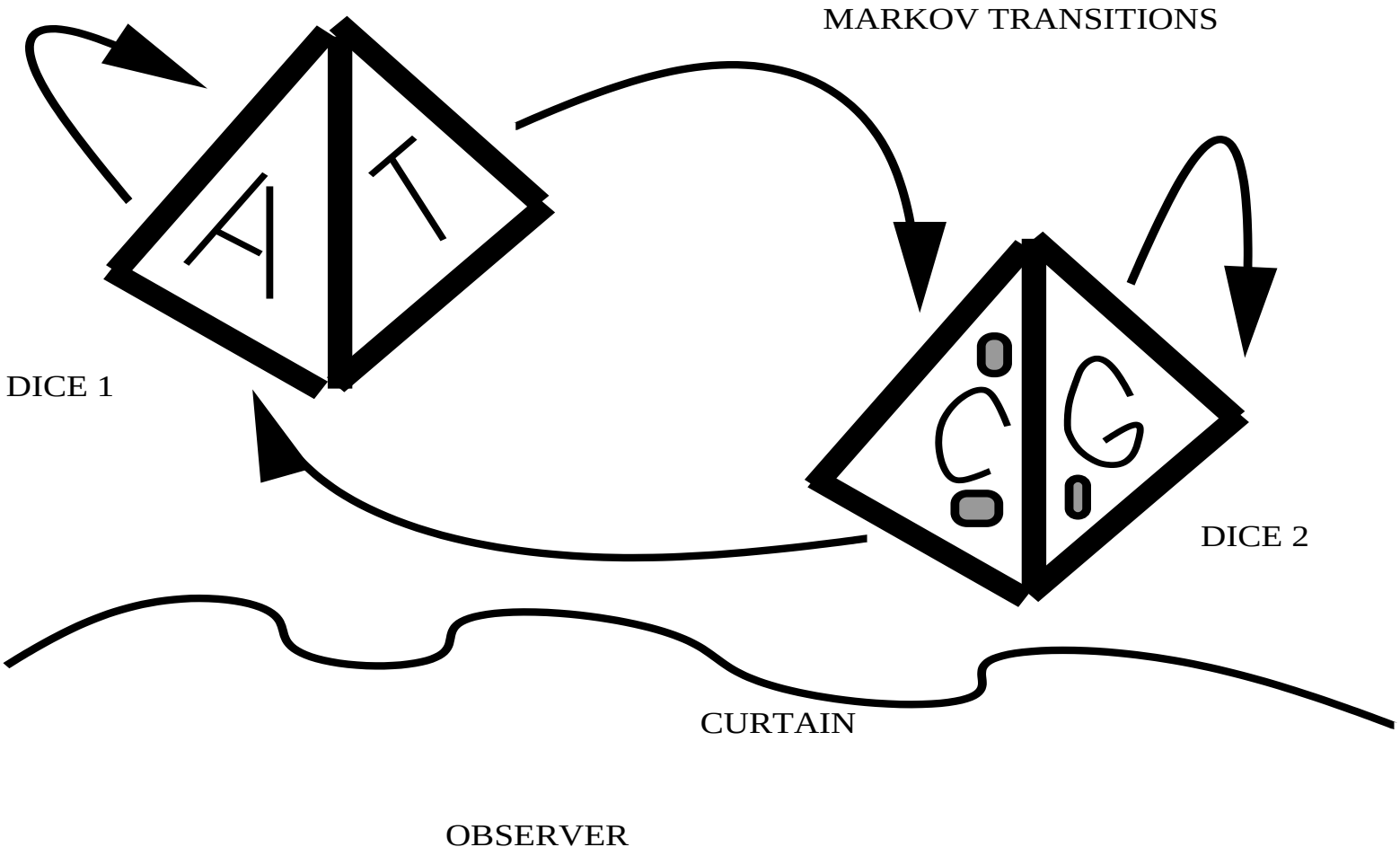


Modelling heterogeneity of DNA sequences

The M.c. $X_n; n = 0, 1, 2, \ldots$ is **hidden** and we see only the emissions $Y_n; n = 0, 1, 2, \ldots$.



Modelling heterogeneity of DNA sequences



The hidden Markov chain has the transition matrix

$$A = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix},$$

where $0 \leq p \leq 1$ and $0 \leq q \leq 1$. Connect this to the emission probability matrix

$$B = \begin{pmatrix} 1-\epsilon-w & w & \epsilon/2 & \epsilon/2 \\ \epsilon/2 & \epsilon/2 & r & 1-\epsilon-r \end{pmatrix},$$

where $\epsilon > 0$, $0 \leq w+\epsilon \leq 1$ and $0 \leq r+\epsilon \leq 1$.

KTH Matematik

Modelling heterogeneity of DNA sequences

The emission probability matrix

$$B = \begin{pmatrix} 1-\epsilon-w & w & \epsilon/2 & \epsilon/2 \\ \epsilon/2 & \epsilon/2 & r & 1-\epsilon-r \end{pmatrix}.$$

is read like $r = P(Y = G \mid X = 1)$, $w = P(Y = T \mid X = 0)$ e.t.c. .

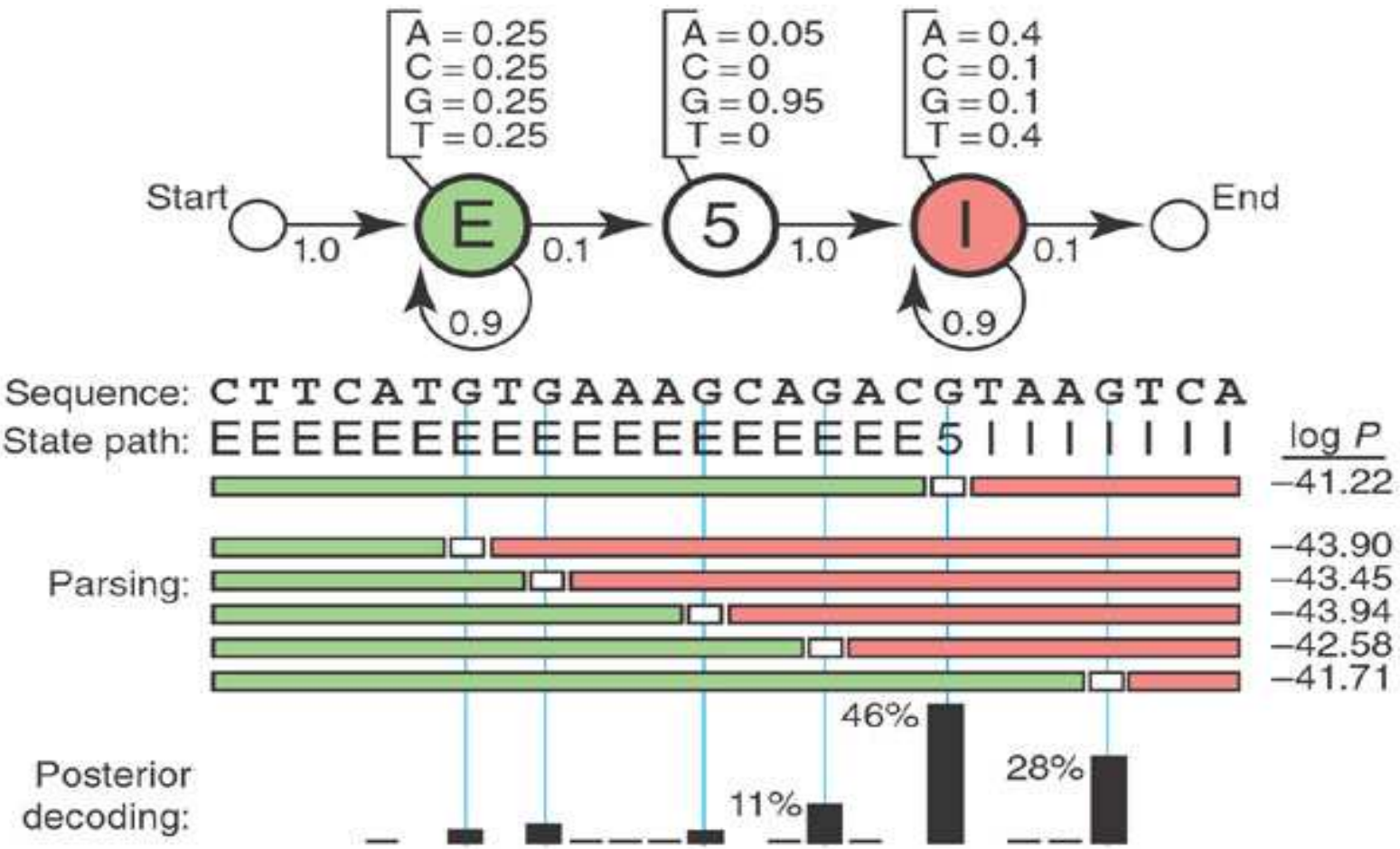
KTH Matematik

Modelling heterogeneity of DNA sequences

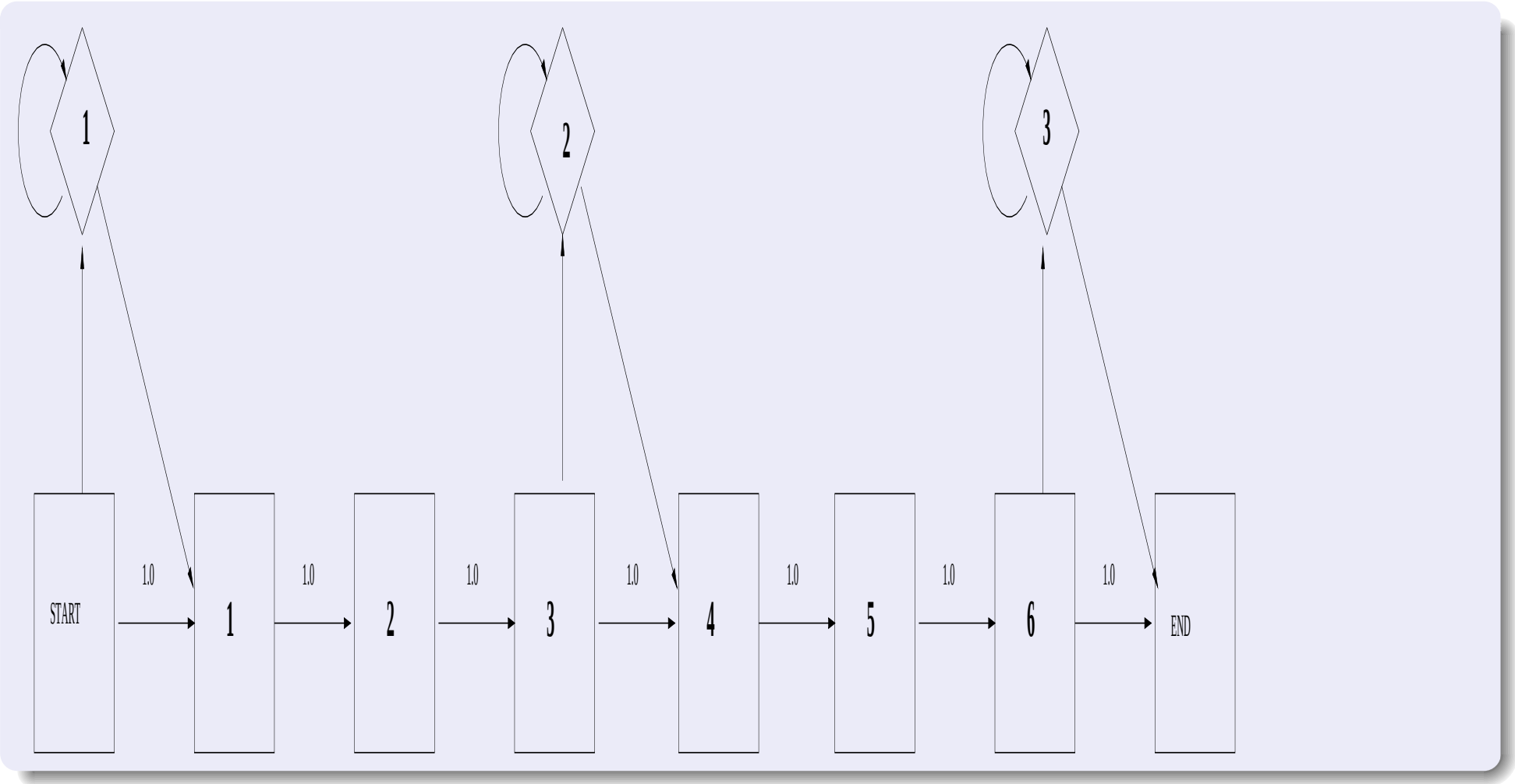
The value of p is chosen close to 0 and q is taken close to 0. Hence the state 0 of the hidden Markov chain persists, once the chain has entered it, and GC has a high probability of being emitted, assuming that ϵ is small, thus generating DNA sequences with GC-rich segments (and AT-rich segments, when the state is 1).

KTH Matematik

S.R. Eddy:What is a hidden Markov model? Nature Biotechnology 22, 1315 - 1316 (2004)



Meta MEME HMM



Motif-based hidden Markov models (HMMs) of families of related biosequences. The program takes as input a set of DNA or protein motif models constructed by MEME and produces as output a single HMM

KTH Matematik

Hidden Markov Models (HMM)

HMM is a model for a sequence of symbols from an alphabet $\mathcal{O} = \{o_1, o_2, \ldots o_K\}$. The model uses the idea of a hidden sequence of state transitions.
HMM has a definition with parts I−III.

KTH Matematik

Hidden Markov Models (HMM) I

(I) **Hidden Markov Chain** $\{X_n\}_{n=0}^{\infty}$ is a Markov chain assuming values in a finite state space $\mathcal{S}= \{1, 2, \ldots, J\}$ with J states. The time-homogeneous conditional probabilities are

$$a_{i|j} = P\left(X_n = j | X_{n-1} = i\right), n \geq 1, i, j \in S$$

and the transition probability matrix is

$$A = \left(a_{i|j}\right)_{i=1,j=1}^{J,J}$$

Hidden Markov Models (HMM) I

A matrix

$$A = \left(a_{i|j}\right)_{i=1,j=1}^{J,J}$$

with the constraints

$$a_{i|j} \geq 0, \sum_{j=1}^{J} a_{i|j} = 1.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Hidden Markov Models (HMM) II

(II) **Emitted Random Process** A random process $\{Y_n\}_{n=0}^{\infty}$ with a finite state space $\mathcal{O} = \{o_1, o_2, \ldots o_K\}$, where K can be $\neq J$. The processes $\{Y_n\}_{n=0}^{\infty}$ and $\{X_n\}_{n=0}^{\infty}$ are for any fixed n related by the conditional probability distributions

$$b_j(k) = P(Y_n = o_k | X_n = j).$$

KTH Matematik

Hidden Markov Models (HMM) II

We set

$$B = \{b_j(k)\}_{j=1,k=1}^{J,K}$$

and call this the *emission probability matrix*. This satisfies

$$b_j(k) \geq 0, \sum_{k=1}^{K} b_j(k) = 1.$$

KTH VETENSKAP OCH KONST
KTH Matematik

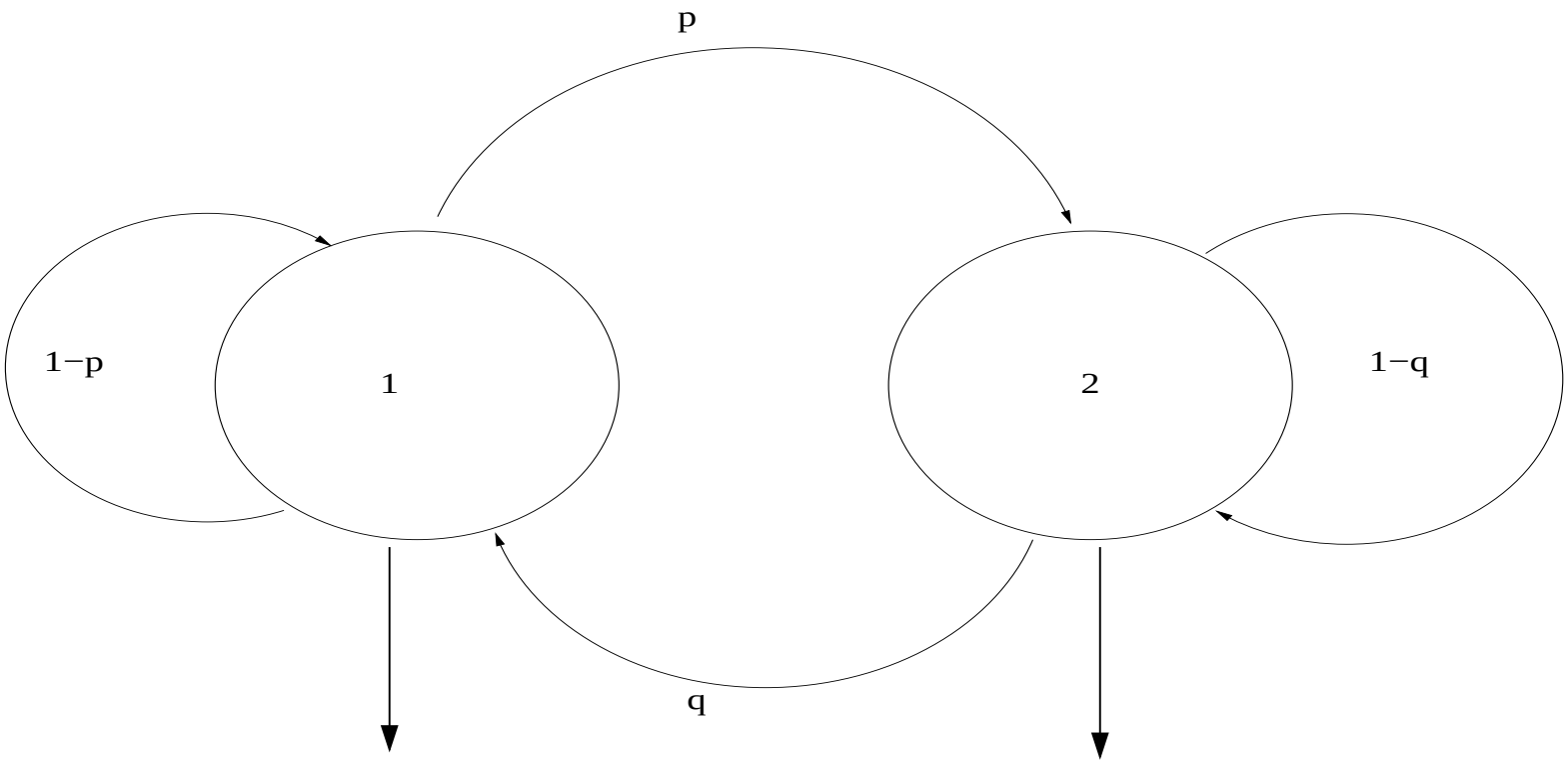
Hidden Markov Models (HMM) III

(III) **Probabilty of an emitted sequence given a state path** For any sequence of states $j_0 j_1 \dots j_n$ the probability of the sequence $o_0 o_1 \dots o_n$ is

$$P\left(Y_0 = o_0, \dots, Y_n = o_n \mid X_0 = j_0, \dots, X_n = j_n, B\right) =$$

$$\prod_{l=0}^{n} b_{j_l}\left(l\right).$$

KTH Matematik



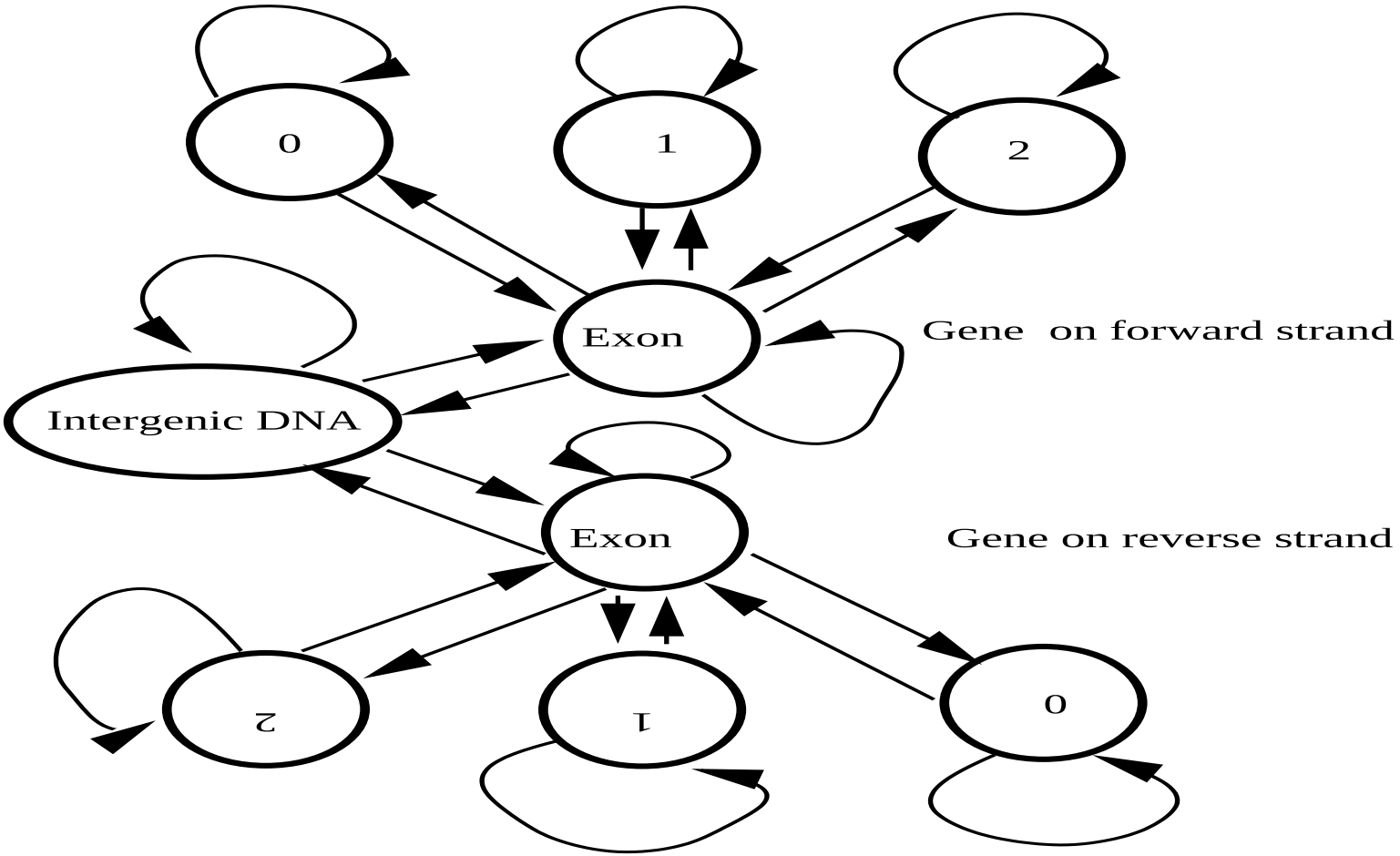
o	AT	CG
b(o|1)	1−w	w

o	AT	CG
b(o|2)	r	1−r

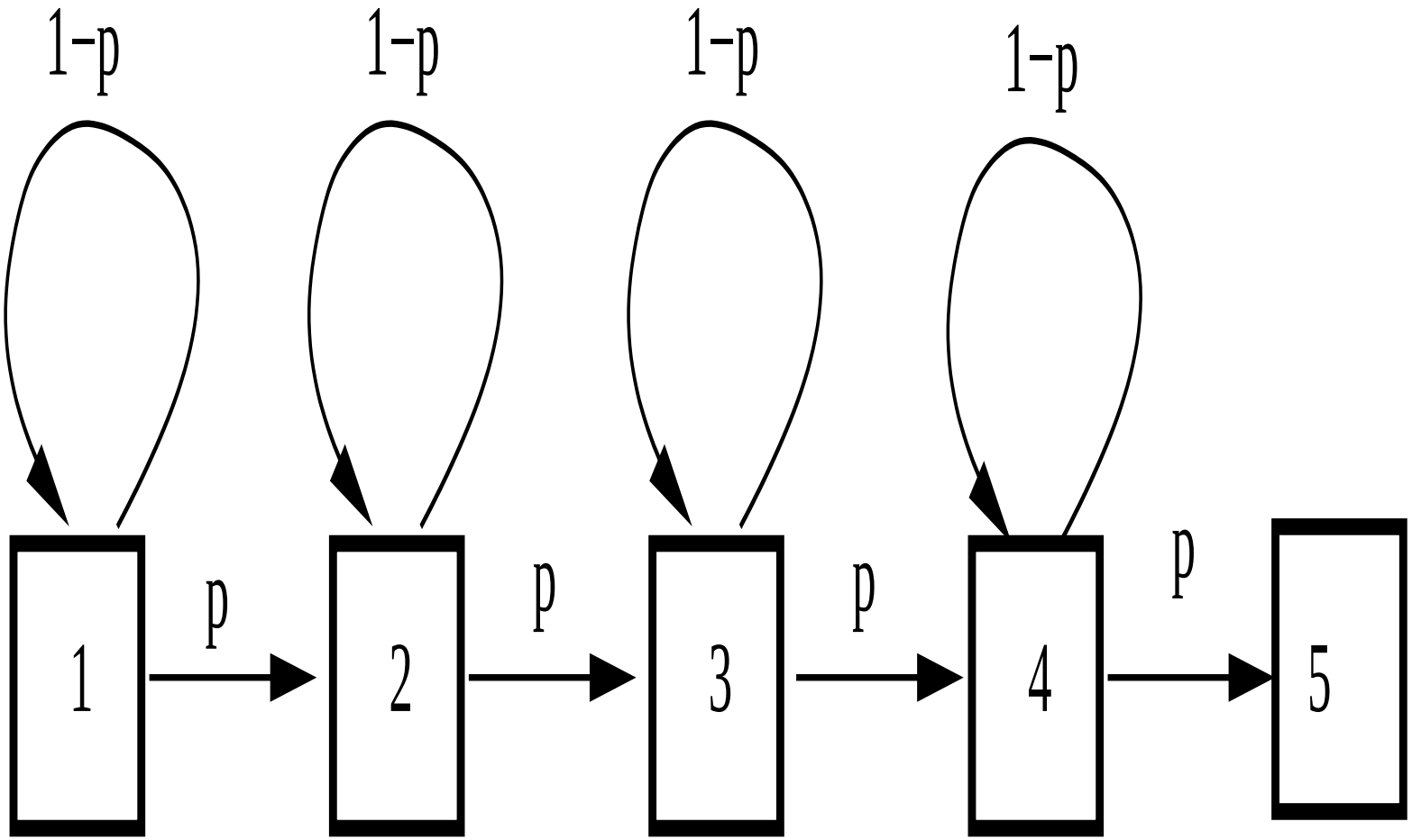
Gene Finding

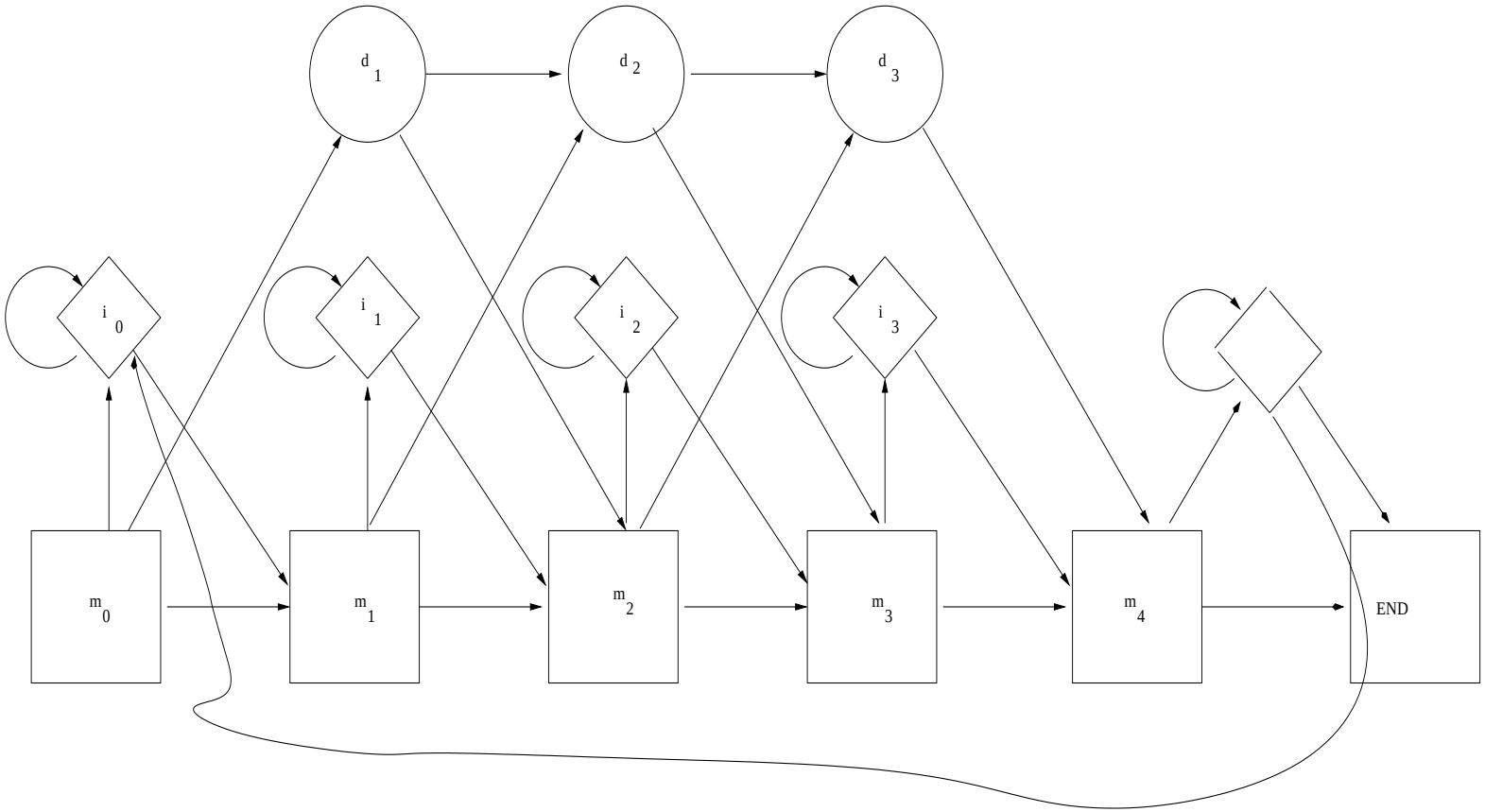
Hidden Markov models have been extensively used for modelling genes. Ab initio HMM gene finders for eukaryotes include software platforms like Genscan, Augustus, HMMgene, GeneMark.HMM-E, Genie, TigrScan and GlimmerHMM, Unveil and Exonomy, SNAP and others.
In gene finding, the states in the hidden Markov chain correspond to intergenic regions and gene structure elements, e.g. coding regions and introns. Each state emits nucleotides that constitute the DNA sequence in corresponding regions. The emissions of bases may be conditional on the occurrence of neighbouring bases within the sequence. This enables the HMM to model higher order dependencies of base frequencies.

KTH Matematik

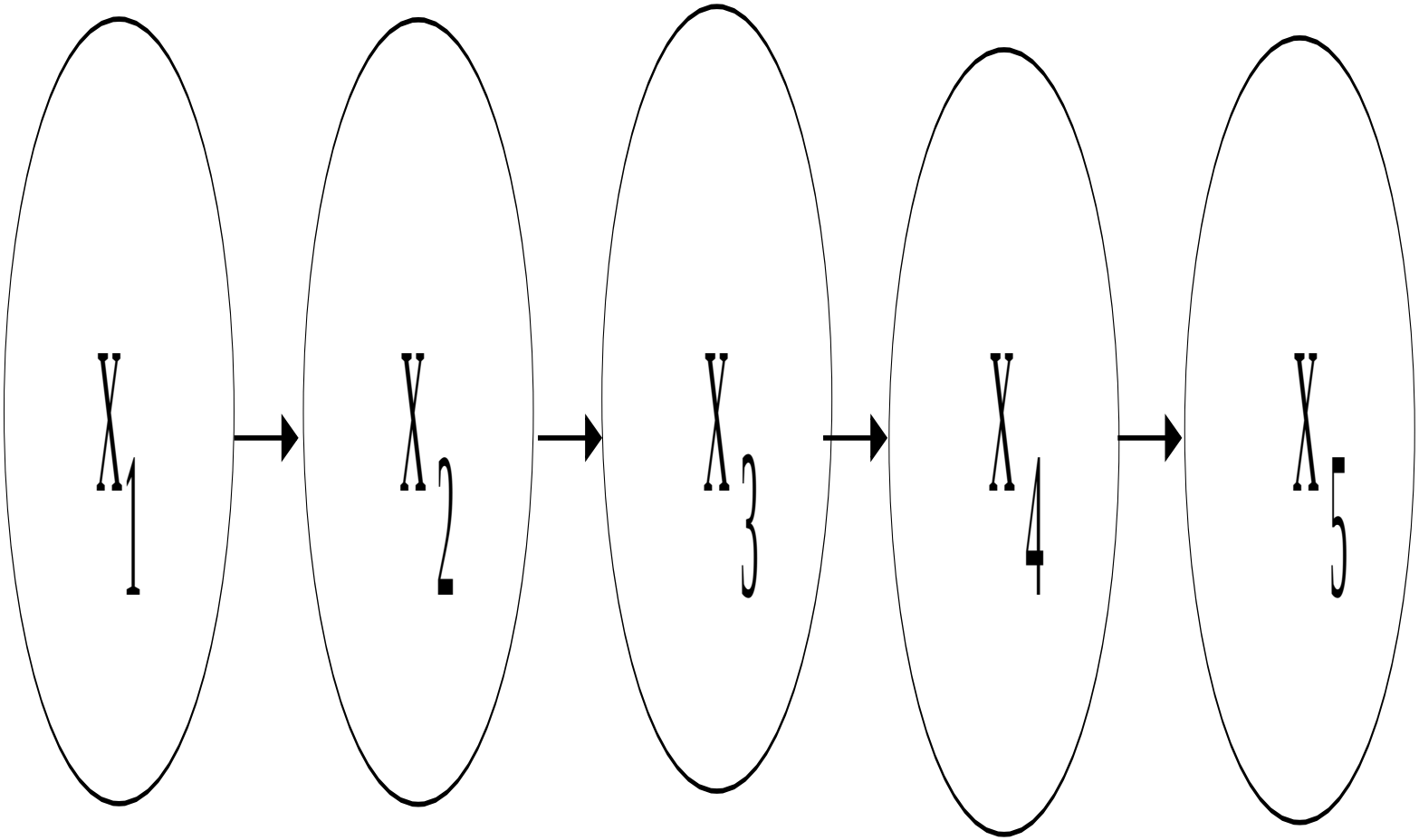


KTH Matematik

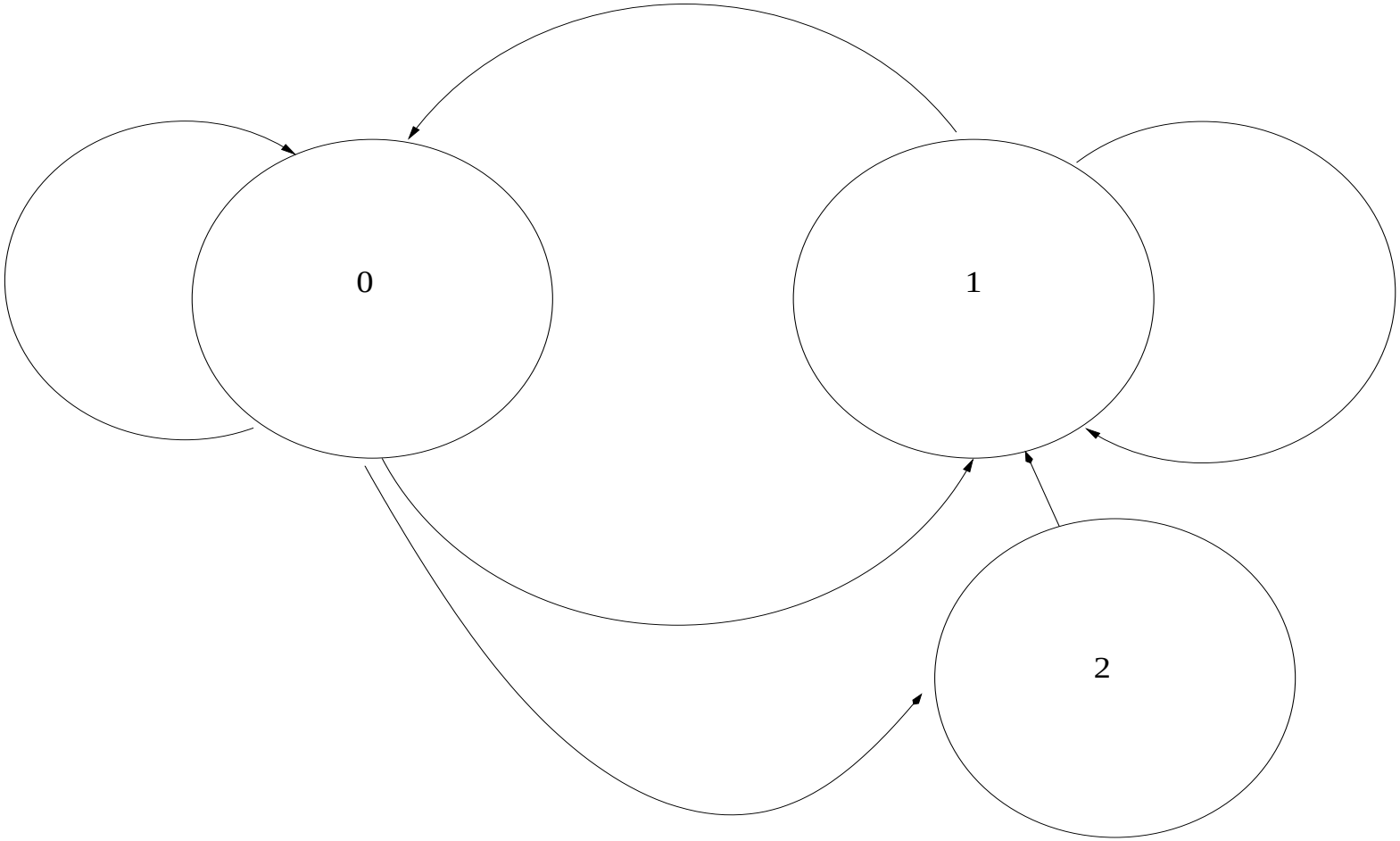

1-p
1-p
1-p
1-p
p
p
p
p
1
2
3
4
5



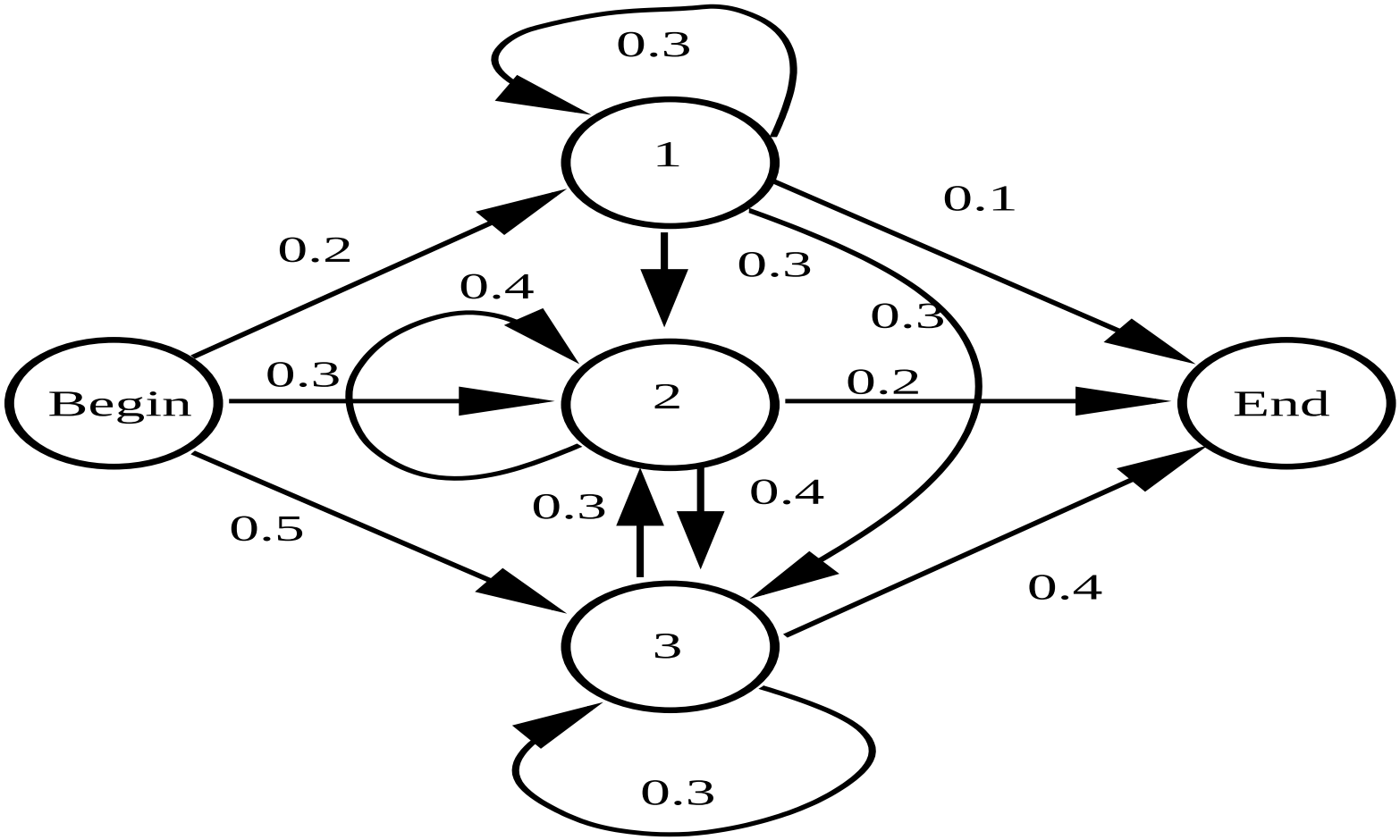
KTH Matematik



KTH Matematik



KTH Matematik

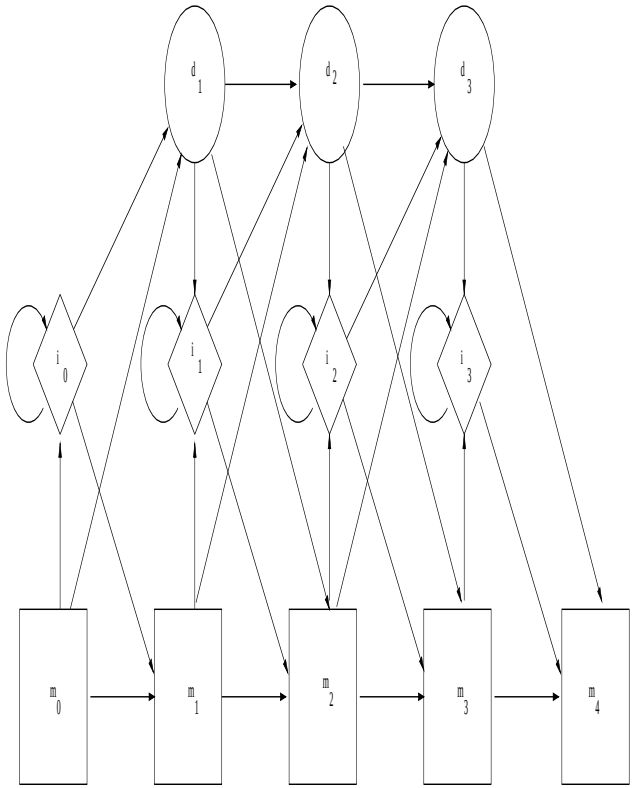

0.3
1
0.1
0.2
0.3
0.4
0.3
Begin
0.3
2
0.2
End
0.3
0.4
0.5
0.4
3
0.3

Prob(Data | Model)

Prob(Data | Model)

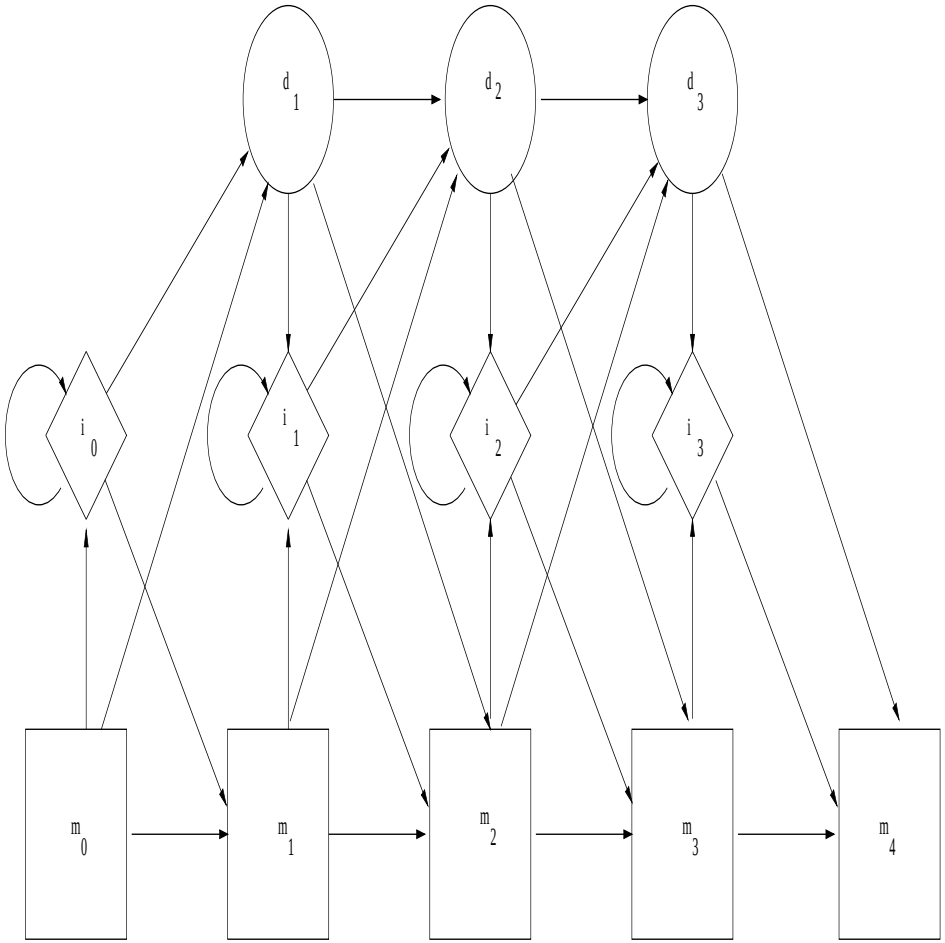
Hidden Markov Chains: Modeling Protein Families

The perhaps main application of Hidden Markov Chains in bioinformatics is detection of remote homologies using the Haussler-Krogh topology of the state space.



HMM: Profile HMM

One main application of HMM in bioinformatics is detection of remote homologies using the Profile HMM.



KTH Matematik

Modelling Protein Families

Proteins are categorized into families that share common function and evolutionary ancestry. Identifying distantly related homologs is a difficult problem, primarily because sequence identity between them is sparse. When these families are considered in the context of a homology search, it becomes possible to identify amino acid variability, which is common to the family members. Distantly related proteins can be found even with low sequence identity, if the similarities and differences are common to the family members.

KTH Matematik

Modelling Protein Families

This type of analysis is powerful because the function of divergent proteins is conserved through evolution even though sequence elements are free to change in some areas. Family-based searches take advantage of the fact that individual members of a family serve as examples of how tolerant this class of proteins is to change, and where. HMMs are one of several family-based search methods.

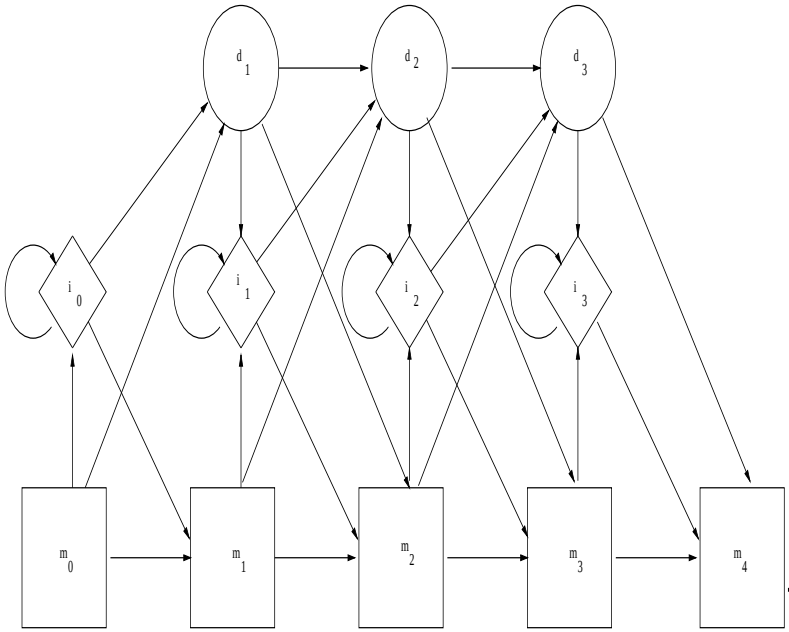
KTH Matematik

Modelling Protein Families

HMM is used to statistically describe a protein family's consensus sequence. This statistical description can be used for sensitive and selective database searching.

KTH VETENSKAP OCH KONST
KTH Matematik

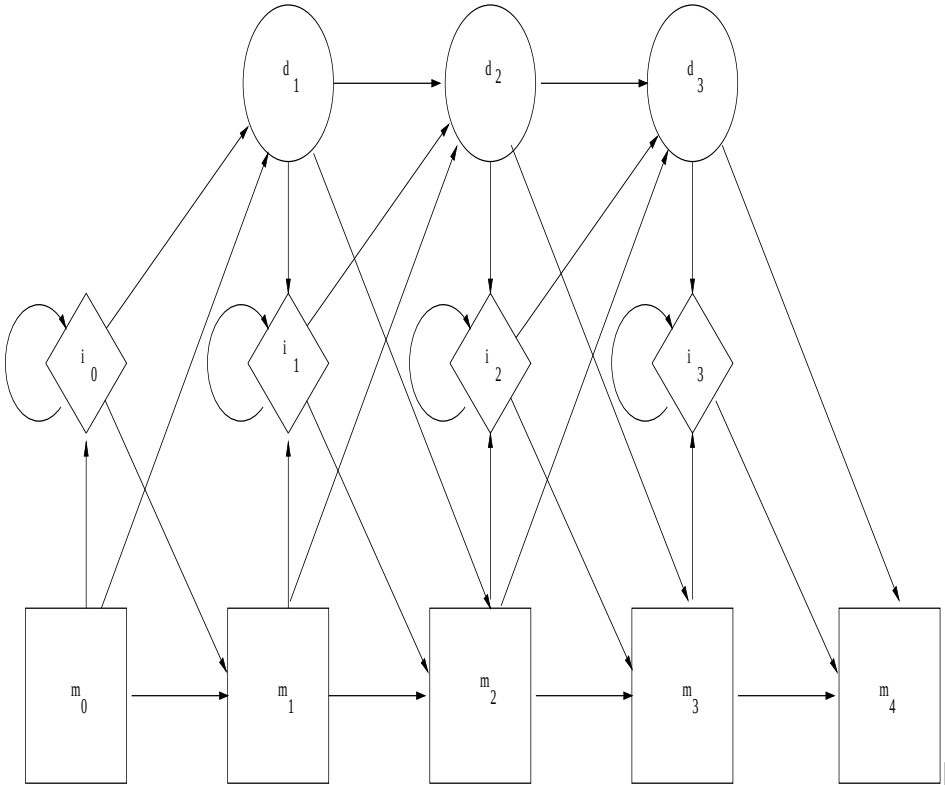
Modelling Protein Families



The model consists of a linear sequence of nodes with a begin state and an end state. although a typical model can contain hundreds of nodes. Each node between the beginning and end states corresponds to a column in a multiple alignment.

KTH VETENSKAP OCH KONST
KTH Matematik

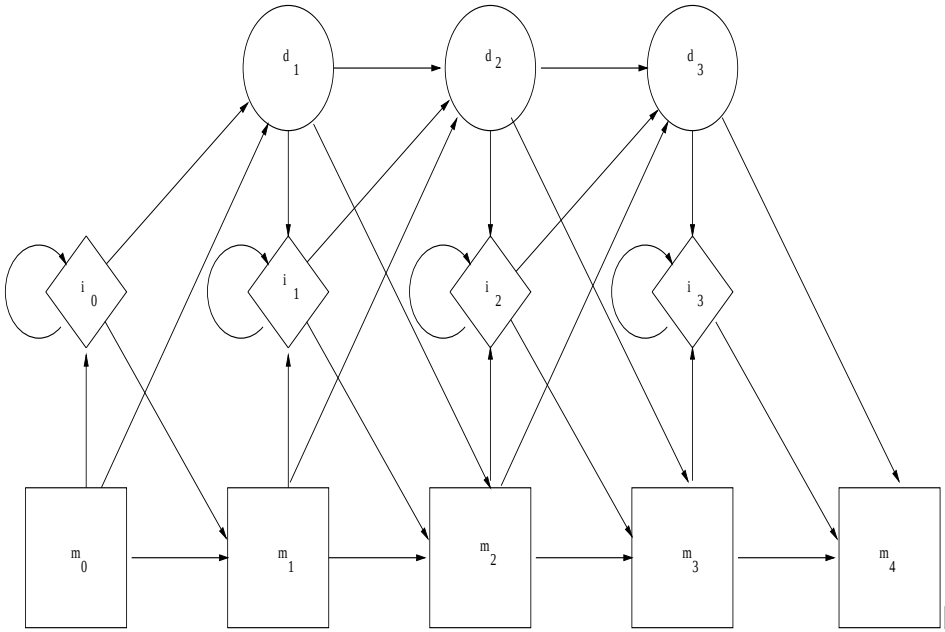
Modelling Protein Families



Each node in an HMM has a match state (M), insert state (I) and delete state (D) with position-specific probabilities for transitioning into each of these states from the previous node.

KTH VETENSKAP OCH KONST
KTH Matematik

Modelling Protein Families



In addition to a transition probability, the match state also has position-specific probabilities for emitting a particular residue.The insert state has probabilities for inserting a residue at the position given by the node. There is also a chance that no residue is associated with a node. That probability is indicated by the probability of transitioning to the delete state.

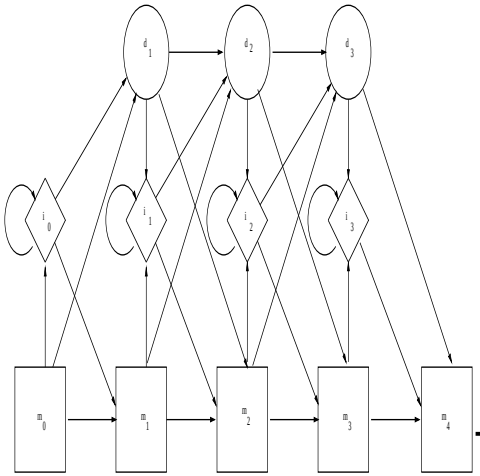
KTH VETENSKAP OCH KONST
KTH Matematik

Modelling Protein Families

Both transition and emission probabilities can be estimated from a multiple alignment of a family of sequences. An HMM can be compared (that is, aligned) with a new sequence to determine the probability that the sequence belongs to the modeled family.

KTH Matematik

Modelling Protein Families

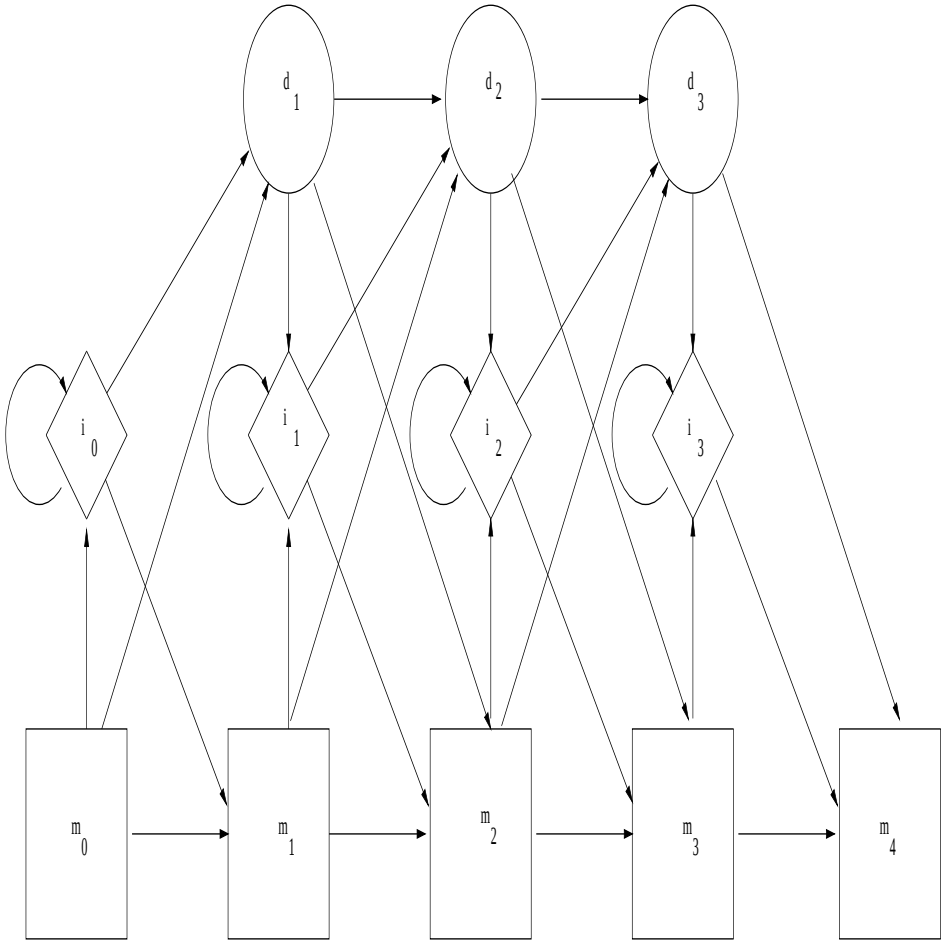


The **most probable path through the HMM** (i.e., which transitions were taken and which residues were emitted at match and insert states) is taken to generate a sequence similar to the new sequence determines the similarity score.

KTH VETENSKAP OCH KONST
KTH Matematik

Topology

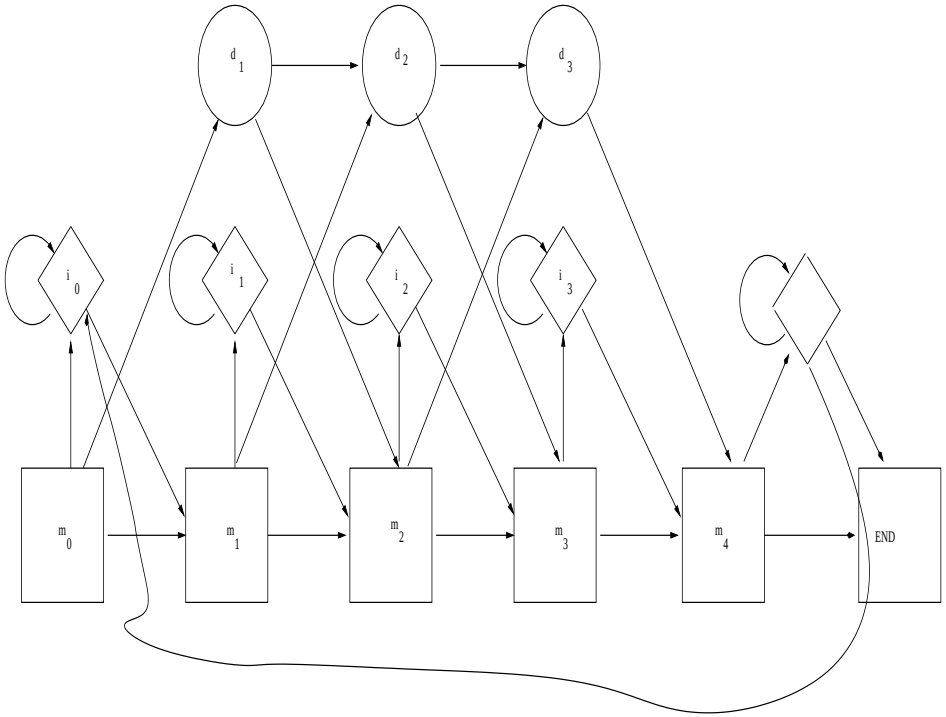
The profile HMM in the figure expresses the Haussler-Krogh **topology** of



the state space

Topology

The profile HMM in the figure has another topology , HMMER - 7.



KTH Matematik

HMM-TOP

Five structural parts were defined in membrane proteins: membrane helix (h), inside and outside helix tail (i and o), inside and outside loop (I and O). Two connected tails form a short loop associated with the membrane, while the tail-loop-tail sequence forms a long loop in the cytosol or in the extra-cytosol. Topology is determined by partitioning amino acid sequence in a way that product of the relative frequencies of amino acids in these structural parts along the sequence should be maximal. This task can be solved by the hidden Markov model (HMM), in which biological constraints can be taken into account by the architecture of HMM. The structural parts correspond to the five states used by the model.

KTH Matematik

The model uses two type of states: one for structural parts, whose length can not be arbitrary (fixed-length type, FL), and one for arbitrary long structural parts (non-fixed-length type, NFL). In the later type there are two possible transitions: one to the same state (elongation) and one to the next state (termination). In case of inside loop the next state is inside tail, and in case of outside loop the next state is outside tail. Structure of FL type states is more complex. Let be the maximum length of an FL type state MAXL, and minimum length MINL. Then let's define MAXL substates.

KTH Matematik

There is only one possible transition from the first MINL substates: transition to the next substate. There are two or three possible transitions between the substate MINL and the substate MAXL: one to the next substate (elongation), one to the next state (termination) and in case of tail following helix the third possibility is to transit to the next tail leaving out the loop state.

KTH Matematik

TMHMM

method to model and predict the location and orientation of alpha helices in membrane-spanning proteins is presented. It is based on a hidden Markov model (HMM) with an architecture that corresponds closely to the biological system. The model is cyclic with 7 types of states for helix core, helix caps on either side, loop on the cytoplasmic side, two loops for the non-cytoplasmic side, and a globular domain state in the middle of each loop. The two loop paths on the non-cytoplasmic side are used to model short and long loops separately, which corresponds biologically to the two known different membrane insertions mechanisms. The close mapping between the biological and computational states allows us to infer which parts of the model architecture are important to capture the information that encodes the membrane topology, and to gain a better understanding of the mechanisms and constraints i

KTH Matematik