Bioinformatics
Lecture : Models of DNA
Timo Koski

TK

02.08.2018

KTH Matematik

Lecture 6

The lecture is based on portions of chapter 5.2, 5.3, 10.3, 11., 11.3 of Ewens and Grant, and some additional material.

KTH Matematik

Lecture 6: Contents

1) Weight Matrix Model
2) Learning Markov models
3) More on Markov Modelling
   - k:th order Markov
   - Frame Dependent MC
   - Hidden Markov Models

KTH Matematik

Weight Matrix Model

A weight matrix $\mathcal{M}_0$ is a simple model often used by molecular biologicists as a representation for a *family of signals*. The sequences containing the signals are supposed to have equal length ($=n$) and to have no *gaps* (no positions are blank).

KTH VETENSKAP OCH KONST
KTH Matematik

Weight Matrix Model

A weight matrix $\mathcal{M}_0$ has as entries the probabilities $p_i\left(x_j\right)$ (e.g. *observed relative frequency*) for that a string should have one of the bases

$$\{x_1, x_2, x_3, x_4\} = \{A, T, C, G\}$$

at position i:

$$\mathcal{M}_0 : \quad \begin{matrix} p_1\left(x_1\right) & \dots & p_n\left(x_1\right) \\ p_1\left(x_2\right) & \dots & p_n\left(x_2\right) \\ p_1\left(x_3\right) & \dots & p_n\left(x_3\right) \\ p_1\left(x_4\right) & \dots & p_n\left(x_4\right) \end{matrix} .$$

The weight matrix model is often called a *profile*.

KTH VETENSKAP OCH KONST
KTH Matematik

Prob(Data | Model)

The probability of a finite sequence $\mathbf{x} = x_{l_1} x_{l_2} \ldots x_{l_n}$ given the model $\mathcal{M}_0$ is given by

$$P\left(\mathbf{x}|\mathcal{M}_0\right) = \prod_{j=1}^{4} \prod_{i=1}^{n} p_i\left(x_j\right)^{I_{i,x_j}(\mathbf{x})},$$

where the indicator $I_{i,x_j}(\mathbf{x})$, a function of $\mathbf{x}$, is 0 if $x_j \neq x_{l_i}$, i.e., if the symbol x_j does not appear in position i in the string $\mathbf{x}$ and is 1 otherwise. Thus the bases in the different positions are *independent* given $\mathcal{M}_0$.

KTH Matematik

Prob(Data | Model)

A sequence of strings $\mathbf{x}^1, \ldots, \mathbf{x}^t$ is training data, i.e., of known cases of members of a signal family. We take them to be generated *independently* given $\mathcal{M}_0$, is by multiplication of the preceding expressions assigned the probability

$$P\left(\mathbf{x}^1, \ldots, \mathbf{x}^t | \mathcal{M}_0\right) = \prod_{s=1}^{t} P\left(\mathbf{x}^s | \mathcal{M}_0\right)$$

$$= \prod_{j=1}^{4} \prod_{i=1}^{n} p_i \left(x_j\right)^{n_i(x_j)},$$

KTH Matematik

Prob(Data | Model)

$$P\left(\mathbf{x}^1, \ldots, \mathbf{x}^t | \mathcal{M}_0\right) = \prod_{s=1}^{t} P\left(\mathbf{x}^s | \mathcal{M}_0\right)$$

$$= \prod_{j=1}^{4} \prod_{i=1}^{n} p_i \left(x_j\right)^{n_i(x_j)},$$

where $n_i\left(x_j\right)$ is the number of times the symbol x_j appears on position i in $\mathbf{x}^1, \ldots, \mathbf{x}^t$.

KTH VETENSKAP OCH KONST
KTH Matematik

Prob(Data | $\text{Model}_{\text{MLE}}$)

$$P\left(\mathbf{x}^1, \ldots, \mathbf{x}^t | \mathcal{M}_0\right) = \prod_{s=1}^{t} P\left(\mathbf{x}^s | \mathcal{M}_0\right)$$

$$= \prod_{j=1}^{4} \prod_{i=1}^{n} p_i\left(x_j\right)^{n_i\left(x_j\right)},$$

where $n_i\left(x_j\right)$ is the number of times the symbol x_j appears on position i in $\mathbf{x}^1, \ldots, \mathbf{x}^t$.
The maximum likelihood estimate is

$$\widehat{p}_i\left(x_j\right) = \frac{n_i\left(x_j\right)}{t}$$

This will be shown during a later 'lektion'.

KTH VETENSKAP OCH KONST
KTH Matematik

Example: Promoter Regions

RNA polymerase molecules start transcription by recognizing and binding to promoter regions upstream of the desired transcription start sites. Unfortunately promoter regions do not follow a strict pattern. It is possible to find a DNA sequence (called the consensus sequence) to which all of them are very similar.

KTH VETENSKAP OCH KONST
KTH Matematik

Example: Promoter Regions

For example, the consensus sequence in the bacterium *E. Coli*, based on the study of 263 promoters, is TTGACA followed by 17 random base pairs followed by TATAAT, with the latter located about 10 bases upstream of the transcription start site. None of the 263 promoter sites exactly match the above consensus sequence.

KTH Matematik

Weight Matrix Model

By constructing the weight matrix of the TATAAT region (n=6) we can compute the probability of a DNA sequence $\mathbf{x} = x_{l_1} x_{l_2} \dots x_{l_6}$ and compute the probability

$$\text{Prob}(\mathbf{x} \mid 'promoter')$$

This means that successive bases are thought as being generated independently from the distributions in the weight matrix table. Similarly, we can compute using, e.g., a weight matrix model of the signal family

$$\text{Prob}(\mathbf{x} \mid 'non-promoter')$$

with weight matrix from a non-promoter region and compare

$$\frac{\text{Prob}(\mathbf{x} \mid 'promoter')}{\text{Prob}(\mathbf{x} \mid 'non-promoter')}.$$

to decide in $\mathbf{x}$ is a member of the family.

KTH Matematik

Bayesian Probability Recalled

Let me directly quote a statement from a webpage of the Center for Genomics Research/Karolinska Institutet

`http://kisac.cgr.ki.se/Pfam/help/scores.html`

"...Life is surprisingly easy for a Bayesian. He or She assumes you want to know something about an unknown variable. Generally you pose the question after you have seen some data which you are (generally) pretty sure about.

KTH VETENSKAP OCH KONST
KTH Matematik

Bayesian Probability (quote continued)

For example, you want to know whether this protein (data which you observe) is an example of this domain, for example a globin. The odd thing about Bayesian statistics is that it assumes that everything unknown is random variable. In this case it assumes that the variable 'is this a globin domain' is a random variable.
The Bayesian will then reply with a probability of that random variable, given the data you have observed. " (End of quote)

KTH Matematik

Markov Model $\mathcal{M}_1$

A Markov model is a statistical 'counter hypothesis' to weight models.

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

The matrix $\mathbf{P}$ contains $16 - 4$ unknown probabilities that will have to be learned from training data, i.e., of known cases of member of a signal family.

KTH Matematik

Prob (Data | $\mathcal{M}_1$)

If

$$\{X_n\}_{n=0}^{\infty} \in \text{Markov}\left(P, p_{X_0}\right),$$

where

$$p_{X_0} = \left(P\left(X_0 = 1\right), \ldots, P\left(X_0 = J\right)\right),$$

then

$$P\left(X_0 = j_0, X_1 = j_1, \ldots, X_n = j_n\right) = p_{X_0}\left(j_0\right) \prod_{l=1}^{n} p_{j_{l-1}|j_l}.$$

KTH Matematik

Learning with Markov chains

$$\mathbf{P} = \begin{pmatrix} p_{1|1} & p_{1|2} & p_{1|2} & p_{1|4} \\ p_{2|1} & p_{2|2} & p_{2|3} & p_{2|4} \\ p_{3|1} & p_{3|2} & p_{3|3} & p_{3|4} \\ p_{4|1} & p_{|2} & p_{4|3} & p_{4|4} \end{pmatrix}$$

We change the interpretation: the function $p\left(\mathbf{x}|\mathcal{M}_1\right)$ is regarded as a function of $\mathbf{P}$ (or the probabilities in $\mathbf{P}$) and called a *likelihood function* and denoted by $L_{\mathbf{x}}\left(\mathbf{P}\right)$

$$L\left(\mathbf{P}\right) = p_{j_0}(0) \prod_{l=1}^{n} p_{j_{l-1}|j_l}.$$

The maximum likelihood estimate of $\mathbf{P}$ is obtained by maximizing $L\left(\mathbf{P}\right)$ as a function of $\mathbf{P}$.

KTH VETENSKAP OCH KONST
KTH Matematik

Learning with Markov chains

The maximum likelihood estimate $\widehat{p}_{i|j}$ of $p_{i|j}$ is

$$\widehat{p}_{i|j} = \frac{n_{i|j}}{n_i}, \text{ for all } i \text{ and } j.$$

Here $n_{i|j}$ is the number of times the sequence contains the pair of bases (i, j) (in this order), i.e., the number of transitions from i to j and n_i is the number of times the base i occurs in the sequence.
This will be shown during a later 'lektion'.

KTH VETENSKAP OCH KONST
KTH Matematik

Modelling with Markov chains

$$\widehat{p}_{i|j} = \frac{n_{i|j}}{n_i}, \text{ for all } i \text{ and } j.$$

Here $n_{i|j}$ is the number of times the sequence contains the pair of bases (i, j) (in this order), i.e., the number of transitions from i to j and n_i is the number of times the base i occurs in the sequence.
Hence Markov models assume that there is biological information contained in the frequency of pairs of bases following each other.

KTH VETENSKAP OCH KONST
KTH Matematik

Learning & Parametric Inference

In maximum likelihood estimation we have regarded the transition probabilities as *parameters* and used training data to *infer* their values.
Inference: the process of deriving a conclusion from from fact and/or premise.
In probabilistic modelling of sequences the facts are the observed sequences, the premise is represented by the model and the conclusions concern unobserved quantities.

KTH Matematik

Markovian Models for Sequences

Following models have been applied to biological sequences and are based on modifications/extensions of the Markov property:

- Markov Chains of k:th order
- Frame Dependent Markov Chains
- Hidden Markov Models (HMM)

KTH Matematik

Markov chains of kth order

A sequence of random variables $\{X_n\}_{n=0}^{\infty}$ is called a **k:th order Markov chain** , if for all $n \geq 1$ and $j_0, j_1, \ldots, j_n \in S$,

$$P\left(X_n = j_n | X_0 = j_0, X_1 = j_1, \ldots, X_{n-1} = j_{n-1}\right) =$$

$$= P\left(X_n = j_n | X_{n-k} = j_{n-k}, \ldots, X_{n-1} = j_{n-1}\right),$$

for a positive integer k.

KTH VETENSKAP OCH KONST
KTH Matematik

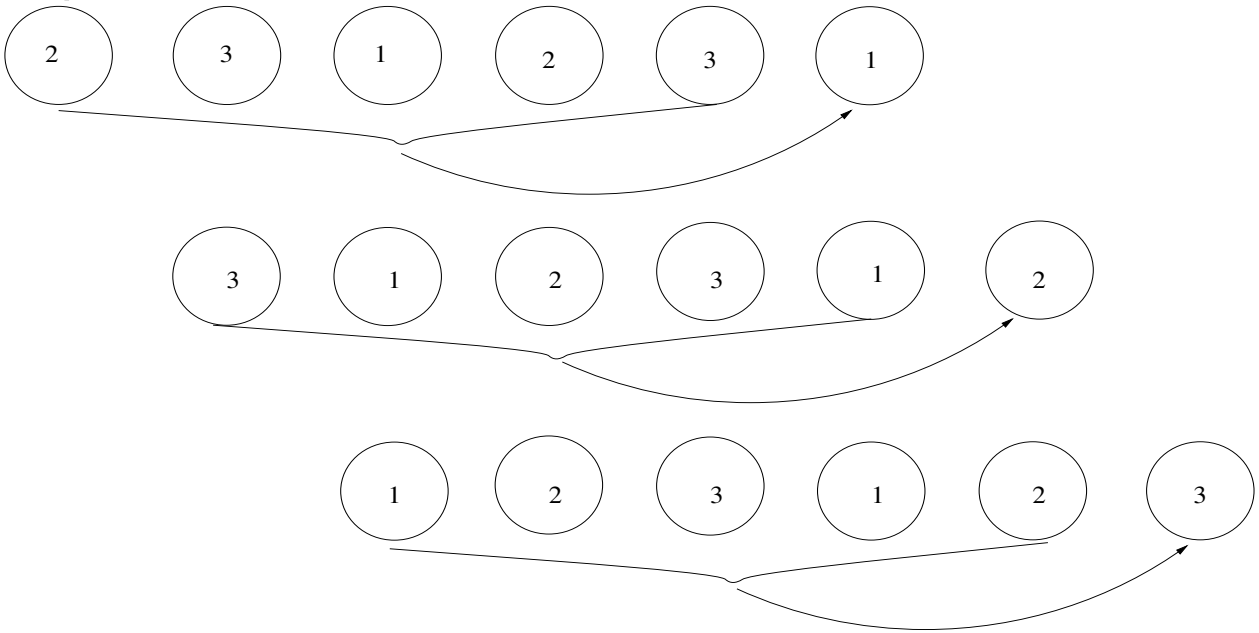
Markov chains of kth order

The MC in the first definition is called a first order Markov chain. An I.I.D process assuming values in S would then be called a Markov chain of zero order. MC:s of order higher than one are frequently used in modelling of DNA sequences. E.g., GeneMark^{TM} uses MC:s of order $k = 5$.

KTH VETENSKAP OCH KONST
KTH Matematik

Frame Dependent Markov Chains

Borodovsky and McIninch (1993), introduced a general model of a *frame dependent* kth order Markov chains.



KTH Matematik

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. Therefore a Markov chain of order k with three stationary 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

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

KTH Matematik

Frame Dependent Markov Chains

The kth order 3-phase MC model is augmented by a homogeneous ($\mathbf{P}$ as above) 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 Matematik

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

`http://genemark.biology.gatech.edu/GeneMark`

KTH Matematik

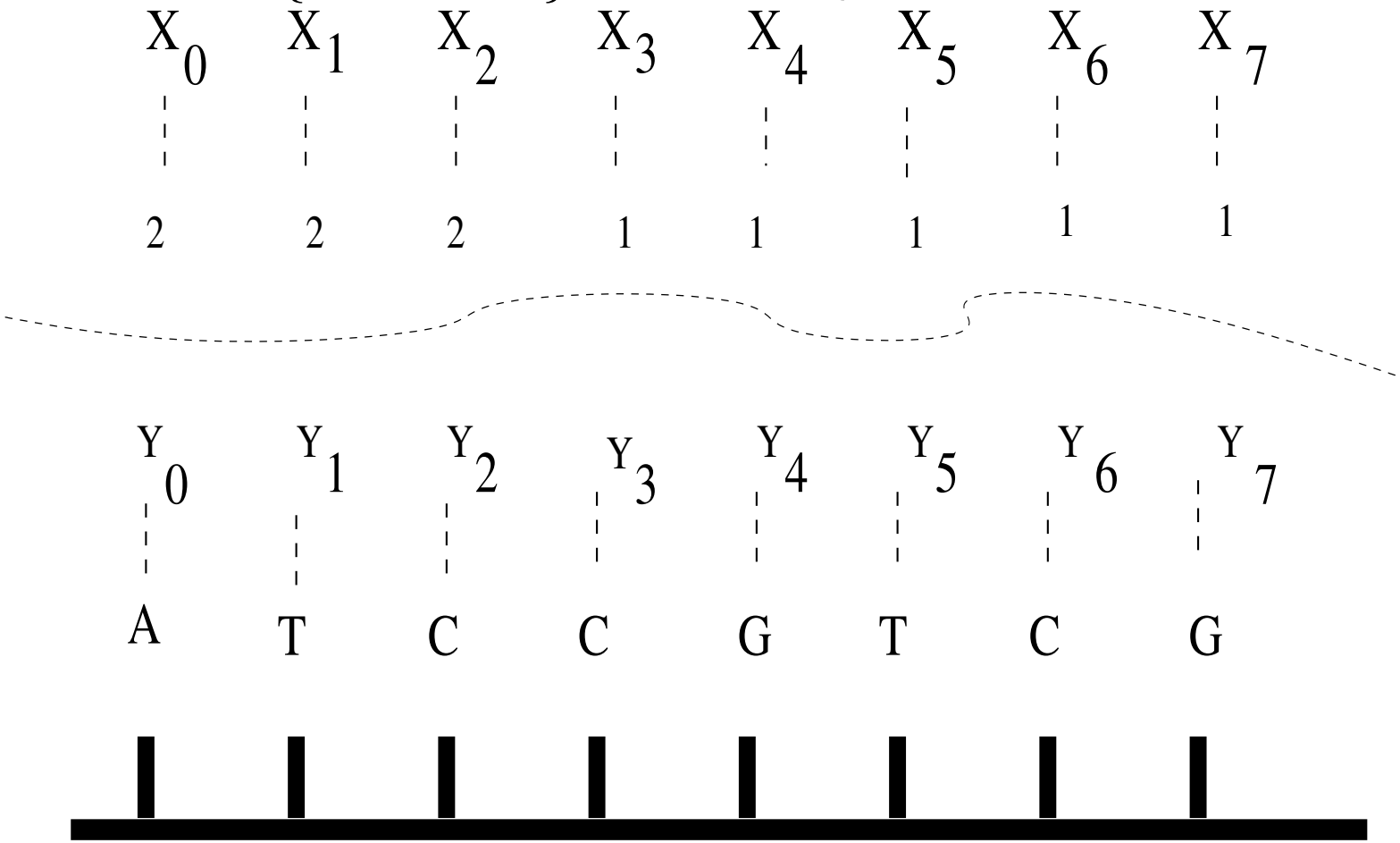
HMM

Next we discuss some simple examples of HMM:

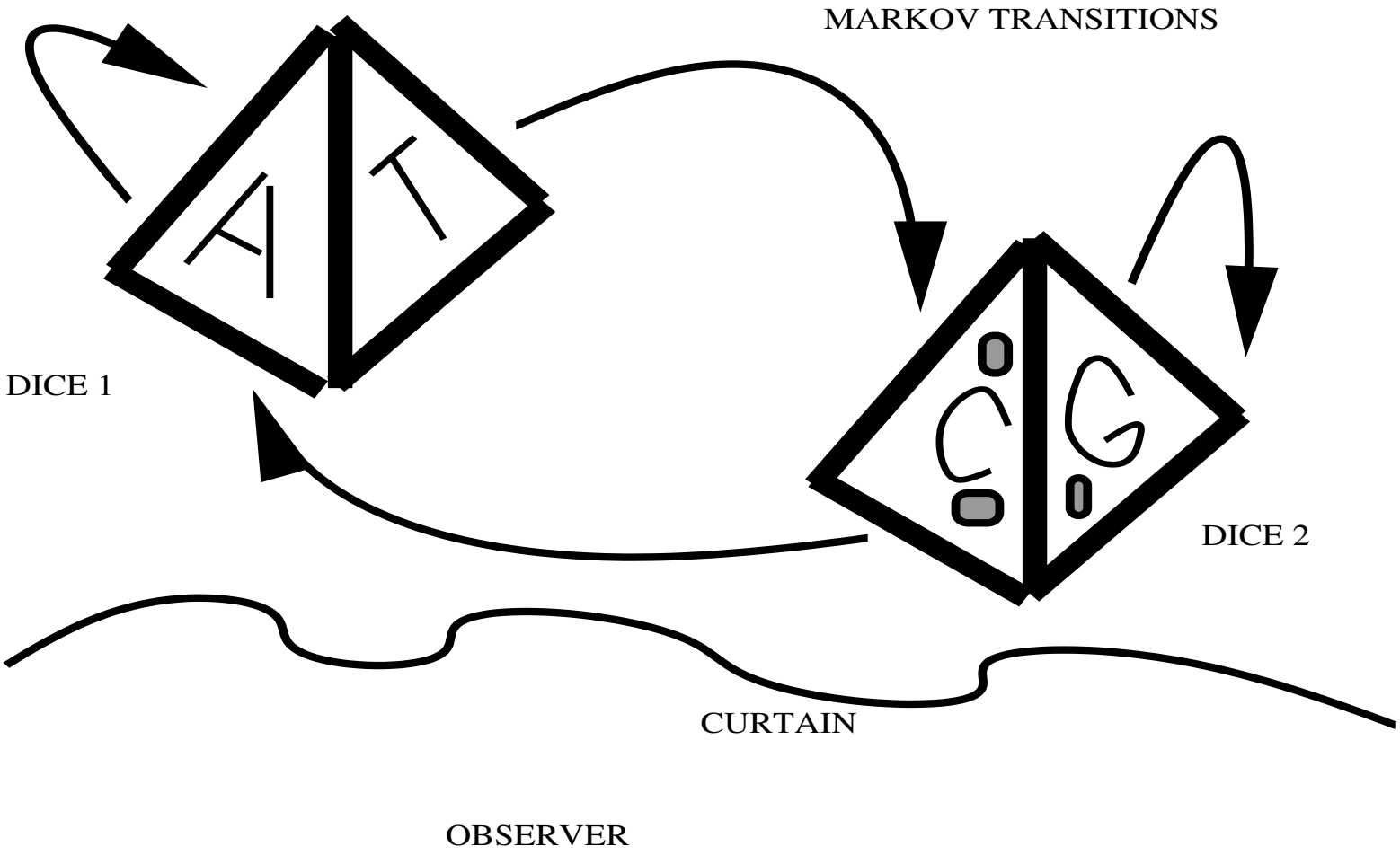
- Modelling DNA heterogenity
- Modelling protein families

KTH Matematik

M.C. X with the state space $S = \{1, 2\}$ and an observed process Y with values $\mathcal{O} = \{A, T, G, C\}$, respectively.



Modelling heterogeneity of DNA sequences



KTH Matematik

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

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

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

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

E.g., $r = P(Y = G \mid X = 2)$.

KTH Matematik

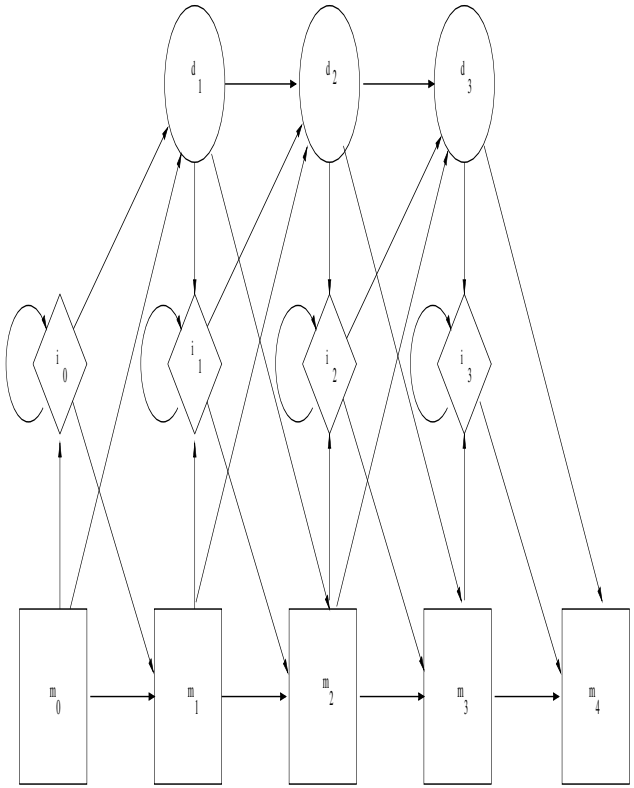
Modelling heterogeneity of DNA sequences

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

KTH Matematik

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.



Hidden Markov Models (HMM)

HMM is a model family 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 VETENSKAP OCH KONST
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 $S = \{1, 2, \ldots, J\}$ 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

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

KTH VETENSKAP OCH KONST
KTH Matematik

Hidden Markov Models (HMM) I

A matrix

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

with the constraints

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

is called a stochastic matrix.

KTH VETENSKAP OCH KONST
KTH Matematik

Hidden Markov Models (HMM) I

At time $n = 0$ the state X_0 is specified by the initial probability distribution $\pi_j(0) = P(X_0 = j)$ with

$$\pi(0) = (\pi_1(0), \ldots, \pi_J(0)).$$

$\pi_j(n) = P(X_n = j)$

$$\pi(n) = (\pi_1(n), \ldots, \pi_J(n)).$$

KTH Matematik

Hidden Markov Models (HMM) II

(II) **Observable 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 is another stochastic matrix in the sense that

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

KTH VETENSKAP OCH KONST
KTH Matematik

Hidden Markov Models (HMM) III

(III) **Conditional independence** 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

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

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

A Formalism

An HMM is designated by

$$\lambda = (A, B, \pi(0)) .$$

KTH Matematik

UNDER THE HMM ASSUMPTIONS THE STRING $\mathbf{o} = o_0 \dots o_n$ HAS THE PROBABILITY

$$P(\mathbf{o}) = P(Y_0 = o_0, \dots, Y_n = o_n; \lambda) =$$

$$\sum_{j_o=1}^{J} \dots \sum_{j_n=1}^{J} P(Y_0 = o_0 \dots, Y_n = o_n, X_0 = j_0, \dots, X_n = j_n; \lambda)$$

Hidden Markov Models, A Formalism (continued)

$$\sum_{j_o=1}^{J} \dots \sum_{j_n=1}^{J} P\left(Y_0 = o_0 \dots, Y_n = o_n, X_0 = j_0, \dots, X_n = j_n; \lambda\right)$$

where

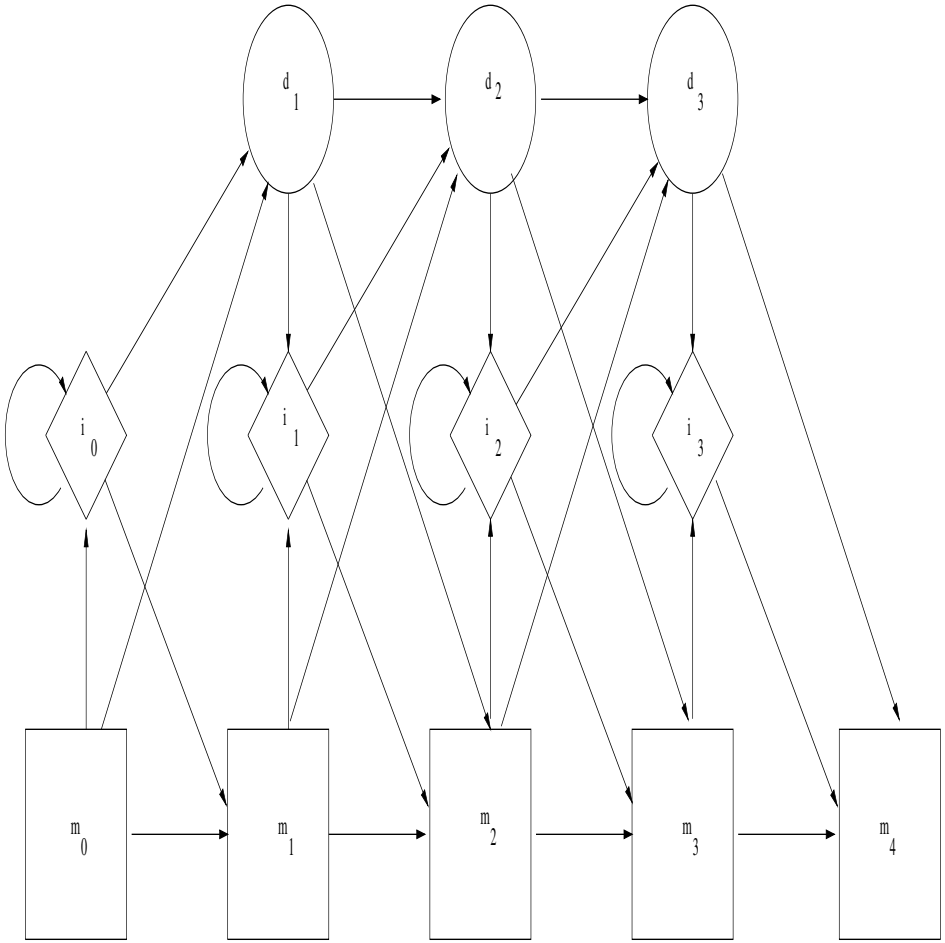
$$P\left(Y_0 = o_0, \dots, Y_n = o_n, X_0 = j_0, \dots, X_n = j_n; \lambda\right) =$$

$$\pi_{j_0}\left(0\right) \cdot \prod_{l=0}^{n} b_{j_l}\left(l\right) \prod_{l=1}^{n} a_{j_{l-1}|j_l}.$$

KTH Matematik

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 VETENSKAP OCH KONST
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.

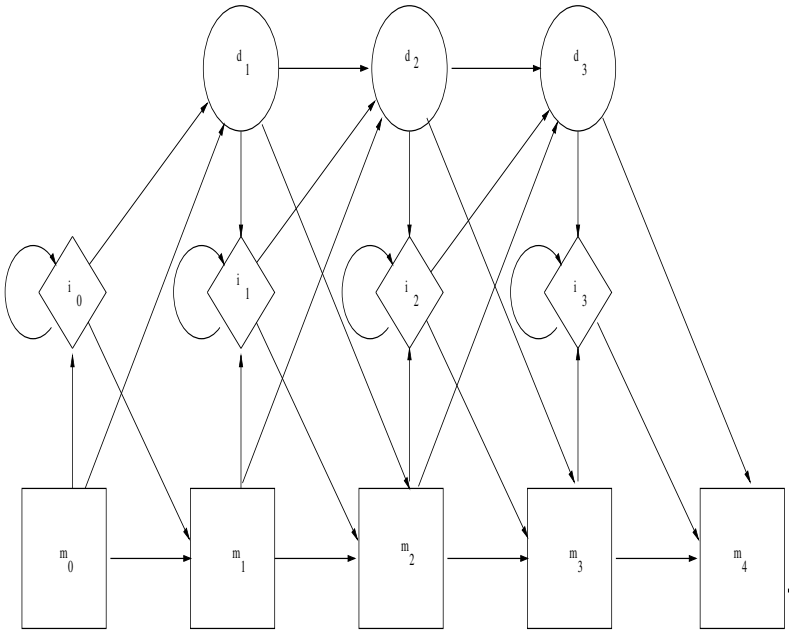
KTH Matematik

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

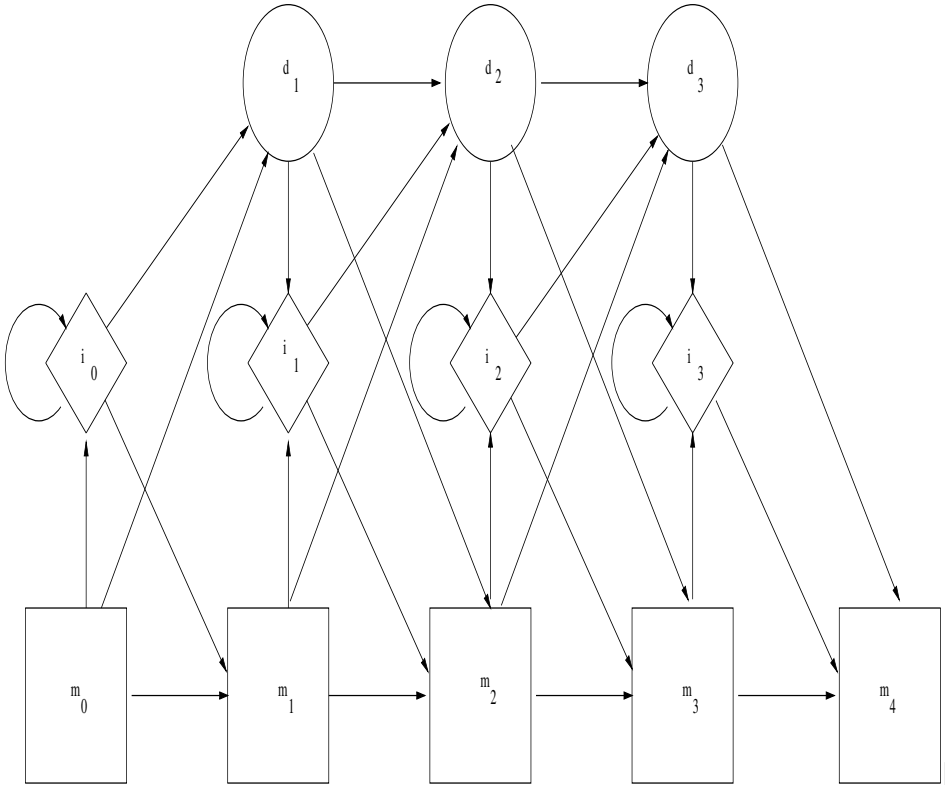
Modelling Protein Families



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

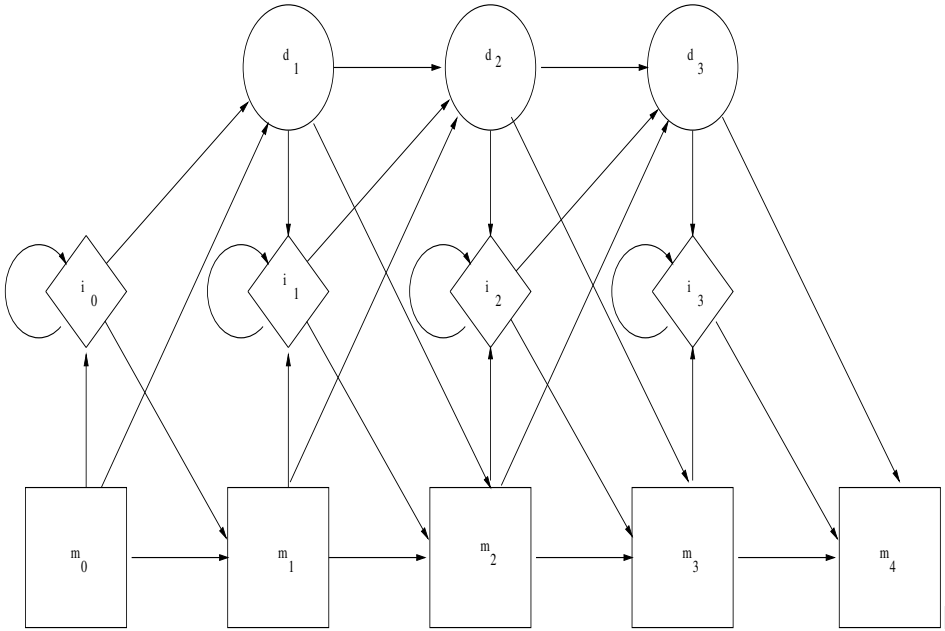
Modelling Protein Families



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.

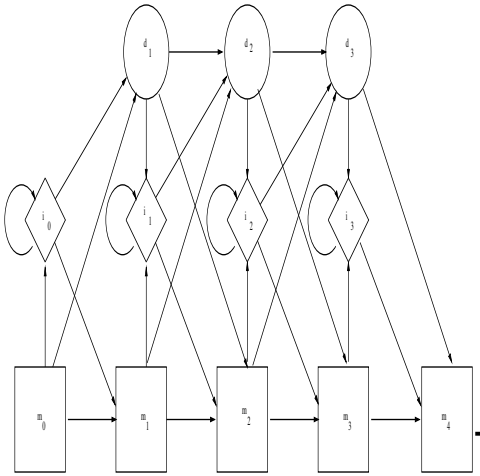
KTH VETENSKAP OCH KONST
KTH Matematik

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 VETENSKAP OCH KONST
KTH Matematik

Modelling Protein Families

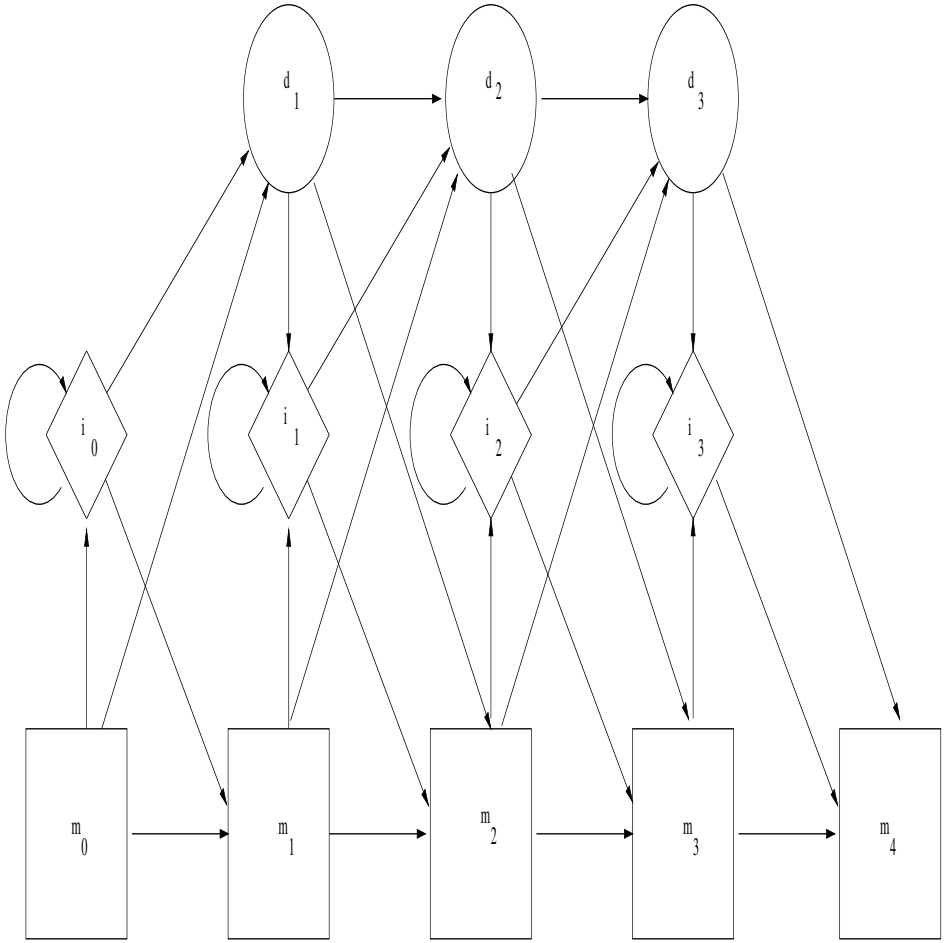


The **most probable path through the HMM** (i.e., which transitions were taken and which residues were emitted at match and insert states) is taken to generate a sequence similar to the new sequence determines the similarity score.

KTH VETENSKAP OCH KONST
KTH Matematik

Topology

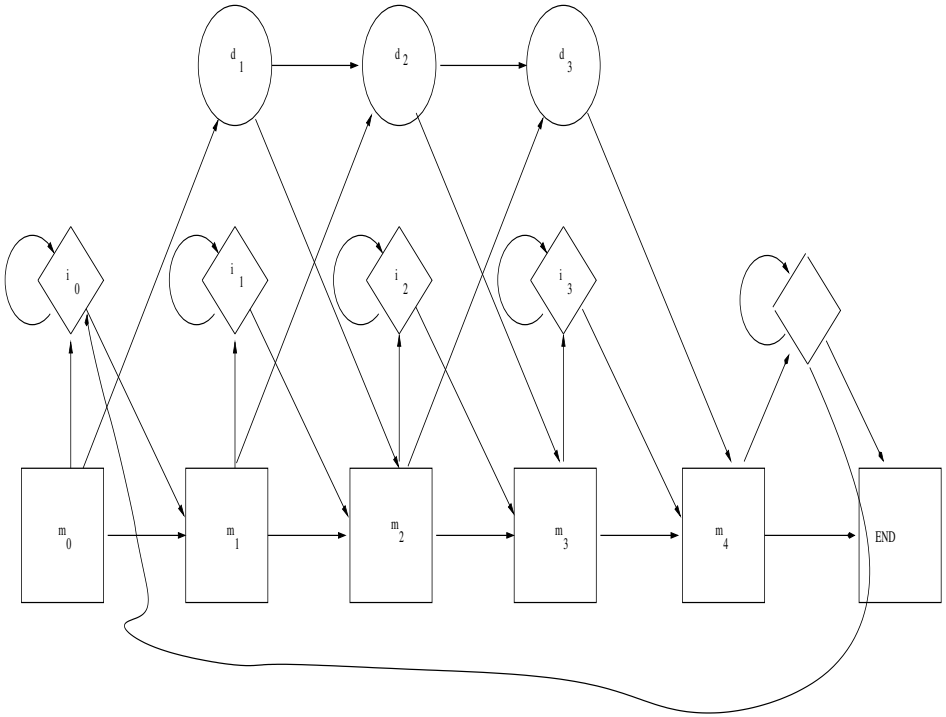
The profile HMM in the figure expresses the Haussler-Krogh **topology** of



the state space

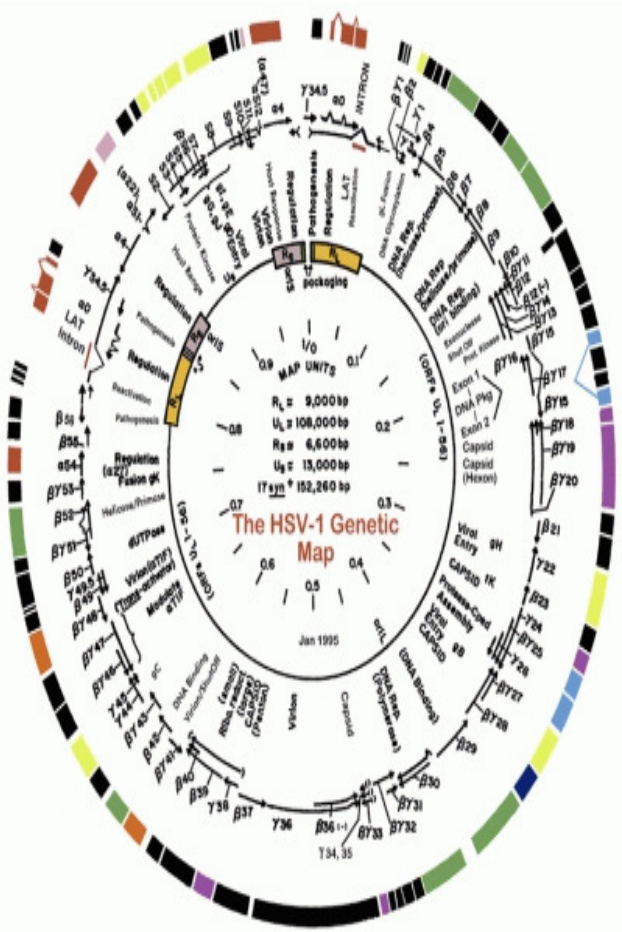
Topology

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



KTH Matematik

End of Lecture



KTH Matematik