Statistical Bioinformatics; Makerere
Markov Chains
Timo Koski

TK

02.08.2018

KTH Matematik

Lecture

This lecture covers some basic relationships of Markov chains with a finite number of states, slightly more extensively than chapter 4 of Ewens and Grant.

Contents

Markov Chains in Bioinformatics

Markov chains are useful at the genome level. It is, however, quite unlikely that a single Markov chain can describe a whole genome. Once a Markov chain has been fitted, no biological mechanism is implied, but useful questions can be answered. The frequency of particular subsequences can be predicted, and, e.g., the expected number of fragments produced by a specific restriction enzyme can be predicted.

Markov Chains in Bioinformatics, an Example: GeneMark^{TM}

GeneMark^{TM} is/was family of gene prediction programs provided by Mark Borodovsky's Bioinformatics Group at the Georgia Institute of Technology, Atlanta, Georgia. The GeneMark program is accessing the protein-coding potential of a DNA sequence (within a sliding window) by using Markov models of coding and non-coding regions.

KTH Matematik

State Space

Consider an alphabet $S = \{E_1, E_2, \ldots, E_J\}$ and sequence of random variables $X_0, X_1, \ldots, X_n, \ldots,$ assuming values in S. The symbols E_j in the alphabet are called *states* and S is also called the state space. We give the state E_j the label j and take for simplicity of typing $S = \{1, 2, \ldots, J\}$.

KTH Matematik

Markov Chain

A sequence of random variables $\{X_n\}_{n=0}^{\infty}$ is called a **Markov chain,(MC)**, if for all $n \geq 1$ and $j_0, j_1, \dots, j_n \in S$,

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

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

$\square$

The condition is known as the *Markov property*.

KTH Matematik

A.A.Markov 1856-1922

Markov Chain: Lack of Memory

The significance of an MC lies in the fact that if $X_n = j_n$ is a future event, then the conditional probability of this event given the past history $X_0 = j_0, X_1 = j_1, \ldots, X_{n-1} = j_{n-1}$ depends only upon the immediate past $X_{n-1} = j_{n-1}$ and not upon the remote past $X_0 = j_0, X_1 = j_1, \ldots, X_{n-2} = j_{n-2}$.
In bioinformatics the index n does not indicate 'time', but a sequence position.

Modelling by Chain Rule

Recall

$$P\left(X_1 = x_{l_1}, \cdots, X_m = x_{l_m}\right) =$$

$$= \prod_{i=1}^{m} P\left(X_i = x_{l_i} \mid X_1 = x_{l_1} \ldots X_{i-1} = x_{l_{i-1}}\right)$$

where

$$P\left(X_1 = x_{l_1} \mid X_0 = x_{l_0}\right) = P\left(X_1 = x_{l_1}\right).$$

This is for obvious reasons unpractical.

KTH Matematik

Markov Chain: some terminology

Let $\{X_n\}_{n=0}^{\infty}$ be Markov chain. If $X_n = j$, we say that the *the chain is in state j at time n* or that *the chain visits the state at time n*. The conditional probabilities

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

are assumed to be independent of n (temporally homogeneous) and are called *(stationary) one-step transition probabilities.* If the conditional probability is not defined, we put $p_{i|j} = 0$.

KTH Matematik

Transition matrix

The numbers $p_{i|j}$ are taken as entries in a matrix

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

$$\begin{array}{llllllll}
 & \text{to } E_1 & \text{to } E_2 & \ldots & \text{to } E_{J-1} & \text{to } E_J & & \\
\text{from } E_1 & p_{1|1} & p_{1|2} & \cdots & p_{1|J-1} & p_{1|J} & & \\
\text{from } E_2 & p_{2|1} & p_{2|2} & \cdots & p_{2|J-1} & p_{2|J} & & \\
\vdots & \vdots & \vdots & \vdots & \ldots & \vdots & \vdots & . \\
\text{from } E_{J-1} & p_{J-1|1} & p_{J-1|2} & \cdots & p_{J-1|J-1} & p_{J-1|J} & & \\
\text{from } E_J & p_{J|1} & p_{J|2} & \cdots & p_{J|J-1} & p_{J|J} & &
\end{array}$$

Transition matrix

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

$$P = \begin{pmatrix} p_{1|1} & p_{1|2} & \cdots & p_{1|J} \\ p_{2|1} & p_{2|2} & \cdots & p_{2|J} \\ \vdots & \vdots & \vdots & \vdots \\ p_{J|1} & p_{J|2} & \cdots & p_{J|J} \end{pmatrix}.$$

Thus P is an $J \times J$ matrix to be called a *transition matrix*.

Transition matrix

The $i : th$ row of P is the conditional probability distribution of X_n given that $X_{n-1} = i$. Clearly the following properties hold true:

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

KTH Matematik

Example

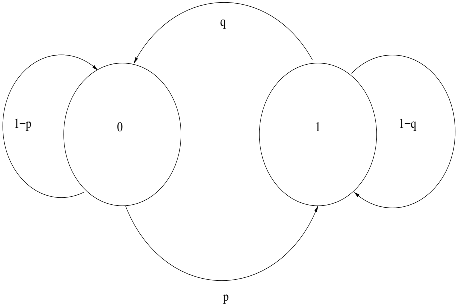
A **binary Markov information source** is a sequential mechanism for which the chance that a certain symbol will be produced depends upon the preceding symbol. Suppose the symbols are 0 and 1. If at some stage 0 is produced, then at the next stage 1 will be produced with probability p and 0 will be produced with probability $1-p$. If a 1 is produced, then at a next stage 0 will be produced with probability q and 1 will be produced with probability $1-q$. This corresponds to the transition matrix

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

KTH Matematik

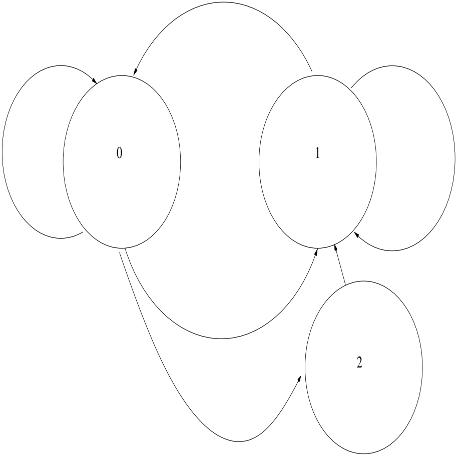
State transition graph

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



Topology of the Graph

The structure of a state transition graph without the probabilities is known



as the **topology of the graph**.

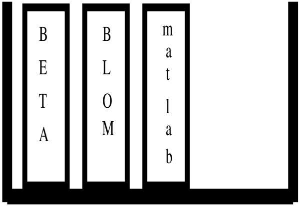
Example

A **binary Markov information source** is a generalization of a sequence of identically and independently $\mathrm{Be}(p)$ - distributed bits. I.I. $\mathrm{Be}(p)$ - distributed bits correspond to the transition matrix

$$P = \begin{pmatrix} 1-p & p \\ 1-p & p \end{pmatrix}.$$

Example: McCabe's library in a special case

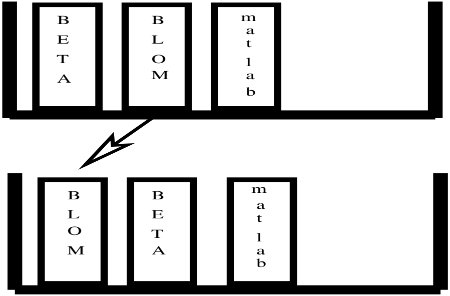
Linnea has a set of three books on a bookshelf. These are (1) L. Råde & B. Westergren: **BETA**, (2) G. Blom: Probability and



Statistics, (3) F. Gustafsson & N. Bergman: MATLAB^{R} for Engineers.

Example: McCabe's library in a special case

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.



Example: McCabe's library in a special case

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

Example: McCabe's library in a special case

The order (from the left) between the books becomes thus the state of a Markov chain, which jumps to a next state every time Linnea returns a book on her shelf.

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

KTH Matematik

Example: McCabe's library in a special case

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

	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 Matematik

Example: McCabe's library in a special case

This is a special case of a known model for self-organization of linear lists of data records and is called McCabe's library.
M. Hofria & H. Schahnai (1991): Self-organizing Lists and Independent References. *Journal of Algorithms*, pp. 533−555,

KTH Matematik

Transition matrix for DNA sequences

$$P = \left(p_{i|j}\right)_{i=1,j=1}^{4,4}$$

or obviously $\begin{array}{c c c c c} & \text{A} & \text{C} & \text{G} & \text{T} \\ \text{A} & p_{\mathrm{A|A}} & p_{\mathrm{A|C}} & p_{\mathrm{A|G}} & p_{\mathrm{A|T}} \\ \text{C} & p_{\mathrm{C|A}} & p_{\mathrm{C|C}} & p_{\mathrm{C|G}} & p_{\mathrm{A|T}} \\ \text{G} & p_{\mathrm{G|A}} & p_{\mathrm{G|C}} & p_{\mathrm{A|G}} & p_{\mathrm{G|T}} \\ \text{T} & p_{\mathrm{T|A}} & p_{\mathrm{T|C}} & p_{\mathrm{T|G}} & p_{\mathrm{T|T}} \end{array}$. How might these numbers be established from genome data ? This will (?) be treated later.

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 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 $k = 5$.

Joint Probability Distribution of an MC

By successive iterations of the definition of conditional probability and by successive uses of the Markov property

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

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

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

KTH Matematik

Joint Probability Distribution of an MC

$$p_{j_{n-1}|j_n} \cdot P\left(X_{n-1} = j_n | X_0 = j_0, \ldots, X_{n-2} = j_{n-2}\right) \cdot P\left(X_0 = j_0, \ldots, X_{n-2} = j_{n-2}\right) =$$

$$\vdots$$

$$= p_{j_{n-1}|j_n} \cdot p_{j_{n-2}|j_{n-1}} \cdots p_{j_0|j_1} \cdot p_{X_0}\left(j_0\right) =$$

$$= p_{X_0}\left(j_0\right) \cdot p_{j_0|j_1} \cdots p_{j_{n-2}|j_{n-1}} \cdot p_{j_{n-1}|j_n}.$$

Thus we have proved:

Joint Probability Distribution of an MC

If $\{X_n\}_{n=0}^{\infty}$ is a Markov chain with stationary transition probabilities, then

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

We set

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

where

$$p_{X_0} = \left(p_1, \ldots, p_J\right).$$

KTH Matematik

The conditional probabilities

$$p_{i|j}(n) = P\left(X_{m+n} = j | X_m = i\right), n \geq 1, i, j \in S$$

are also independent of m. The probabilities $p_{i|j}(n)$ are called the *n -step transition probablities*. Then

$$P(n) = \left(p_{i|j}(n)\right)_{i=1,j=1}^{J,J}$$

is the *n -step transition matrix*. We define

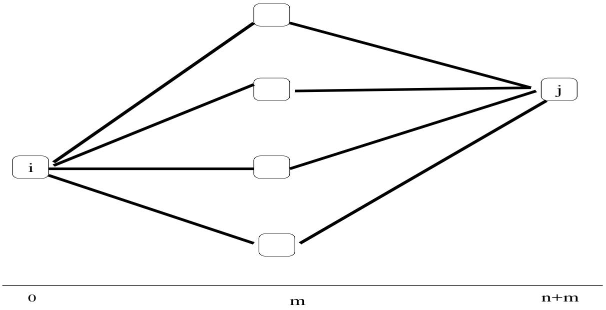
$$p_{i|j}(0) = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i. \end{cases}$$

KTH Matematik

Chapman-Kolmorogorov equations

For all $m, n \geq 1$ and $i, j \in S$,

$$p_{i|j}(m+n) = \sum_{k=1}^{J} p_{i|k}(m) \cdot p_{k|j}(n).$$



Chapman-Kolmorogorov equations, Proof:

We observe that

$$p_{i|j}(m+n) = P\left(X_{m+n} = j | X_0 = i\right) =$$

$$\sum_{k=1}^{J} \frac{P\left(X_{m+n} = j, X_m = k, X_0 = i\right)}{P\left(X_0 = i\right)}.$$

We try now to express $P\left(X_{m+n} = j, X_m = k, X_0 = i\right)$ in a suitable way.

KTH Matematik

Chapman-Kolmorogorov equations, Proof (contnd):

We consider the following identity obtained by definition of conditional probability

$$P\left(X_0 = j_0, \ldots, X_m = j_m, \ldots X_n = j_n\right) =$$

$$P\left(X_{m+1} = j_{m+1}, .., X_n = j_n | X_0 = j_0, X_1 = j_1, .., X_m = j_m\right) P\left(X_0 = j_0, .., X_m = j_m\right)$$

Chapman-Kolmorogorov equations, Proof (contnd):

We can show that

$$P\left(X_{m+1} = j_{m+1}, \ldots X_n = j_n | X_0 = j_0, X_1 = j_1, \ldots, X_m = j_m\right) =$$

$$P\left(X_{m+1} = j_{m+1}, \ldots X_n = j_n | X_m = j_m\right).$$

This is intuitively plausible, and is left as an exercise.

Chapman-Kolmorogorov equations, Proof (contnd):

$$P\left(X_0 = j_0, \ldots, X_m = j_m, \ldots, X_n = j_n\right) =$$

$$P\left(X_{m+1} = j_{m+1}, .., X_n = j_n | X_m = j_m\right) P\left(X_1 = j_1, .., X_m = j_m | X_0 = j_0\right) P\left(X_0 = j_0\right).$$

KTH Matematik

Chapman-Kolmorogorov equations, Proof (contnd):

$$P\left(X_0 = j_0, \ldots, X_m = j_m, \ldots, X_n = j_n\right) =$$

$$P\left(X_{m+1} = j_{m+1}, .., X_n = j_n | X_m = j_m\right) P\left(X_1 = j_1, .., X_m = j_m | X_0 = j_0\right) P\left(X_0 = j_0\right).$$

If we next sum over $j_1, \ldots, j_{m-1}, j_{m+1} \ldots j_{n-1}$ we get

$$P\left(X_0 = j_0, X_m = j_m, X_n = j_n\right) =$$

$$P\left(X_n = j_n | X_m = j_m\right) P\left(X_m = j_m | X_0 = j_0\right) P\left(X_0 = j_0\right).(*)$$

KTH Matematik

Chapman-Kolmorogorov equations, Proof (contnd):

As shown above

$$p_{i|j}(m+n) = P\left(X_{m+n} = j | X_0 = i\right) =$$

$$\sum_{k=1}^{J} \frac{P\left(X_{m+n} = j, X_m = k, X_0 = i\right)}{P\left(X_0 = i\right)}.$$

Replacing n by $n+m$, j_0 by i, j_m by k and j_n by j in $(*)$ above, we get from this

KTH Matematik

Chapman-Kolmorogorov equations, Proof (finished):

$$p_{i|j}(m+n) = \sum_{k=1}^{J} P\left(X_{m+n} = j | X_m = k\right) P\left(X_m = k | X_0 = i\right) =$$

$$= \sum_{k=1}^{J} p_{k|j}(n) \cdot p_{i|k}(m).$$

Chapman-Kolmorogorov equations, Matrix Form:

Using a matrix notation we can write the Chapman - Kolmogorov equation as the following matrix multiplication

$$P(n+m) = P(m) \cdot P(n).$$

n-step Transition Matrix as Matrix Power :

$$P(n) = P^n.$$

Proof: This is easily proved by induction, since where $P^0 = I$ ($=$ the $J \times J$ identity matrix), $P^1 = P$, $P^2 = P \cdot P$, $P^3 = P \cdot P^2$ and so on.

KTH Matematik

Chapman-Kolmorogorov equations

Chapman - Kolmogorov equation can be written as

$$P^{n+m} = P^m \cdot P^n.$$

State probabilities

Let the distribution of X_0 be denoted by $\phi(0)$. In other words,

$$\phi(0) = \left(p_{X_0}\left(1\right), \ldots, p_{X_0}\left(J\right)\right).$$

This will be called the *initial distribution*. Let us denote by

$$\phi(n) = \left(p\left(X_n = 1\right), \ldots, p\left(X_n = J\right)\right)$$

the $1 \times J$ vector of the probabilities that the chain is in state j at time n.

KTH Matematik

State probabilities

By marginalization

$$p\left(X_n = j\right) = \sum_{k=1}^{J} p_{k|j} \cdot p\left(X_{n-1} = k\right).$$

This we write using a matrix notation as

$$\phi(n) = \phi(n-1)P.$$

Stationary distribution

A Markov chain $\{X_n\}_{n=0}^{\infty}$ is called *stationary*, if the probability $p\left(X_n = j\right)$ is independent of n for all j in the state space. A distribution ϕ an *invariant* or *stationary distribution*, with

$$\phi = \left(\phi_1, \ldots, \phi_J\right),$$

if $p\left(X_0 = j\right) = \phi_j$ for all j implies that $p\left(X_1 = j\right) = \phi_j$ for all j.

KTH Matematik

Stationary distribution

Let $\{X_n\}_{n=0}^{\infty} \in \mathrm{Markov}\left(P, \phi(0)\right)$. Every stationary (invariant) distibution satisfies the equation

$$\phi = \phi P$$

(ϕ is a row vector) with the constraints

$$\sum_{j=1}^{J} \phi_j = 1, \phi_j \geq 0.$$

KTH Matematik

Proof:

Assume first that ϕ is an invariant distribution. Then $\sum_{j=1}^{J} \phi_j = 1$ and $\phi_j \geq 0$ are clear. Since ϕ is invariant, by the definition above we must have $\phi(0) = \phi$ and $\phi(1) = \phi$. But since

$$\phi(n) = \phi(n-1)P,$$

we get that

$$\phi = \phi P.$$

KTH Matematik

Proof:

Assume now that ϕ satisfies $\phi = \phi P$ and the other constraints. Let $\phi(0) = \phi$. Then

$$\phi(1) = \phi(0)P = \phi P = \phi$$

and ϕ is an invariant distribution. $\square$

Existence of a stationary distribution

Every MC with a finite state space has *at least one* invariant distribution *Proof:* We give only an outline of the proof. Let p be an arbitrary probability distribution on S. Set

$$p^{(n)} = \frac{1}{n}\left(p + pP + pP^2 + \ldots + pP^{n-1}\right)$$

This is a sequence of probability distributions, i.e. vectors with components with values between zero and one. Thus the well known theorem of Bolzano and Weierstrass shows that we can pick a convergent subsequence $p^{(n_V)}$ which converges componentwise to the vector ϕ. We can show that ϕ is a probability distribution.

KTH Matematik

Existence of a stationary distribution

By our construction we have the recursion relations

$$p^{(n+1)} = \frac{n}{n+1} p^{(n)} + \frac{1}{n+1} pP^n$$

and

$$p^{(n+1)} = \frac{n}{n+1} p^{(n)} P + \frac{1}{n+1} p.$$

From the recursion above we get that

$$p^{(n_v+1)} \to \phi$$

and then we get that

$$\phi = \phi P,$$

which proves the claim.

KTH Matematik

Stationary distribution

The components in the stationary distribution can be interpreted as the asymptotic percentages of ‘time’ the chain spends in each of the states.

KTH Matematik

Stationary distribution

Is there convergence to a stationary distribution for any $\phi(0)$? Let $\{X_n\}_{n=0}^{\infty} \in \mathrm{Markov}\left(P, p_{X_0}\right)$. Let us assume that

$$\lim_{n \to \infty} \phi(n) = a,$$

where $a = (a_1, \ldots, a_J)$ is a probability distribution. Then a is an invariant distribution.

Proof:

Taking of limits yields

$$a = \lim_{n \to \infty} \phi(n) = \lim_{n \to \infty} \phi(n+1) =$$

$$= \lim_{n \to \infty} \left(\phi(n) P\right) = \left(\lim_{n \to \infty} \phi(n)\right) P = aP.$$

periodic and irreducible MC:s

(a) An MC is aperiodic if there is no state such that return to that state is possible only after t_0, $2t_0$, $3t_0$ ... steps later.

(b) An MC is irreducible means that every state can eventually be reached from any other state, if not in one step, but then after several steps.

These assumptions hold for many ((?) almost all according to Ewens and Grant) MC's in bioinformatics.

Convergence to an invariant distribution

If a finite MC is aperiodic and irreducible, then for any $\phi(0)$

$$\lim_{n \to \infty} \phi(n) = \phi,$$

where ϕ is a probability distribution that satisfies

$$\phi = \phi P.$$

One of the many possible proofs of the theorem is found on pp. $40-42$ of P.Clote & R. Backofen (2000): *Computational Molecular Biology*, Wiley.

KTH Matematik

Convergence of P^n

Under the conditions of the preceding theorem we have that

$$P^n \to \begin{pmatrix} \phi_1 & \phi_2 & \dots & \phi_J \\ \phi_1 & \phi_2 & \dots & \phi_J \\ \vdots & \vdots & \vdots & \vdots \\ \phi_1 & \phi_2 & \dots & \phi_J \end{pmatrix}.$$

as $n \to +\infty$.

Stationary distribution and McCabe's library

Consider McCabe's library again.

(a) Explain why the chain is irreducible and aperiodic.

(b) Will the distribution of the chain converge to a stationary distribution?

(c) What is the expression of the invariant distribution for the case in the preceding ?

KTH Matematik

The right hand side might be

$$\sum_{i=1}^{J} \phi_i \sum_{j=1}^{J} p_{i|j} \log \frac{p_{i|j}}{q_{i|j}}$$

in the present notation.

`http://www.biology.gatech.edu/bioinformatics/whatis.html`

End of Lecture