

BIOLOGICAL SEQUENCE ANALYSIS: A RAPID SURVEY

Master 2 in Systems & Synthetic Biology – Semester 3, 2011-2012

MIKAEL FALCONNET
mikael.falconnet@genopole.cnrs.fr

Many problems in biological sequence analysis have the same structure: based on a sequence of symbols from some alphabet, capture some information about the complexity of the molecule, its history, etc. We introduce standard Markov models and hidden-Markov models which are helpful in this purpose.

1 Markov chains

All along this section we focus on a concrete example (Doudou le hamster) to illustrate rigorous definitions and results about Markov chains.

1.1 Definition and Markov property

Doudou is the hamster of your little cousin Kevin. Doudou has only three activities in his life.

- Sleeping (alone).
- Eating (pellets).
- Running (in its wheel).

We are going to use a *Markov chain* to describe Doudou's way of life and answer this important question :

“How long a day does Doudou sleep?”

Understand “Is Doudou as lazy as he seems?”.

Why a Markov chain? A Markov chain is a random process characterized by a main property : it retains *no memory* of where it has been in the past. This means that only the current state of the process can influence where it goes next.

Definition 1.1. Let S be a finite or countable space. Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in $(S, \mathcal{P}(S))$. We say that (X_n) is an homogeneous Markov chain if

- (Markov property) for $n \in \mathbb{N}$ and $x_0, \dots, x_{n+1} \in S$

$$\mathbb{P}(X_{n+1} = x_{n+1} | X_{0:n} = x_{0:n}) = \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n);$$

- (Homogeneity) the transition probability $\mathbb{P}(X_{n+1} = y | X_n = x)$ does not depend on n , for any $x, y \in S$. We denote this transition probability by $p(x, y)$.

The distribution of X_0 is called the initial distribution of the Markov chain, and we call transition matrix, the matrix (with possibly infinite dimensions) defined as

$$P = (p(x, y))_{x, y \in S}.$$

The matrix P fullfills

$$p(x, y) \geq 0 \quad \text{and} \quad \sum_y p(x, y) = 1.$$

Such a matrix is called stochastic.

In our case, a state is an activity for Doudou and $S = \{\text{Sleeping, Eating, Running}\}$.

To satisfy the Markov property, we assume that every minute, Doudou can change his activity or can do it again, and it only depends of his current activity.

Remark 1.2. In the case of Doudou, the lack of memory is quite obvious.

Behaviour of Doudou.

- When Doudou is sleeping, he does not wake up the next minute with probability 0.9.
- When Doudou wakes up, he has lunch or makes some jogging with equal probability.
- He just eats during one minute (pellets are not so tasty), after he moves.
- After lunch, he makes exercise with probability 0.3 or he has a nap with probability 0.7 (Doudou is a lazy hamster).

- Jogging is hard, hence after one minute of jogging he takes some rest with probability 0.8. But with probability 0.2, he forgets that he is tired and jogs for one minute again.

The corresponding matrix P of the description above is given by

$$P = \begin{matrix} & \begin{matrix} S & E & R \end{matrix} \\ \begin{matrix} S \\ E \\ R \end{matrix} & \begin{pmatrix} 0.9 & 0.05 & 0.05 \\ 0.7 & 0 & 0.3 \\ 0.8 & 0 & 0.2 \end{pmatrix} \end{matrix}.$$

Markov chains are always best described by diagrams as illustrated on Figure 1. We might have drawn a loop from Sleeping to itself with label 0.9 but since the total probability on jumping from S must be equal to 1, this does not convey any more information and we prefer to leave the loops out.

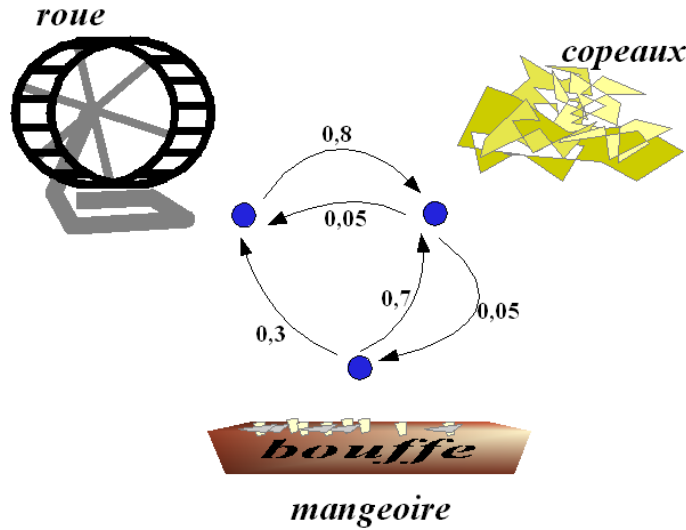


Figure 1: Doudou's French way of life

Theorem 1.3. *If a discrete time random process (X_n) is Markov(μ, P), then for any $n \geq 0$ and any $x_{0:n} \in S^{n+1}$*

$$\mathbb{P}(X_{0:n} = x_{0:n}) = \mu(x_0)p(x_0, x_1)p(x_1, x_2) \dots p(x_{n-1}, x_n).$$

Proof. By induction. By definition, $\mathbb{P}(X_0 = x_0) = \mu(x_0)$. Now, fix $n \geq 0$ and assume that for any $x_{0:n} \in S^{n+1}$

$$\mathbb{P}(X_{0:n} = x_{0:n}) = \mu(x_0)p(x_0, x_1)p(x_1, x_2) \dots p(x_{n-1}, x_n).$$

Let $x_{0:n+1} \in S^{n+2}$. We have

$$\mathbb{P}(X_{0:n+1} = x_{0:n+1}) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_{0:n} = x_{0:n}) \mathbb{P}(X_{0:n} = x_{0:n}).$$

Using Markov property yields

$$\mathbb{P}(X_{0:n+1} = x_{0:n+1}) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n) \mathbb{P}(X_{0:n} = x_{0:n}).$$

Since

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n) = p(x_n, x_{n+1}),$$

and

$$\mathbb{P}(X_{0:n} = x_{0:n}) = \mu(x_0) p(x_0, x_1) p(x_1, x_2) \dots p(x_{n-1}, x_n),$$

this achieves the proof. $\square$

1.2 Distribution of X_n

What is the probability that after n steps our Markov chain is in a given state? For instance, in our context, what is the probability that after n minutes Doudou sleeps?

We shall see that this problem reduces to calculating entries in the n th power of the transition matrix.

Proposition 1.4. *Let (X_n) be Markov(μ, P). Then*

$$\mathbb{P}(X_n = x) = (\mu P^n)(x), \quad \forall x \in S.$$

Proof. Let $\mu^{(n)}$ denote the distribution of X_n . We have

$$\mu^{(n)}(x) = \mathbb{P}(X_n = x) = \sum_y \mathbb{P}(X_n = x, X_{n-1} = y).$$

Conditioning on $[X_{n-1} = y]$ yields

$$\mu^{(n)}(x) = \sum_y \mathbb{P}(X_n = x \mid X_{n-1} = y) \mathbb{P}(X_{n-1} = y) = \sum_y p(y, x) \mu^{(n-1)}(y) = (\mu^{(n-1)} P)(x).$$

We achieve the proof by induction. $\square$

Assume that Doudou sleeps the first minute. Then,

$$\mu^{(0)} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}.$$

After one minute, one has

$$\mu^{(1)} = \mu^{(0)} P = \begin{pmatrix} 0.9 & 0.05 & 0.05 \end{pmatrix}.$$

And after two and three minutes

$$\mu^{(2)} = \begin{pmatrix} 0.885 & 0.045 & 0.07 \end{pmatrix} \quad \text{and} \quad \mu^{(3)} = \begin{pmatrix} 0.884 & 0.04425 & 0.07175 \end{pmatrix}.$$

One can see that $\mu^{(2)}$ and $\mu^{(3)}$ are close. We now explain why it is not by accident.

1.3 Convergence to equilibrium

Many of the long-time properties of Markov chains are connected with the notion of an invariant distribution or measure.

Definition 1.5. We say that a distribution λ is invariant for P if

$$\lambda P = \lambda.$$

For our example, the distribution π defined as

$$\pi = \left(\frac{160}{181} \quad \frac{8}{181} \quad \frac{13}{181} \right)$$

is invariant for P . Yet, one has

$$\frac{160}{181} \approx 0.884, \quad \frac{8}{181} \approx 0.044, \quad \text{and} \quad \frac{13}{181} \approx 0.072.$$

Some computations give us

$$P^2 = \begin{pmatrix} 0.885 & 0.045 & 0.07 \\ 0.87 & 0.035 & 0.095 \\ 0.88 & 0.04 & 0.08 \end{pmatrix},$$

and

$$P^3 = \begin{pmatrix} 0.884 & 0.04425 & 0.07175 \\ 0.8835 & 0.0435 & 0.073 \\ 0.884 & 0.044 & 0.072 \end{pmatrix}.$$

We have the impression that

$$p^{(n)}(x, y) \rightarrow \pi(y) \quad \text{as } n \rightarrow \infty.$$

This phenomenon really happens under some conditions that we detail now.

Class structure. It is sometimes possible to break a Markov chain into smaller pieces, each of which is relatively easy to understand, and which together give an understanding of the whole. This is done by identifying the communicating classes of the chain.

Definition 1.6. Let x and y be two states in S . We say that x leads to y and write $x \rightarrow y$ if there exists $n \geq 0$ such that $p^{(n)}(x, y) > 0$. We say that x communicates with y and write $x \leftrightarrow y$ if both $x \rightarrow y$ and $y \rightarrow x$.

It is clear (really?) that $x \rightarrow y$ and $y \rightarrow z$ imply $x \rightarrow z$. Also $x \rightarrow x$ for any $x \in S$. So $\leftrightarrow$ satisfies the conditions for an equivalence relation on S , and thus partitions S into communicating classes.

Definition 1.7. We say that a class C is closed if

$$x \in C, x \rightarrow y \quad \text{imply } y \in C.$$

Thus, a closed class is one from which there is no escape. A state x is absorbing if $\{x\}$ is a closed class.

Definition 1.8. A Markov chain with transition matrix P is called irreducible when S is a single class.

Exercise 1.9. Find the communicating classes of the following transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{2} & 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}$$

Explain why irreducibility is one of the condition necessary for a convergence to equilibrium.

Aperiodicity. We shall investigate the limiting behaviour of the n -step transition probabilities $p^{(n)}(x, y)$ as $n \rightarrow \infty$. Consider the two-state chain with transition matrix

$$P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then $P^2 = I$, so $P^{2n} = I$ and $P^{2n+1} = P$ for all n . Thus, $p^{(n)}(x, y)$ fails to converge for all x, y .

Definition 1.10. A state x is called aperiodic if $p^{(n)}(x, x) > 0$ for all sufficiently large n .

Lemma 1.11. Suppose P is irreducible and has an aperiodic state x . Then, for all states y and z , $p^{(n)}(y, z) > 0$ for all sufficiently large n . In particular, all states are aperiodic.

Proof. There exists $r, s \geq 0$ with $p^{(r)}(y, x), p^{(s)}(x, z) > 0$. Then

$$p^{(r+n+s)}(y, z) \geq p^{(r)}(y, x)p^{(n)}(x, x)p^{(s)}(x, z) > 0. \quad \square.$$

Here is one important result for Markov chains.

Theorem 1.12. Let P be irreducible and aperiodic, and suppose that P has an invariant distribution π . Let μ be any distribution. Suppose that (X_n) is Markov(μ, P). Then

$$\mathbb{P}(X_n = x) \rightarrow \pi(x) \quad \text{as } n \rightarrow \infty \text{ for all } x,$$

and

$$p^{(n)}(x, y) \rightarrow \pi(y) \quad \text{as } n \rightarrow \infty \text{ for all } x, y.$$

Proof. See [Nor97] page 41-42. $\square$

Exercise 1.13. A fair die is thrown repeatedly. Let X_n denote the sum of the first n throws. Find

$$\lim_{n \rightarrow \infty} \mathbb{P}(X_n \text{ is a multiple of } 13).$$

The speed of this convergence is of order $|\zeta|^n$ where ζ is the biggest eigenvalue of P different from 1.

In our example, P satisfies the hypothesis of the theorem above and $|\zeta| = 0.05$. Hence the stationary regime is reached really fast, as shown by the computations below

$$P^2 = \begin{pmatrix} 0.885 & 0.045 & 0.07 \\ 0.87 & 0.035 & 0.095 \\ 0.88 & 0.04 & 0.08 \end{pmatrix} \quad \text{and} \quad P^3 = \begin{pmatrix} 0.884 & 0.04425 & 0.07175 \\ 0.8835 & 0.0435 & 0.073 \\ 0.884 & 0.044 & 0.072 \end{pmatrix}.$$

One can say

Every minute, the probability that Doudou sleeps is about 88.4%.

Does it mean that Doudou sleeps about 88.4% of the day?

1.4 Ergodic theorems

Ergodic theorems concern the limiting behaviour of averages over time. An essential tool in probability theory is the following ergodic theorem for independent random variables.

Theorem 1.14 (Strong law of large numbers). *Let (Y_n) be a sequence of i.i.d. real valued random variables such that $\mathbb{E}(|Y_1|) < \infty$. Then*

$$\frac{Y_1 + \cdots + Y_n}{n} \xrightarrow[n \rightarrow \infty]{a.s.} \mathbb{E}(Y_1),$$

that is,

$$\mathbb{P}\left(\frac{Y_1 + \cdots + Y_n}{n} \xrightarrow[n \rightarrow \infty]{} \mathbb{E}(Y_1)\right) = 1.$$

In general, the random variables (X_n) are not independent for a Markov chain, but there is an equivalent of the strong law of large numbers.

Theorem 1.15. *Suppose S finite. Let P be an irreducible and aperiodic transition matrix. Then,*

- P has a unique invariant distribution π .
- For any bounded function $f : S \rightarrow \mathbb{R}$, we have

$$\frac{1}{n} \sum_{i=1}^n f(X_i) \xrightarrow[n \rightarrow \infty]{a.s.} \bar{f} = \sum_{x \in S} \pi(x) f(x).$$

Let $V_S(n)$ denote the number of minutes Doudou has slept before n . One can write

$$V_S(n) = f(X_1) + \cdots + f(X_n), \quad \text{with} \quad f(x) = \delta_S(x).$$

Then, $V_S(n)/n$ is the proportion of time before n Doudou slept and

$$\mathbb{P} \left(\frac{V_S(n)}{n} \rightarrow \pi(S) \approx 88.4\% \text{ as } n \rightarrow \infty \right) = 1.$$

Conclusion.

Doudou sleeps about 88.4% of the day.

Understand “Doudou is really **une grosse feignasse**”.

One could argue that the choice of parameters we made in the transition matrix P could not have provided another result. This is right but, in fact it is possible to estimate the parameters of the transition matrix with the following theorem.

Theorem 1.16. *Suppose S finite. Let P be an irreducible and aperiodic transition matrix. Suppose that (X_n) is Markov(π, P) with π the unique invariant distribution of P . Then*

$$\frac{1}{n} \sum_{i=1}^n f(X_{i-1}, X_i) \xrightarrow[n \rightarrow \infty]{a.s.} \sum_{x,y \in S} f(x,y) \pi(x) p(x,y),$$

for any bounded function $f : S \times S \rightarrow \mathbb{R}$.

Hence,

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}\{X_{i-1} = x, X_i = y\} \xrightarrow[n \rightarrow \infty]{a.s.} \pi(x) p(x,y).$$

1.5 Modelling DNA sequences

Let us turn to a simple example to understand why Markov models are important for DNA sequences.

Example: CpG islands. One knows that regions of the DNA are more prone to mutations than others and that some of them are related to the dinucleotide CpG. Here and later on in this lecture, we use the notation “CpG” as a shorthand for “5' – CG – 3'”. In CpG dinucleotides of mammalian genomes, the cytosine is frequently methylated¹ (see [Bir80]). Methylated CpG dinucleotides may change into TpG with higher frequency, and consequently into CpA on the complementary strand. There exist other hotspots in DNA, but this one is of particular interest to the biologist because of the existence of CpG islands in mammalian genomes (see

¹Cytosines in CpG dinucleotides are methylated by DNA methyltransferases in many eukaryotic organisms to form 5-methylcytosine. In mammals, 70% to 80% of CpG cytosines are methylated.

[Bir86], [AB91b], [AB91a]). A CpG island is a region of DNA that have a higher concentration of CpG sites, and frequently these regions are functional ones (see [AB99]). This suggests that the process of methylation is repressed in such regions, and the comprehension of this phenomenon is important in biology.

What sort of probabilistic model might we use for CpG island region? We know that dinucleotides are important. We therefore want a model that generates sequences in which the probability of a symbol depends on the previous symbol. The simplest such model is a standard Markov chain.

From now on, the random vector $X_{1:N}$ represents a genomic sequence of length N , that is, a random string of N symbols taken in the nucleotidic alphabet $\mathcal{A} = \{a, t, c, g\}$.

Estimation of the order of the model. Definition 1.1 contains the description of an homogeneous Markov chain of order 1, that is, a process where the next state depends only on the current state. This definition can be extended. For instance if we wish the next state to depend not only on the current state but also on the state just before. Such a process is also a Markov chain, but with order 2. In this section, we explain how to choose the “right” order for the sequence we observe.

Definition 1.17. We say that (Y_n) is an homogeneous Markov chain of order m if

- (Markov property) for $n \geq m$ and $y_1, \dots, y_{n+1} \in S$

$$\mathbb{P}(Y_{n+1} = y_{n+1} \mid Y_{1:n} = x_{1:n}) = \mathbb{P}(Y_{n+1} = y_{n+1} \mid Y_{n-m+1:n} = y_{n-m+1:n});$$
- (Homogeneity) the transition probability $\mathbb{P}(Y_{n+1} = y \mid Y_{n-m+1:n} = x)$ does not depend on n , for any $(x, y) \in S^m \times S$. We denote this transition probability by $p(x, y)$.

The distribution of $Y_{1:m}$ is called the initial distribution of the Markov chain, and we call transition matrix, the matrix (with possibly infinite dimensions) defined as

$$P = (p(x, y))_{x, y \in S^m \times S}.$$

The matrix P fullfills

$$p(x, y) \geq 0 \quad \text{and} \quad \sum_{x \in S^m} p(x, y) = 1.$$

Remark 1.18. When $S = \mathcal{A}$, the number of free parameters for a Markov($\mu, P, 1$) is 15, 3 for the invariant distribution and 4×3 for the transition matrix. For a Markov(μ, P, m), this number is $(4^m - 1) + 4^m \times 3$. One can see the cost of the extension of the memory of a Markov chain.

From Theorem 1.3, we know that the probability to observe the DNA sequence $x_{1:N}$ under a Markov chain (μ, P) of order 1 is

$$\mathbb{P}(X_{1:N} = x_{1:N}) = \mu(x_0)p(x_0, x_1)p(x_1, x_2) \dots p(x_{N-1}, x_N).$$

Such a formula can be generalized for Markov chains of order m

Theorem 1.19. *If a discrete time random process (Y_n) is Markov (μ, P, m) , then for any $n \geq m$ and any $y_{1:n} \in S^n$*

$$\mathbb{P}(Y_{1:n} = y_{0:n}) = \mu(y_{1:m})p(y_{1:m}, y_{m+1})p(y_{2:m+1}, y_{m+2}) \cdots p(y_{n-m+1}, y_n).$$

Example 1.20. *Consider the sequence of length 8*

$$\text{seq} = \text{aatatagc}.$$

If we assume that this sequence is the realisation of a Markov $(\mu, P, 1)$, then we have

$$(1.1) \quad \mathbb{P}(X_{1:8} = \text{seq}) = \mu(\text{a})p(\text{a}, \text{a})p(\text{a}, \text{t})p(\text{t}, \text{a})p(\text{a}, \text{t})p(\text{t}, \text{a})p(\text{a}, \text{g})p(\text{g}, \text{c}).$$

If we assume that this sequence is the realisation of a Markov $(\mu, P, 2)$, then we have

$$(1.2) \quad \mathbb{P}(X_{1:8} = \text{seq}) = \mu(\text{aa})p(\text{aa}, \text{t})p(\text{at}, \text{a})p(\text{ta}, \text{t})p(\text{at}, \text{a})p(\text{ta}, \text{g})p(\text{ag}, \text{c}).$$

If we assume that this sequence is the realisation of a Markov $(\mu, P, 3)$, then we have

$$(1.3) \quad \mathbb{P}(X_{1:8} = \text{seq}) = \mu(\text{aat})p(\text{aat}, \text{a})p(\text{ata}, \text{t})p(\text{tat}, \text{a})p(\text{ata}, \text{g})p(\text{tag}, \text{c}).$$

Equation (1.1) can be rewritten

$$\mathbb{P}(X_{1:8} = \text{seq}) = \mu(\text{a})p(\text{a}, \text{a})^1 p(\text{a}, \text{t})^2 p(\text{a}, \text{g})^1 p(\text{g}, \text{c})^1 p(\text{t}, \text{a})^2.$$

One can see that the exponent corresponding to $p(\text{a}, \text{a})$ is equal to 1 which is exactly the number of times the dinucleotide aa appears in the sequence seq. The exponent corresponding to $p(\text{a}, \text{t})$ is equal to 2, the number of times the dinucleotide at appears in the sequence seq. In general, for a Markov chain of order 1

$$\mathbb{P}(X_{1:N} = s) = \mu(s_1) \prod_{(x,y) \in \mathcal{A} \times \mathcal{A}} p(x,y)^{N_s(xy)},$$

where $N_s(xy)$ is the number of apparitions of the dinucleotide xy in the sequence s .

Similarly, Equation (1.2) can be rewritten

$$\mathbb{P}(X_{1:8} = \text{seq}) = \mu(\text{aa})p(\text{aa}, \text{t})^1 p(\text{at}, \text{a})^2 p(\text{ta}, \text{t})^1 p(\text{ta}, \text{g})^1 p(\text{ag}, \text{c})^1.$$

Proposition 1.21. *If the discrete time random vector $X_{1:N}$ is Markov (μ, P, m) with $N \geq m$, then for any $s \in \mathcal{A}^N$*

$$(1.4) \quad \mathbb{P}(X_{1:N} = s) = \mu(s_{1:m}) \prod_{(x,y) \in \mathcal{A}^m \times \mathcal{A}} p(x,y)^{N_s(xy)},$$

where $N_s(xy)$ is the number of apparitions of the word xy in the sequence s .

Given a sequence s , we wonder which Markov model fits better s , that is maximizes $\mathbb{P}(X_{1:N} = s)$. First, we can fix m and look for the “better” Markov model with order m . It suffices to find the set of parameters $\theta = (\mu, P)$ such that

$$\ell_m(\theta|s) = \mu(s_{1:m}) \prod_{(x,y) \in \mathcal{A}^m \times \mathcal{A}} p(x,y)^{N_s(xy)},$$

is maximum. The function $\theta \mapsto \ell_m(\theta|s)$ is called the *likelihood function*. The maximum likelihood theory tells us that the set of parameters $\hat{\theta}_m$ which maximizes ℓ_m given s is

$$\hat{\mu}(x) = \mathbf{1}\{x = s_{1:m}\}, \quad \forall x \in \mathcal{A}^m \quad \text{and} \quad \hat{p}(x,y) = \frac{N_s(xy)}{N_s(x\bullet)}, \quad \forall (x,y) \in \mathcal{A}^m \times \mathcal{A}.$$

Furthermore, we have

$$\ell_{m+1}(\hat{\theta}_{m+1}|s) \geq \ell_{m+1}(\hat{\theta}_m|s),$$

hence the larger the order is, the “better” the model fits. But, the number of parameters increases exponentially with m so we want to penalize it. In the end, we would use a Bayesian information criterion (BIC) to choose our order and as a consequence our model. The BIC is given by

$$BIC = -L_m + 3 \times 4^m \log(N),$$

where L_m is the log-likelihood of the model defined as

$$L_m = \log \left(\prod_{(x,y) \in \mathcal{A}^m \times \mathcal{A}} \hat{p}(x,y)^{N_s(xy)} \right) = \sum_{(x,y) \in \mathcal{A}^m \times \mathcal{A}} N_s(xy) \log \left[\frac{N_s(xy)}{N_s(x\bullet)} \right].$$

2 Hidden Markov models

Standard Markov models are useful to study *homogeneous* stretches of genomic sequences. We already know that mammalian genomes are not homogeneous because of the existence of CpG islands, and we cannot use a standard Markov model.

To understand why, consider that we are in a normal region of the genome, that is, a region where the cytosines are frequently methylated. In such region, there is a lack of dinucleotides *cg*, hence the transition probability $p(c, g)$ is low. However, in a CpG island region, the process of methylation is repressed, and the concentration of dinucleotides *cg* is higher which implies that the transition probability $p(c, g)$ is high.

To take into account this phenomenon, we need to define two Markov chains, one for CpG island region and one for the rest of the genome. But given a genomic sequence, we do not know where the different spots, they are hidden. This is where HMMs come to rescue us.

2.1 Definitions

We consider two random processes which are strongly dependent. One corresponds to the observed process, that is, in our context the DNA sequence. The other one corresponds to the hidden process related to the hidden structure of the observed process, in our context, CpG island region or not. One of our purpose later in this section is to find a way to rebuild this structure.

The hidden process. The hidden process, denoted by $Z_{1:N}$, is a sequence of hidden states with values in a discrete space S . In the context of CpG island regions, $S = \{+, -\}$ where $+$ means that methylation is normal and $-$ means that the methylation is repressed.

Definition 2.1. *The process $Z_{1:N}$ is an homogeneous Markov chain whose parameters are denoted by q where*

$$q(u, v) = \mathbb{P}(Z_{i+1} = v | Z_i = u), \quad \forall u, v \in S$$

and

$$q(u) = \mathbb{P}(Z_1 = u), \quad \forall u \in S.$$

As a consequence,

$$\mathbb{P}(Z_{1:N} = z_{1:N}) = q(z_1) \prod_{i=1}^{N-1} q(z_i, z_{i+1}).$$

The observed process. The observed process, denoted by $X_{1:N}$, is the sequence of the observed symbols with values in a finite alphabet $\mathcal{X}$, in our context nucleotides. In a biological context, it would be nice to keep the dependence between the present symbol and the next one, but for a better understanding of the model we start without this dependence. We assume that for a given state $z \in S$ there are emission probabilities p defined as

$$p_z(x) = \mathbb{P}(X_i = x | Z_i = z), \quad \sum_{x \in \mathcal{X}} p_z(x) = 1, \quad \forall z \in S.$$

Definition 2.2. *We say that a process $X_{1:N}$ is a hidden Markov model M1M0 if*

- *there exists a hidden process $Z_{1:N}$ defined according to Definition 2.1 with parameters q ;*
- *conditional on the sequence $Z_{1:N}$, the random variables $X_1, \dots, X_N$ are independent and*

$$\mathbb{P}(X_{1:N} = x_{1:N} | Z_{1:N} = z_{1:N}) = \prod_{i=1}^N \mathbb{P}(X_i = x_i | Z_i = z_i) = \prod_{i=1}^N p_{z_i}(x_i).$$

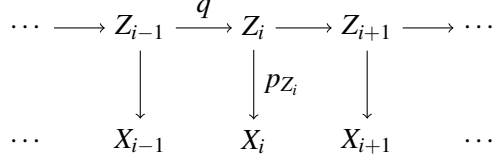


Figure 2: Directed Acyclic Graph of M1M0 model

Such a process can be visualized on Figure 2.

Likelihood of the M1M0 model. Let $\theta = (q, p)$ denote the set of parameters of the M1M0 model. Under θ , the joint probability $\mathbb{P}_\theta(X_{1:N} = x_{1:N}, Z_{1:N} = z_{1:N})$ is given by

$$(2.1) \quad \mathbb{P}_\theta(X_{1:N} = x_{1:N}, Z_{1:N} = z_{1:N}) = q(z_1) \prod_{i=1}^{N-1} q(z_i, z_{i+1}) \prod_{i=1}^N p_{z_i}(x_i).$$

Indeed conditioning on $[Z_{1:N} = z_{1:N}]$ yields

$$\mathbb{P}_\theta(X_{1:N} = x_{1:N}, Z_{1:N} = z_{1:N}) = \mathbb{P}_\theta(X_{1:N} = x_{1:N} | Z_{1:N} = z_{1:N}) \mathbb{P}_\theta(Z_{1:N} = z_{1:N}).$$

On one hand, one has

$$\mathbb{P}_\theta(X_{1:N} = x_{1:N} | Z_{1:N} = z_{1:N}) = \prod_{i=1}^N p_{z_i}(x_i),$$

and on the other hand,

$$\mathbb{P}_\theta(Z_{1:N} = z_{1:N}) = q(z_1) \prod_{i=1}^{N-1} q(z_i, z_{i+1}).$$

A consequence of (2.1) is

$$(2.2) \quad \mathbb{P}_\theta(X_{1:N} = x_{1:N}) = \sum_{z_{1:N} \in \mathcal{S}^N} q(z_1) \prod_{i=1}^{N-1} q(z_i, z_{i+1}) \prod_{i=1}^N p_{z_i}(x_i).$$

2.2 Estimation of θ

Our goal is to describe a procedure to estimate θ given a sequence of observations $x_{1:N} \in \mathcal{X}^N$. We could estimate maximum likelihood theory to do so but we would have to maximize the right hand of (2.2) which is computationally impossible.

Expectation-Maximization. In statistics, an expectation–maximization (EM) algorithm is an iterative method for finding maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model

depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

EM ALGORITHM

Initialization. Given an initial value θ' , given a sequence $x = x_{1:N}$.

Step 1 (Expectation). Compute the function

$$\theta \mapsto Q(\theta, \theta') = \mathbb{E}_{\theta'} (\log [\mathbb{P}_{\theta}(Z_{1:N}, X_{1:N})] | X_{1:N} = x).$$

Step 2 (Maximization). Compute θ'' such that

$$Q(\theta'', \theta') \geq Q(\theta, \theta'), \quad \forall \theta.$$

Step 3. Stop the algorithm or go back to **Step 1** with $\theta' = \theta''$.

E step. The first question is how to compute $Q(\theta, \theta')$. From (2.1), we have

$$\begin{aligned} (2.3) \quad \log [\mathbb{P}_{\theta}(X_{1:N}, Z_{1:N})] &= \log [\mathbb{P}_{\theta}(Z_1)] + \sum_{i=1}^N \log [\mathbb{P}_{\theta}(X_i | Z_i)] + \\ &\quad + \sum_{i=1}^{N-1} \log [\mathbb{P}_{\theta}(Z_{i+1} | Z_i)] \\ &= \sum_{z \in S} \log [q(z)] \mathbf{1}\{Z_1 = z\} + \sum_{i=1}^N \sum_{z \in S} \log [p_z(X_i)] \mathbf{1}\{Z_i = z\} + \\ &\quad + \sum_{i=1}^{N-1} \sum_{y, z \in S} \log [q(y, z)] \mathbf{1}\{Z_i = y, Z_{i+1} = z\}. \end{aligned}$$

As a consequence,

$$\begin{aligned} (2.4) \quad Q(\theta, \theta') &= \sum_{z \in S} \log [q(z)] \mathbb{P}_{\theta'}(Z_1 = z | x) + \sum_{i=1}^N \sum_{z \in S} \log [p_z(x_i)] \mathbb{P}_{\theta'}(Z_i = z | x) + \\ &\quad + \sum_{i=1}^{N-1} \sum_{y, z \in S} \log [q(y, z)] \mathbb{P}_{\theta'}(Z_i = y, Z_{i+1} = z | x). \end{aligned}$$

To compute $Q(\theta, \theta')$ we need to compute

$$\mathbb{P}_{\theta'}(Z_i = z | x), \quad \forall z \in S, \forall i \in \{1, N\},$$

and

$$\mathbb{P}_{\theta'}(Z_i = y, Z_{i+1} = z | x), \quad \forall y, z \in S, \forall i \in \{1, N-1\}.$$

To do this we use a forward-backward procedure.

FORWARD PROCEDURE

Initialization. Given initial values $\mathbb{P}_{\theta'}(Z_1 = z)$, given a sequence $x = x_{1:N}$, given $i = 1$,

Step 1 (Prediction). Compute

$$\mathbb{P}_{\theta'}(Z_{i+1} = y | x_{1:i}) = \sum_{z \in S} q'(z, y) \mathbb{P}_{\theta'}(Z_i = z | x_{1:i}).$$

Step 2 (Filter). Compute

$$\mathbb{P}_{\theta'}(Z_{i+1} = z | x_{1:i+1}) = \frac{p'_j(x_{i+1}) \mathbb{P}_{\theta'}(Z_{i+1} = z | x_{1:i})}{\sum_{z \in S} p'_j(x_{i+1}) \mathbb{P}_{\theta'}(Z_{i+1} = z | x_{1:i})}.$$

Step 3. If $i = N - 1$, stop the algorithm, else go back to **Step 1** with $i = i + 1$.

In the end of the forward procedure, we know

$$\mathbb{P}_{\theta'}(Z_i = z | x_{1:i}) \quad \forall i = 1 \dots N, \quad \forall z \in S,$$

and

$$\mathbb{P}_{\theta'}(Z_{i+1} = z | x_{1:i}) \quad \forall i = 1 \dots N - 1, \quad \forall z \in S.$$

BACKWARD PROCEDURE

Initialization. Given a sequence $x = x_{1:N}$, given initial values from the forward procedure, given $i = N$,

Step 1 (Smoothing). Compute

$$\mathbb{P}_{\theta'}(Z_{i-1} = y, Z_i = z | x) = \frac{\mathbb{P}_{\theta'}(Z_{i-1} = y | x_{1:i}) q'(y, z) \mathbb{P}_{\theta'}(Z_i = z | x)}{\mathbb{P}_{\theta'}(Z_i = z | x_{1:i-1})},$$

and

$$\mathbb{P}_{\theta'}(Z_{i-1} = y | x) = \sum_{z \in S} \mathbb{P}_{\theta'}(Z_{i-1} = y, Z_i = z | x).$$

Step 2. If $i = 2$, stop the algorithm, else go back to **Step 1** with $i = i - 1$.

M step. The maximization step is almost for free because of the shape of $Q(\theta, \theta')$. Indeed, from (2.4) we have

$$\begin{aligned} Q(\theta, \theta') &= \sum_{z \in S} \log[q(z)] \mathbb{P}_{\theta'}(Z_1 = z | x) + \sum_{z \in S} \sum_{y \in \mathcal{Y}} \log[p_z(y)] \sum_{i: x_i = y} \mathbb{P}_{\theta'}(Z_i = z | x) + \\ &\quad + \sum_{y, z \in S} \log[q(y, z)] \sum_{i=1}^{N-1} \mathbb{P}_{\theta'}(Z_i = y, Z_{i+1} = z | x). \end{aligned}$$

As a consequence, $Q(\theta, \theta')$ is maximum at $\theta = (q, p)$ defined as

$$q(z) = \mathbb{P}_{\theta'}(Z_1 = z | x),$$

and

$$q(y, z) = \sum_{i=1}^{N-1} \mathbb{P}_{\theta'}(Z_i = y, Z_{i+1} = z | x),$$

and

$$p_z(y) = \sum_{i: x_i=y} \mathbb{P}_{\theta'}(Z_i = z | x).$$

To understand why, we focus on one example which also helps to understand what happened in the E step. Of course, this is not a proof.

Example. Imagine that P_θ and $P_{\theta'}$ are two probability measures on $\{0, 1\}$ such that

$$P_\theta(1) = 1 - P_\theta(0) = \theta, \quad \text{and} \quad P_{\theta'}(1) = 1 - P_{\theta'}(0) = \theta'.$$

P_θ can be seen as a function on $\{0, 1\}$. As a consequence, for any random variable ξ with values in $\{0, 1\}$, $\log[P_\theta(\xi)]$ is a random variable. Since ξ can take two values, $\log[P_\theta(\xi)]$ can take two values which are

$$\log[P_\theta(1)] = \log \theta \quad \text{and} \quad \log[P_\theta(0)] = \log(1 - \theta).$$

What is the probability of the event $[\log[P_\theta(\xi)] = \log \theta]$? It is exactly the probability of the event $[\xi = 1]$. Similarly, the probability of the event $[\log[P_\theta(\xi)] = \log(1 - \theta)]$ is given by the probability of the event $[\xi = 0]$.

As a consequence, if the distribution of ξ is given by $P_{\theta'}$, we have

$$\mathbb{P}(\log[P_\theta(\xi)] = \log \theta) = \mathbb{P}(\xi = 1) = P_{\theta'}(1) = \theta',$$

and

$$\mathbb{P}(\log[P_\theta(\xi)] = \log(1 - \theta)) = \mathbb{P}(\xi = 0) = P_{\theta'}(0) = 1 - \theta',$$

which yields

$$\mathbb{E}(\log[P_\theta(\xi)]) = \theta' \log \theta + (1 - \theta') \log(1 - \theta).$$

The derivative of the function $\theta \mapsto \mathbb{E}(\log[P_\theta(Z)])$ is given by

$$\theta \mapsto \frac{\theta' - \theta}{\theta(1 - \theta)},$$

which implies that $\mathbb{E}(\log[P_\theta(Z)])$ reaches its maximum at $\theta = \theta'$.

Results. The EM algorithm converges toward a local maximum of the function $\theta \mapsto \mathbb{P}_\theta(X_{1:N} = x)$. To reach the maximum of the function and the corresponding θ^* , it is necessary to run the EM algorithm with different initial values, compute the corresponding log-likelihood and check what happens. The computation of the log-likelihood can be done during the forward procedure because

$$\mathbb{P}_\theta(X_{1:i+1}) = \sum_z p_z(x_{i+1}) \mathbb{P}_\theta(Z_{i+1} = z | x_{1:i}).$$

Right now, it is not possible to find the structure of the sequence given the observations. To do so, we introduce a new procedure which is a stochastic version of the EM algorithm.

2.3 Reconstruction of the hidden structure

Once we have an estimation $\hat{\theta}$ of θ , it is possible to compute an approximation $\mathbb{P}_{\theta}(Z_{1:N} = z|x)$ for any z in S^N because

$$\mathbb{P}_{\theta}(Z_{1:N} = z_{1:N}|x) = q(z_1) \prod_{i=1}^{N-1} q(z_i, z_{i+1}).$$

The problem is to find z such that this quantity is maximum. Because of the too large possibilities for z , it is not possible to compute all the quantities and make a comparison. One possible compromise is the stochastic EM algorithm.

SEM ALGORITHM

Initialization. Given an initial value θ' , given a sequence $x = x_{1:N}$.

Step 1 (Expectation). Compute

$$\mathbb{P}_{\theta'}(Z_i = z|x), \quad \forall z \in S, \forall i \in \{1, N\},$$

with the forward-backward procedure.

Step 2 (Simulation). Simulate one realization $z'_{1:N}$ of a random vector $Z'_{1:N}$ such that

$$\mathbb{P}(Z'_i = z) = \mathbb{P}_{\theta'}(Z_i = z|x).$$

Step 3 (Maximization). Compute θ'' such that

$$R(\theta'', \theta') \geq R(\theta, \theta'), \quad \forall \theta,$$

with

$$R(\theta, \theta') = \log[q(z'_1)] + \sum_{i=1}^{N-1} \log[q(z'_i, z'_{i+1})] + \sum_{i=1}^N \log[p_{z'_i}(x_i)].$$

Step 4. Stop the algorithm or go back to **Step 1** with $\theta' = \theta''$.

Note that $R(\theta, \theta')$ is the computation of the logarithm of the complete likelihood

$$\mathbb{P}_{\theta}(X_{1:N} = x_{1:N}, Z_{1:N} = z'_{1:N}).$$

Results. In the end of the SEM algorithm, we obtain the posterior distribution of the states (Z_i) given the sequence x . Hence it is possible to simulate the hidden structure under this distribution or compute the *maximum a posteriori* estimate.

2.4 M1M1 model

In the M1M0 model, conditional on the sequence $Z_{1:N}$, the random variables $X_1, \dots, X_N$ are independent. To fit genomic data, we would like to modify the M1M0 model such that X_{i+1} depends on X_i . This is the M1M1 model.

Definition 2.3. We say that a process $X_{1:N}$ is a hidden Markov model M1M1 if

- there exists a hidden process $Z_{1:N}$ defined according to Definition 2.1 with parameters q ;
- conditional on the sequence $Z_{1:N}$, the joint distribution of the random variables $X_1, \dots, X_N$ is

$$\begin{aligned} \mathbb{P}(X_{1:N} = x_{1:N} | Z_{1:N} = z_{1:N}) \\ &= \mathbb{P}(X_1 = x_1 | Z_1 = z_1) \prod_{i=1}^{N-1} \mathbb{P}(X_{i+1} = x_{i+1} | Z_{i+1} = z_{i+1}, X_i = x_i) \\ &= p_{z_1}(x_1) \prod_{i=1}^{N-1} p_{z_{i+1}}(x_i, x_{i+1}), \end{aligned}$$

where $(p_z(x))_{x \in \mathcal{X}}$ is an emission probability for any $z \in S$ and $(p_z(x, y))_{x, y \in \mathcal{X}}$ is a transition matrix for any $z \in S$.

Such a process can be visualized on Figure 3.

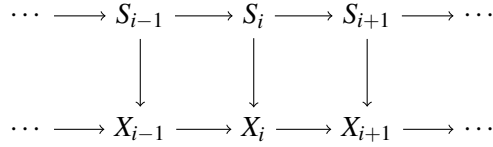


Figure 3: directed Acyclic Graph of M1M1 model

Likelihood of the M1M1 model. Let $\theta = (q, p)$ denote the set of parameters of the M1M1 model. Under θ , the joint probability $\mathbb{P}_\theta(X_{1:N} = x_{1:N}, Z_{1:N} = z_{1:N})$ is given by

$$(2.5) \quad \mathbb{P}_\theta(X_{1:N} = x_{1:N}, Z_{1:N} = z_{1:N}) = q(z_1) p_{z_1}(x_1) \prod_{i=1}^{N-1} q(z_i, z_{i+1}) \prod_{i=1}^{N-1} p_{z_{i+1}}(x_i, x_{i+1}).$$

A consequence of (2.5) is

$$(2.6) \quad \mathbb{P}_\theta(X_{1:N} = x_{1:N}) = \sum_{z_{1:N} \in S^N} q(z_1) p_{z_1}(x_1) \prod_{i=1}^{N-1} q(z_i, z_{i+1}) \prod_{i=1}^{N-1} p_{z_{i+1}}(x_i, x_{i+1}).$$

Results. All the procedures introduced for the M1M0 model can be adapted to the M1M1 model. See [DEKM98] and [Nic03] for more details.

3 One example of nucleotidic substitution process

The Jukes-Cantor model [JC69] is the simplest model of nucleotidic substitution processes in DNA sequences. Relying on the theory of Markov processes, this probabilistic model describes changes between nucleotides. In this section, we present a mathematical description of the model and the main tools to estimate distances between DNA sequences.

3.1 Mathematical description of the model

The Jukes-Cantor model (JC) is a continuous-time Markov chain. In JC, one assumes that the nucleotidic sites evolve independently from the others, in a similar way. Hence, the evolution of the DNA sequence is ruled by the independent parallel evolutions of N nucleotidic sites. Another assumption of JC is that every substitution occurs at the same rate.

3.1.1 Dynamics of one site

We present now the dynamics of one nucleotidic site in JC.

Definition 3.1. Let $X_i(t)$ denote the random value of the nucleotide at site i and time t . In JC, the process $(X_i(t))_{t \geq 0}$ is a Markov process on $\mathcal{A}$ and there exists a parameter $\lambda > 0$ such that the infinitesimal generator $Q = (q_{xy})_{xy}$ is given by the 4×4 matrix of substitution rates

$$Q = \begin{matrix} & \begin{matrix} A & T & C & G \end{matrix} \\ \begin{matrix} A \\ T \\ C \\ G \end{matrix} & \begin{pmatrix} -3\lambda & \lambda & \lambda & \lambda \\ \lambda & -3\lambda & \lambda & \lambda \\ \lambda & \lambda & -3\lambda & \lambda \\ \lambda & \lambda & \lambda & -3\lambda \end{pmatrix} \end{matrix}.$$

Each off-diagonal entry q_{xy} is the rate of substitution of the nucleotide x by the nucleotide y , that is, during a small interval of time dt , x is replaced by y with probability $q_{xy}dt$. Each number $-q_{xx}$ can be interpreted as the rate of substitution of the nucleotide x . This means that site i occupied by nucleotide x is modified after an exponentially distributed random time with mean $-1/q_{xx}$, and that, when it is modified, it becomes occupied by nucleotide y with probability $-q_{xy}/q_{xx}$, for each $y \neq x$.

In JC, this means for example that nucleotide C becomes an A , a T or a G with probability $\frac{1}{3}$ each and that the rate of substitution of C is 3λ . The parameter λ rules the overall rate of evolution.

The infinitesimal generator Q fully determines the dynamics of the Markov chain. For instance, let $p_{xy}(t)$ denote the probability that site i is occupied by y at time t

given that it is occupied by x at time 0, that is,

$$p_{xy}(t) = \mathbb{P}(X_i(t) = y | X_i(0) = x).$$

Then, the transition-probability matrix over any time t , denoted $P(t) = (p_{xy}(t))$, is determined Q as the unique solution of the initial value problem

$$\frac{dP(t)}{dt} = QP(t), \quad P(0) = \text{Identity},$$

The solution is

$$P(t) = e^{Qt}.$$

The reader will find a complete presentation of continuous-time Markov chains on discrete spaces in [Nor97]. We now provide the transition-probability matrix for JC.

Proposition 3.2. *In JC, for every x and every $y \neq x$ in $\mathcal{A}$, $p_{xy}(t) = \frac{1}{3}p(t)$ and $p_{xx}(t) = 1 - p(t)$, with*

$$p(t) = \frac{3}{4} \left(1 - e^{-4\lambda t} \right).$$

Proof. Let I denote the 4×4 identity matrix and J the 4×4 matrix whose every coefficient is $\frac{1}{4}$. Then, $IJ = JI = J^2 = J$ and $Q = 4\lambda(J - I)$. Hence,

$$e^{Qt} = e^{-4\lambda t I} e^{4\lambda t J} = e^{-4\lambda t} e^{4\lambda t J}.$$

For every integer $n \geq 1$, $J^n = J$, hence

$$e^{4\lambda t J} = \sum_{n \geq 0} \frac{(4\lambda t)^n}{n!} J^n = I + (e^{4\lambda t} - 1)J.$$

This means that $e^{Qt} = e^{-4\lambda t} I + (1 - e^{-4\lambda t})J$, which implies the result. $\square$

The transition-probability matrix coupled with an initial distribution for $X_i(0)$ fully determines the law of $X_i(t)$. Indeed, assume that $X_i(0)$ has the initial distribution $\pi = (\pi_x)_{x \in \mathcal{A}}$, that is

$$\mathbb{P}(X_i(0) = x) = \pi_x,$$

then the law of $X_i(t)$, denoted $\pi(t)$, is given by $\pi(t) = \pi P(t)$, that is, for every nucleotide x

$$\pi_x(t) = \mathbb{P}(X_i(t) = x) = \sum_{y \in \mathcal{A}} \pi_y p_{yx}(t).$$

3.1.2 Dynamics of the DNA sequence

Until now, we have presented the evolution of one site. Let $X_{1:N}(t)$ be a shorthand for $(X_1(t), X_2(t), \dots, X_N(t))$ and denote the random value of the N nucleotides from site 1 to site N at time t in the DNA sequence.

Definition 3.3. In JC, the process $X_{1:N}(t)$ is Markov, with initial distribution π , where π is a distribution on $\mathcal{A}^N$ and transition kernel $Q^{\otimes N}$. In other words, one has for every $x_{1:N} = (x_1, x_2, \dots, x_N) \in \mathcal{A}^N$

$$\mathbb{P}(X_{1:N}(t) = x_{1:N}) = \prod_{i=1}^N \mathbb{P}(X_i(t) = x_i) = \prod_{i=1}^N (\pi e^{Qt})_{x_i}.$$

Definition 3.3 is just a translation of the independence of the N sites and their identically distributions.

When the length N of the sequence is large, the number π_x can be interpreted as the initial proportion of nucleotides x in the sequence, as a consequence of the law of large numbers. Indeed, one has

$$\frac{1}{N} \sum_{i=1}^N \mathbf{1}\{X_i(0) = x\} \xrightarrow[N \rightarrow +\infty]{a.s.} \mathbb{E}(\mathbf{1}\{X_1(0) = x\}) = \pi_x.$$

Similarly, $\pi_x(t)$ can be interpreted as the proportion of nucleotides x at time t in the sequence.

The distribution π^* on $\mathcal{A}$, defined as $\pi_x^* = 1/4$ for every nucleotide x , is *stationary* for the Jukes-Cantor process on one site. This means that if the initial distribution of X_i is π^* , then for every time t the distribution of $X_i(t)$ is π^* . Furthermore, the Jukes-Cantor process is *ergodic*, which means for example that the law of $X_i(t)$ converges to π^* when t tends to infinity, for any initial distribution of X_i .

As a consequence, the proportion of nucleotides x in the DNA sequence when t is large is almost $1/4$, and one can imagine that the sequence is really blended in comparison of the initial sequence, and that pick some information in the sequence may be difficult.

3.2 Distance estimation

Since we have presented the Jukes-Cantor model, we now discuss the possibility to calculate pairwise distances between DNA sequences.

3.2.1 One Estimator of the time elapsed between ancestral and present sequences

Consider two DNA sequences of length N . Assume that one sequence is ancestral to the other and that the present sequence has evolved under the Jukes-Cantor

model since an unknown time t . This means that every nucleotide site i in the past sequence is ancestral to nucleotide site i in the present sequence. We want to provide an estimator of the elapsed time between these two sequences.

First of all, note that the law of $X_i(t)$ under the Jukes-Cantor model with parameter λ is similar to the law of $X_i(t/2)$ under the Jukes-Cantor model with parameter 2λ . Indeed, in the transition-probability matrix $P(t)$, these quantities appear only in the form of a product λt . Thus, without external information on λ , there is no hope to estimate t in function of λ . Yet, 3λ is the global rate of substitutions per site per unit of time. Given the organism, the nature of DNA, or the area of DNA we look at, maybe some information can be provided on λ , but that is not our purpose here.

In Jukes-Cantor model, the estimator used for the elapsed time is simply based on the proportion of different sites in the two sequences. This is a direct application of the maximum likelihood method.

Precisely, let P_{obs} denote the observed quantity defined as

$$P_{\text{obs}} = \frac{1}{N} \sum_{i=1}^N K_i(t), \quad \text{with} \quad K_i(t) = \mathbf{1}\{X_i(t) \neq X_i(0)\}.$$

The random variables $(K_i(t))_{i=1}^N$ are Bernoulli random variables identically distributed, and their common mean is $p(t)$. Indeed, for every initial distribution π on $\mathcal{A}$, one has

$$\mathbb{P}(X_i(t) = X_i(0)) = \sum_{x \in \mathcal{A}} \pi_x p_{xx}(t) = \sum_{x \in \mathcal{A}} \pi_x (1 - p(t)) = 1 - p(t).$$

The random variables $(K_i(t))_{i=1}^N$ are also independent, then the law of large numbers and central limit theorem provide

$$(3.1) \quad P_{\text{obs}} \xrightarrow[N \rightarrow +\infty]{a.s.} p(t), \quad \sqrt{N}(P_{\text{obs}} - p(t)) \xrightarrow[N \rightarrow +\infty]{d.} \mathcal{N}(0, p(t)(1 - p(t))),$$

where $\mathcal{N}(\alpha, \sigma^2)$ stands for a normal law with mean α and variance σ^2 .

Now we explain how it is possible to compute an estimator for λt from P_{obs} . The function $d \mapsto \frac{3}{4}(1 - e^{-4d})$ is increasing on $[0, +\infty[$ from the value 0 at $d = 0$ to the value $3/4$ at $d = +\infty$, and every value p in the interval $[0, 3/4[$ corresponds to a unique value d in the interval $[0, +\infty[$ via this function.

This one to one correspondence allows to define the estimator D of the time elapsed.

Definition 3.4. Let D denote the estimator of the time elapsed defined as the solution in d of the equation

$$\frac{3}{4}(1 - e^{-4d}) = P_{\text{obs}}.$$

3.2.2 Consistency of the estimator and asymptotic confidence interval

Consistency of estimator D is a consequence of the almost convergence of P_{obs} and the continuity of the reciprocal function of $d \mapsto \frac{3}{4}(1 - e^{-4d})$.

Proposition 3.5. *The random variable D converges almost surely to λt when N tends to infinity. Hence, D is a consistent estimator of λt .*

Consistency is a nice property for an estimator, but we are also interested in an asymptotic confidence interval for estimator D . Thanks to central limit theorem, we already have an asymptotic confidence interval for $p(t)$, and we now use the delta method [vdV98] to provide an asymptotic confidence interval for t .

Citing [vdV98], the delta method consists of using a Taylor expansion to approximate a random vector of the form $\varphi(T_N)$ by the polynomial $\varphi(\theta) + \varphi'(\theta)(T_N - \theta) + \dots$ in $T_N - \theta$. It is a simple but useful method to deduce the limit law of $\varphi(T_N) - \varphi(\theta)$ from that of $T_N - \theta$.

Proposition 3.6 (delta method). *Let $\varphi : I \subset \mathbb{R} \rightarrow \mathbb{R}$ be a map defined on an interval I of $\mathbb{R}$ and differentiable at θ . Let (T_N) be random variables taking their values in the domain of φ . Assume that $r_N(T_N - \theta)$ converges in distribution to a random variable T , with $r_N \rightarrow +\infty$.*

Then $r_N(\varphi(T_N) - \varphi(\theta))$ converges in distribution to $\varphi'(\theta)T$.

The situation we have with D and P_{obs} is exactly these of proposition 3.6. Indeed, let μ denote the reciprocal function of $d \mapsto \frac{3}{4}(1 - e^{-4d})$. Function μ is differentiable on $[0, 3/4[$, and

$$D = \mu(P_{\text{obs}}), \quad \lambda t = \mu(p(t)).$$

From convergence in distribution (3.1), we deduce

Corollary 3.7. *The random variables $\sqrt{N}(D - \lambda t)$ converges in distribution to the centered normal law with variance $\sigma^2(t)$, where*

$$\sigma^2(t) = \frac{p(t)(1 - p(t))}{(3 - 4p(t))^2}.$$

To build a confidence interval for λt from corollary 3.7 requires to know the value of $p(t)$ which depends on the quantity λt to be estimated. Slutsky's lemma (see [vdV98]) allows to bypass this difficulty. Indeed, Slutsky's lemma states that if two sequence of random variable $(X_N)_N$ and $(Y_N)_N$ are such that $(X_N)_N$ converges in distribution to a random variable X and $(Y_N)_N$ converges in probability to a constant c then the sequence $(X_N Y_N)_N$ converges in distribution to the random variable cX . In our situation, we apply this lemma by using the fact that

$$\frac{3 - 4P_{\text{obs}}}{\sqrt{P_{\text{obs}}(1 - P_{\text{obs}})}} \xrightarrow[N \rightarrow +\infty]{a.s.} \frac{3 - 4p(t)}{\sqrt{p(t)(1 - p(t))}}.$$

Hence from central limit theorem for P_{obs} , delta method and Slutsky's lemma, one has the final result.

Theorem 3.8. *In the Jukes-Cantor model,*

$$(3 - 4P_{\text{obs}}) \sqrt{\frac{N}{P_{\text{obs}}(1 - P_{\text{obs}})}} (D - \lambda t) \xrightarrow[N \rightarrow +\infty]{d.} \mathcal{N}(0, 1).$$

At the moment, we have only provided an estimator of the time elapsed between an ancestral sequence and a present one. Remember that in most cases, data are not available for extinct ancestors. We now explain how it is possible to derive an estimator of the time of divergence between two present sequences, that is, a phylogenetic distance between these two DNA sequences

3.2.3 Estimation of the time of divergence between two present sequences

In the case of the Jukes-Cantor model, we provide an estimator for the time of divergence between two present sequences directly from the estimator of the time elapsed between an ancestral sequence and a present one. However, it is necessary to assume that the ancestral sequence is at stationarity.

The Jukes-Cantor model is a time-reversible Markov chain. Reversibility means that the dynamics will look the same whether time runs forward or backward. To check if a Markov process is time-reversible, one can use the generalized Kolmogorov criterion (Kendall [Ken59]), that is, one needs to check that for each loop in the state space, the product of the rates is the same whatever direction in the loop is chosen.

Proposition 3.9 (Generalized Kolmogorov criterion). *A stationary Markov process with infinitesimal generator Q and state space $\mathcal{S}$ is reversible if and only if the entries of Q satisfy*

$$q_{x_1 x_2} q_{x_2 x_3} \cdots q_{x_{n-1} x_n} q_{x_n x_1} = q_{x_1 x_n} q_{x_n x_{n-1}} \cdots q_{x_3 x_2} q_{x_2 x_1},$$

for any elements $x_1, x_2, \dots, x_n \in \mathcal{S}$.

As a result, given two sequences at stationarity, the probability of data in a state is the same whether one sequence is ancestral to the other or both are descendants of an ancestral sequence at stationarity. Roughly speaking, for every x and y that belong to $\mathcal{A}$, going from a x at time t to 0 then back to a y at time t on another branch, is equivalent to going from a x at time 0 to a y at time $2t$, as illustrated in figure 4.

To understand why stationarity is necessary, imagine the following case. Let the time run forward largely from the ancestral sequence to sequence 1. Then, sequence 1 is almost at stationarity. As a consequence, the proportion of each nucleotide in

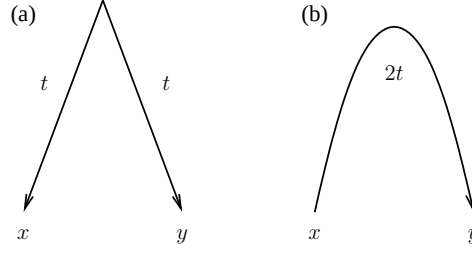


Figure 4: A tree for two sequences showing the observed nucleotides x and y at one site. (a) Two sequences diverged from a common ancestor. (b) Sequence 1 is ancestral to sequence 2.

sequence 1 is almost $1/4$. Now, let the time run backward from sequence 1 to the ancestral sequence, as the time is large, the ancestral sequence is also at stationarity. Stationarity is necessary for time-reversible process.

Let $X_{1:N}^k(t)$ denote for every $k \in \{1, 2\}$, the random value of the N nucleotides from site 1 to site N at time t in the DNA sequence k .

Let $\tilde{P}_{\text{obs}}$ denote the observed quantity defined as

$$\tilde{P}_{\text{obs}} = \frac{1}{N} \sum_{i=1}^N \tilde{K}_i(t), \quad \text{with} \quad \tilde{K}_i(t) = \mathbf{1}\{X_i^1(t) \neq X_i^2(t)\}.$$

Definition 3.10. Let $\tilde{D}$ denote the estimator of the time elapsed defined as the solution in d of the equation

$$\frac{3}{4}(1 - e^{-8d}) = \tilde{P}_{\text{obs}}.$$

The main result of this section is the following.

Theorem 3.11. Assume that two present DNA sequences have diverged under the Jukes-Cantor model with parameter λ from a common ancestral DNA sequence at equilibrium since a time t .

Then, $\tilde{D}$ is a consistent estimator of λt , and

$$(6 - 8\tilde{P}_{\text{obs}}) \sqrt{\frac{N}{\tilde{P}_{\text{obs}}(1 - \tilde{P}_{\text{obs}})}} (\tilde{D} - \lambda t) \xrightarrow[N \rightarrow +\infty]{d.} \mathcal{N}(0, 1).$$

References

- [AB91a] Brahim Aïssani and Giorgio Bernardi, *CpG islands, genes and isochores in the genomes of vertebrates*, Gene **106** (1991), 185–195.

- [AB91b] Brahim Aïssani and Giorgio Bernardi, *CpG islands: features and distribution in the genome of vertebrates*, *Gene* **106** (1991), 173–183.
- [AB99] Francisco Antequera and Adrian P. Bird, *CpG islands as genomic footprints of promoters that are associated with replication origins*, *Current Biology* **9** (1999), R661–R667.
- [Bir80] Adrian P. Bird, *DNA methylation and the frequency of CpG in animal DNA*, *Nucleic Acids Research* **8** (1980), 1499–1504.
- [Bir86] ———, *CpG-rich islands and the function of DNA methylation*, *Nature* **321** (1986), 209–213.
- [DEKM98] Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison, *Biological sequence analysis Probabilistic models of proteins and nucleic acids*, Cambridge University Press, New York, 1998.
- [JC69] Thomas Hughes Jukes and Charles R. Cantor, *Mammalian protein metabolism*, ch. Evolution of Protein Molecules, pp. 21–132, Academic Press, New York, 1969.
- [Ken59] David G. Kendall, *Unitary dilations of one-parameter semigroups of Markov transition operators, and the corresponding integral representations for Markov processes with countable infinity of states*, *Proceedings of the London Mathematical Society* **9(3)** (1959), 417–431.
- [Nic03] Pierre Nicolas, *Mise au point et utilisation de modèles de chaînes de Markov cachées pour l'étude des séquences ADN*, Ph.D. thesis, Université d'Évry, 2003.
- [Nor97] James R. Norris, *Markov Chains*, ch. Continuous-time Markov chains I, Cambridge university press, 1997.
- [vdV98] Aad W. van der Vaart, *Asymptotic Statistics*, Cambridge University Press, 1998.