

Summary

Continuous-time finite-state-space Markov chains are stochastic processes that are widely used to model the process of nucleotide substitution. This chapter aims to present much of the mathematics behind the models, but we stop short of a completely rigorous introduction to continuous-time Markov chains. This first section will summarize the main results of the chapter and introduce notation. Details supporting the summary statements follow.

A *stochastic process* $\{X(t) : t \geq 0\}$ with finite state space Ω is a family of discrete random variables taking values in Ω indexed by a continuous parameter t , thought of as *time*. Throughout the chapter, it will be useful to think about Ω as being the set $\{A, C, G, T\}$ for the four DNA bases, although all of the results are equally valid for other different finite sets (such as the twenty amino acids or 61 non-stop codons). Furthermore, many of the results hold for countably infinite state spaces. We use the symbol $d = |\Omega|$ to represent the number of symbols in Ω .

Informally, the *Markov property* says that given the present, the past and the future are independent. Suppose that we know the value of the process at the k increasing time points $t_0 < t_1 < \dots < t_{k-1}$ and we desire the conditional distribution of the process at time $t_k > t_{k-1}$. The Markov property says that this conditional distribution depends only on the most recently observed time and is independent of past events.

$$\mathbb{P}(X(t_k) = i_k \mid X(t_0) = i_0, X(t_1) = i_1, \dots, X(t_{k-1}) = i_{k-1}) = \mathbb{P}(X(t_k) = i_k \mid X(t_{k-1}) = i_{k-1})$$

We also consider only *homogeneous* Markov chains where the probabilities do not change after shifts in time, namely

$$\mathbb{P}(X(s+t) = j \mid X(s) = i) = \mathbb{P}(X(t) = j \mid X(0) = i)$$

for all $s, t \geq 0$.

Each finite-state-space continuous-time Markov chain is completely described by an *initial probability vector* $\mathbf{p}(0)$, whose i th element is the probability that the chain begins in state i at time 0, and an *infinitesimal rate (or generator) matrix* $\mathbf{Q} = \{q_{ij}\}$. These two objects determine the distribution of $\mathbf{p}(t)$, the vector of probabilities for each state at time t . We will consider $\mathbf{p}(t)$ to be a $1 \times d$ row vector while \mathbf{Q} is a $d \times d$ matrix.

A word on notation. It is conventional to represent vectors as columns of numbers. Row vectors are typically denoted with a superscript T . To ease notation, in this chapter we defy convention for the vectors $\mathbf{p}(t)$, $\mathbf{p}(0)$, and $\boldsymbol{\pi}$, all of which should be thought of as row vectors despite the missing superscript Ts. However, other vectors such as $\mathbf{0}$, $\mathbf{1}$, and \mathbf{v} should be thought of as column vectors. We follow conventional mathematics notation where matrices are boldface capital letters and vectors are boldface lower case letters.

The off-diagonal elements of \mathbf{Q} are nonnegative and can be interpreted as the rate of transitions from i directly to j given the process is in state i . The diagonal elements of \mathbf{Q} equal the negative row sums of the off diagonal elements, so that

$$q_{ii} = - \sum_{j \neq i} q_{ij} .$$

The $d \times d$ *Markov probability transition matrix* $\mathbf{P}(t) = \{p_{ij}(t)\}$ contains the probabilities that the chain is in state j at time t given we began in state i for each pair of states. Transition matrices

satisfy the expression

$$\mathbf{P}(s + t) = \mathbf{P}(s)\mathbf{P}(t)$$

for each $s, t \geq 0$. We can also represent the transition matrices with a *matrix exponential*.

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \sum_{k=0}^{\infty} \frac{\mathbf{Q}^k t^k}{k!}$$

for \mathbf{Q} defined as above where we understand \mathbf{Q}^0 to be the $d \times d$ identity matrix \mathbf{I} with ones down the main diagonal and zeros elsewhere. We can write $\mathbf{p}(t) = \mathbf{p}(0)\mathbf{P}(t)$

The *dwell time* (waiting time) in state i before a transition to some other state is exponentially distributed with rate $q_i = -q_{ii}$. Given that a transition from i occurs, the probability that the transition is to state j is q_{ij}/q_i .

If we assume that the Markov chain is *irreducible*, meaning that it is possible to get from any state i to any state j in a finite amount of time, it follows that the Markov chain is *positive recurrent*, meaning that the expected return time for each state is finite. (For infinite state spaces, irreducibility need not imply positive recurrence.) These two conditions suffice to imply the existence of a unique *stationary distribution* $\boldsymbol{\pi} = \{\pi_i\}$. If $\mathbf{p}(0) = \boldsymbol{\pi}$, then $\mathbf{p}(t) = \boldsymbol{\pi}$ for all t . In addition, the limiting transition probabilities as time goes to infinity do not depend on the initial state and take the values of the stationary distribution.

$$\lim_{t \rightarrow \infty} \mathbf{P}(t) = \mathbf{1}\boldsymbol{\pi}$$

Notice please that the last expression is a $d \times d$ matrix where each row is $\boldsymbol{\pi}$.

The Markov chain will be *time-reversible* if the probability of any sequence of events is the same in forward or backward time. This condition is equivalent to the expression

$$\pi_i q_{ij} = \pi_j q_{ji}, \quad \text{for all } i, j$$

which formally says that the rate of transitions from i to j equals the rate of transitions from j to i . A time-reversible matrix \mathbf{Q} can always be represented by a symmetric matrix \mathbf{R} with zeros on the main diagonal and the stationary distribution $\boldsymbol{\pi}$.

$$\mathbf{Q} = \mathbf{R}\boldsymbol{\Pi} - \text{diag}(\mathbf{R}\boldsymbol{\Pi})$$

where $\boldsymbol{\Pi}$ is the diagonal matrix with values of $\boldsymbol{\pi}$ along the main diagonal.

Mathematical Derivations

The following sections provide mathematical arguments to support results and interpretations in the previous summary. First, we derive results on waiting times and interpretations of parameters in the \mathbf{Q} matrix.

Next, we will show two separate arguments for the form of the Markov probability transition matrix being a matrix exponential. One argument uses the Markov property to show that

$$\mathbf{P}(s + t) = \mathbf{P}(s)\mathbf{P}(t)$$

for all $s, t \geq 0$ and argues that the matrix exponential is the only non-trivial solution by analogy to the real case. A second approach uses differential equations where the matrix exponential is the only solution. We go through both of these arguments but without complete mathematical rigor.

Waiting Times

Suppose we begin a Markov chain in state i . Let T_i be the time that the Markov chain remains in state i before the first transition elsewhere. Since we have a homogeneous Markov chain, it follows that $\mathbb{P}(T_i > t) = \mathbb{P}(T_i > s + t \mid T_i > s)$ for any $s > 0$. Using the definition of conditional probability, we find the following.

$$\begin{aligned} \mathbb{P}(T_i > s + t \mid T_i > s) &= \frac{\mathbb{P}(T_i > s + t \cap T_i > s)}{\mathbb{P}(T_i > s)} \\ &= \frac{\mathbb{P}(T_i > s + t)}{\mathbb{P}(T_i > s)} \end{aligned}$$

Thus,

$$\mathbb{P}(T_i > s + t) = \mathbb{P}(T_i > s) \mathbb{P}(T_i > t)$$

for all $s, t \geq 0$. There are very few functions f that satisfy $f(s+t) = f(s)f(t)$ for all s, t . Two trivial solutions are $f \equiv 0$ and $f \equiv 1$. The only other solutions are exponential functions. It follows that the dwell time distributions of a continuous-time Markov chain must be exponentially distributed. For each state i , we will let the rate of this distribution be q_i . Thus, the probability density of the dwell time in state i is $q_i e^{-q_i t}$, for $t > 0$ and the probability that the dwell time exceeds a time t is $e^{-q_i t}$.

Limit Results

By ignoring dwell-times, we can create a discrete-time Markov chain where there is a transition (or jump) at each discrete time point. Suppose that the transition probabilities of this jump chain are $\{a_{ij}\}$, so that $\sum_{j \neq i} a_{ij} = 1$ for each i . We will define $q_{ij} = q_i a_{ij}$ and then show that these q_{ij} can be interpreted as rates. Notice that it follows that

$$\sum_{j \neq i} q_{ij} = \sum_{j \neq i} q_i a_{ij} = q_i \sum_{j \neq i} a_{ij} = q_i$$

First, consider the value of $p_{ii}(t)$, the probability that a Markov chain beginning in state i is in state i at time t . The probability of this is the probability of no transitions in time t plus the probability of two or more transitions that result in a return to state i . We will represent the probability of all nonempty sequences of transitions leading from i back to i as $o(t)$, a notation that represents a generic function with the property

$$\lim_{t \downarrow 0} \frac{o(t)}{t} = 0$$

This means that the probability of two or more transitions returning to state i is small relative to t . The probability of no change is $e^{-q_i t}$, so

$$p_{ii}(t) = e^{-q_i t} + o(t)$$

We can simplify this expression by writing the exponential function as the first two terms of its Taylor series expansion around 0 plus an error term. We have

$$e^{-q_i t} = \sum_{k=0}^{\infty} \frac{(-q_i t)^k}{k!} = 1 - q_i t + o(t)$$

so that

$$p_{ii}(t) = 1 - q_i t + o(t)$$

The infinitesimal rate of change in time t is then

$$\begin{aligned} \lim_{t \downarrow 0} \frac{1 - p_{ii}(t)}{t} &= \lim_{t \downarrow 0} \frac{q_i t + o(t)}{t} \\ &= q_i \end{aligned}$$

so it is correct to think of q_i as the rate at which we leave state i when in state i . Similarly, we can represent $p_{ij}(t) = a_{ij} q_i t + o(t)$ for $j \neq i$ and

$$\begin{aligned} \lim_{t \downarrow 0} \frac{p_{ij}(t)}{t} &= \lim_{t \downarrow 0} \frac{a_{ij} q_i t + o(t)}{t} \\ &= a_{ij} q_i \\ &= q_{ij} \end{aligned}$$

so q_{ij} is the rate at which we enter state j directly from state i when in state i .

Markov Probability Transition Matrix Result

Let $\mathbf{P}(t) = \{p_{ij}(t)\}$ be the $d \times d$ Markov probability transition matrix. We will manipulate the probability distribution over a time $s + t$ and represent it in terms of transitions over times s and t to derive a property of these matrices.

An arbitrary element of the matrix $\mathbf{P}(s + t)$ is

$$p_{ij}(s + t) = \mathbb{P}(X(s + t) = j \mid X(0) = i)$$

where $s, t \geq 0$. We can rewrite this probability as a sum over all of the possibilities for the process at intermediate time s and then use properties of homogeneous Markov chains to derive the result.

$$\begin{aligned} p_{ij}(s + t) &= \mathbb{P}(X(s + t) = j \mid X(0) = i) \\ &= \sum_k \mathbb{P}(X(s + t) = j, X(s) = k \mid X(0) = i) \\ &= \sum_k \mathbb{P}(X(s) = k \mid X(0) = i) \mathbb{P}(X(s + t) = j \mid X(0) = i, X(s) = k) \\ &= \sum_k \mathbb{P}(X(s) = k \mid X(0) = i) \mathbb{P}(X(t) = j \mid X(0) = k) \\ &= \sum_k p_{ik}(s) p_{kj}(t) \end{aligned}$$

This says that each ij element of the matrix $\mathbf{P}(s + t)$ is the dot product of the i th row of $\mathbf{P}(s)$ and the j th column of $\mathbf{P}(t)$. Thus,

$$\mathbf{P}(s + t) = \mathbf{P}(s)\mathbf{P}(t) \quad \text{for all } s, t \geq 0.$$

We argued above that the only non-trivial real valued functions with this form were exponential functions. We assert by analogy that

$$\mathbf{P}(t) = e^{\mathbf{Q}t} = \sum_{k=0}^{\infty} \frac{\mathbf{Q}^k t^k}{k!}$$

is the form of continuous-time Markov probability transition matrix where $\mathbf{Q}^0 = \mathbf{I}$. To ensure that $\mathbf{P}(t)$ is a probability matrix, we need for its row sums to be one. This occurs if we assume that the rows of \mathbf{Q} sum to 0, in other words that $\mathbf{Q}\mathbf{1} = \mathbf{0}$. Notice the following.

$$\begin{aligned} \mathbf{P}(t)\mathbf{1} &= \left(\sum_{k=0}^{\infty} \frac{\mathbf{Q}^k t^k}{k!} \right) \mathbf{1} \\ &= \left(\mathbf{I} + \left(\sum_{k=1}^{\infty} \frac{\mathbf{Q}^{k-1} t^k}{k!} \right) \mathbf{Q} \right) \mathbf{1} \\ &= \mathbf{I}\mathbf{1} + \left(\sum_{k=1}^{\infty} \frac{\mathbf{Q}^{k-1} t^k}{k!} \right) \mathbf{Q}\mathbf{1} \\ &= \mathbf{1} + \mathbf{0} \\ &= \mathbf{1} \end{aligned}$$

To assert that the elements of $\mathbf{P}(t)$ are positive, we assume further that the off-diagonal elements of \mathbf{Q} must be nonnegative so that the diagonal elements of \mathbf{Q} are negative. This condition with the row-sum condition together imply that the \mathbf{Q} matrix is diagonally dominant (but not strictly). Results from some advanced texts on matrix analysis can be used to show non-negativity of $\mathbf{P}(t)$, but we do not show those results here.

Differential Equation Approach

The conventional mathematical approach to introduce continuous-time Markov chains is through differential equations. We show that approach here. The benefit is that the assertions we made earlier about the \mathbf{Q} matrix will be seen explicitly.

We begin by assuming the existence of rates q_{ij} for entering state j directly from state i and the rate $q_i = \sum_{j \neq i} q_{ij}$ which is the overall rate of leaving state i .

If we then want an expression for $p_i(t + dt)$, the probability that the process is in state i an infinitesimal time after time t , we need to consider that probability that the process is in state i at time t minus the probability that it leaves plus the probability that the process enters state i from some other state. As we will let dt become small, we need not consider multiple transitions. So,

$$p_i(t + dt) = p_i(t) - p_i(t)q_i dt + \sum_{j \neq i} p_j(t)q_{ji} dt$$

must hold for all states i . In matrix form, these d equations are represented succinctly as

$$\mathbf{p}(t + dt) = \mathbf{p}(t) + \mathbf{p}(t)\mathbf{Q}dt$$

where $\mathbf{Q} = \{q_{ij}\}$ and $q_{ii} = -q_i$. Some algebra and taking limits shows the following.

$$\mathbf{p}'(t) = \lim_{dt \downarrow 0} \frac{\mathbf{p}(t + dt) - \mathbf{p}(t)}{dt} = \mathbf{p}(t)\mathbf{Q}$$

If we replace $\mathbf{p}(t)$ by $\mathbf{p}(0)\mathbf{P}(t)$, we have the expression

$$\mathbf{p}(0)\mathbf{P}'(t) = \mathbf{p}(0)\mathbf{P}(t)\mathbf{Q}$$

The preceding expression must hold for all possible initial distributions $\mathbf{p}(0)$, in particular the elementary vectors with a single 1 and $d - 1$ 0s. It follows that

$$\mathbf{P}'(t) = \mathbf{P}(t)\mathbf{Q}$$

For real valued functions, the exponential functions are the solutions to this form of differential equation. We appeal to a similar result for matrix functions and find the following.

$$\mathbf{P}(t) = e^{\mathbf{Q}t}$$

Stationary Distributions and Limiting Distributions

A Markov chain is *irreducible* if it is possible to get from any state i to any state j . Notice that this need not imply that all q_{ij} are positive, but we typically do make this more stringent assumption for nucleotide substitutions. (Models for codons typically assume that changes that require two or more nucleotide substitutions have rate 0, but these models are irreducible nonetheless.)

Finite-state-space irreducible continuous-time Markov chains are *ergodic* and have a common limiting and stationary distribution, denoted $\boldsymbol{\pi}$. We say that $\boldsymbol{\pi}$ is the stationary distribution if $\mathbf{p}(t) = \boldsymbol{\pi}$ for all $t \geq 0$. We say that $\boldsymbol{\pi}$ is the limiting distribution if

$$\lim_{t \rightarrow \infty} \mathbf{P}(t) = \mathbf{1}\boldsymbol{\pi}$$

We assert that ergodic Markov chains have a unique stationary distribution, and that this distribution may be found by finding the right-eigenvector of \mathbf{Q} associated with eigenvalue 0, so that

$$\boldsymbol{\pi}\mathbf{Q} = \mathbf{0}^T$$

Notice that $\mathbf{1}$ is an eigenvector of \mathbf{Q} , also with eigenvalue 0, because

$$\mathbf{Q}\mathbf{1} = \mathbf{0} \times \mathbf{1} = \mathbf{0} .$$

A note on eigenvalues and linear algebra

A square matrix \mathbf{A} is said to have an *eigenvector* \mathbf{v} if $\mathbf{v} \neq \mathbf{0}$ and the product $\mathbf{A}\mathbf{v}$ is a multiple of \mathbf{v} , namely $\lambda\mathbf{v}$. The scalar λ is called the *eigenvalue* associated with the eigenvector. If \mathbf{v} is an eigenvector, then any non-zero multiple of \mathbf{v} will also be an eigenvector, so eigenvectors are only known up to a constant.

A square matrix \mathbf{A} is *invertible* if there exists a matrix \mathbf{A}^{-1} such that $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. A matrix is invertible if and only if it has no eigenvalues equal to 0.

The *spectral decomposition* of a matrix can be formed when the eigenvectors are *linearly independent*, meaning that no eigenvector can be written as a linear combination of the others. Let the matrix \mathbf{U} contain the (assumed) linearly independent eigenvectors of the matrix \mathbf{A} as its columns, let $\mathbf{V} = \mathbf{U}^{-1}$ have as its rows a set of right-eigenvectors, and let $\mathbf{\Lambda}$ be a diagonal matrix where the eigenvalues of \mathbf{A} are along the main diagonal. Assume as well that the i th column of \mathbf{U} and the i th row of \mathbf{V} correspond to the eigenvalue Λ_{ii} . Then

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}$$

This decomposition can be useful for computing matrix exponentials. Suppose that $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}$. Notice that $\mathbf{Q}^k = \mathbf{U}\mathbf{\Lambda}^k\mathbf{V}$ for $k \geq 0$ and that raising diagonal matrices to powers is simple.

We can assume if we wish that the first column of \mathbf{U} is $\mathbf{1}$, the first element of $\mathbf{\Lambda}$ is 0 and the first row of \mathbf{V} is $\boldsymbol{\pi}$. Let $\boldsymbol{\lambda}$ be the eigenvalues of \mathbf{Q} written as a vector so that $\mathbf{\Lambda} = \text{diag}(\boldsymbol{\lambda})$. Then,

$$\begin{aligned} \mathbf{P}(t) &= e^{\mathbf{Q}t} \\ &= \sum_{k=0}^{\infty} \frac{\mathbf{Q}^k t^k}{k!} \\ &= \mathbf{U} \left(\sum_{k=0}^{\infty} \frac{\mathbf{\Lambda}^k t^k}{k!} \right) \mathbf{V} \\ &= \mathbf{U} \text{diag}(e^{\boldsymbol{\lambda}t}) \mathbf{V} \end{aligned}$$

where $e^{\boldsymbol{\lambda}t}$ is the vector whose i th element is the exponential of the i th element of $\boldsymbol{\lambda}$.

If we wish to compute matrix exponentials of \mathbf{Q} for many different t (as we would in a maximum likelihood optimization), it can be most efficient to work out the spectral decomposition of \mathbf{Q} . If we need only a single calculation, there are better computational methods rather than the full spectral decomposition.

Time-Reversibility

A Markov chain is *time-reversible* if and only if the probability of any sequence of events (jumps at specific times) is the same for time moving forward as it is for it moving backward. It turns out that Markov chains are time-reversible if and only if they satisfy *detailed balance*, which says that the stationary rate of transitions from i to j must equal the stationary rate of transitions from j to i for every pair of states i and j . In equation form, this is

$$\pi_i q_{ij} = \pi_j q_{ji} \quad \text{for all states } i \text{ and } j .$$

We can show that for a matrix \mathbf{Q} to be time-reversible, we must be able to write it as

$$\mathbf{Q} = \mathbf{R}\mathbf{\Pi} - \text{diag}(\mathbf{R}\mathbf{\Pi})$$

where $\mathbf{\Pi}$ is the diagonal matrix with values of $\boldsymbol{\pi}$ along the main diagonal and \mathbf{R} is a symmetric matrix with zeros on the diagonal. Notice that the first term on the right-hand-side of the equation is a matrix with zeros on the main diagonal and that the second term merely puts the negative row

sums in the proper place. So, we need only check that our time-reversibility condition is met by the matrix $\mathbf{R}\mathbf{\Pi}$.

If $q_{ij} = r_{ij}\pi_j$ for $i \neq j$, we need to check that

$$\pi_i(r_{ij}\pi_j) = \pi_j(r_{ji}\pi_i)$$

for all i and j which is obviously true if and only if r is symmetric.

This parameterization is useful for seeing how many free parameters there are in the most general time-reversible model. For $d = 4$, there are six free choices in the matrix \mathbf{R} and three free choices for $\boldsymbol{\pi}$ because the four values are constrained to sum to one, making nine in all. In general, there are $(d - 1)^2$ free parameters in the most general time-reversible model (at stationarity).

Notational Differences to Galtier

I will hand out a book chapter by Nicolas Galtier on likelihood models that covers much of the same material as these notes. I attempt to use the most standard notation that is commonly used in the phylogenetics literature. Galtier is far more complete than these notes. His chapter includes many topics that I do not, such as calculating likelihoods on trees. His chapter contains citations and far more verbiage to motivate the reader, as well as examples of specific models.

This section of the notes is simply meant to highlight the differences in our notation to aid your reading of both.

Galtier uses	I use
$\mathbf{F}(t)$	$\mathbf{p}(t)$
M^T	\mathbf{Q}
\mathcal{E}	Ω
$\mathbf{\Pi}$	$\boldsymbol{\pi}$
μ_{xy}	q_{ij}
s_{xy}	r_{ij}
$e^{Mt}\mathbf{F}(0)$	$\mathbf{p}(0)e^{\mathbf{Q}t}$
QDQ^{-1}	$\mathbf{U}\boldsymbol{\Lambda}\mathbf{V}$
Pr	$\mathbb{P}()$