Universität Wien

Max F. Perutz Laboratories

Master's Project

MATHEMATICAL TOPICS IN PHYLOGENETICS

Cassius Manuel Pérez de los Cobos Hermosa

Supervisor: Prof. Arndt Von Haeseler

I

¡Tiempo perdido, di cuánto durabas
mientras iba cantando tu sueño!

Ismena, 2º Acto, Coro de niños.

Contents

Structure of the work

Our main goals along this project were two: first of all, understand and explain, from a mathematically rigorous point of view, the modeling of an evolutionary process as a continuous Markov process, and why Lie Markov models are an adequate option for this task. Secondly, we wanted to improve a previous implementation of Lie Markov models in IQ-TREE, an algorithm for inferring phylogenies.

The distribution of the chapters follows this aim:

- Chapter 1 contains a description of every object necessary to deal with continuous-time Markov models. The reference we used is [1], which we recommend for students who start learning the matter.
- Chapter 2 is nothing but a detailed explanation of some parts of the article [4]. Since we did not have the space limits and the formal requirements which a paper normally has, we could give abundant explanations when we considered convenient, and omit technical results which are not necessary for the understanding of the theory.
- Chapter 3 includes a plot of the article [5], plus our results regarding the implementation of Lie Markov models to IQ-TREE. For a description of this software, see [9].
- Chapter 4 is composed by mathematical results which can be used in much broader contexts than ours, and whose description could distract the attention of the reader when following the rest of the chapters. Apart from the mentioned references, the main source of results is the excellently written book "The symmetric group", [8].

Apart from these references, we used notes from the course Mathematical Models of Biology, given by Marta Casanellas and Jesús Fernández Sánchez.

Notation

There is only one notation convention which we follow and must be carefully taken into account by the reader. Our transition matrices do not follow the stochastic convention of having rows which sum 1, but have columns which sum 1. We do this in order to follow the articles [4] and [5], for otherwise the reader of both would be confused. Moreover, it is justified to follow this convention, since it allows us to write most of biologically meaningful vectors as rows, which is more comfortable for mathematicians. In any case, we ask to the reader not to forget about this fact.

Chapter 1

First steps

1.1 The discrete model

Given a set of m species with a common ancestor, the evolutionary process suffered by the original gene is modeled by a tree T which is rooted and has m labeled leaves. For example, if $m = 3$ and the root is called π, we could have:

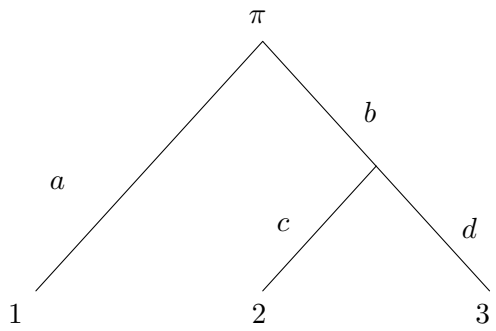

Figure 1.1: Tree of three leaves.

We assume than only substitution of bases can occur (i.e. no deletion nor addition) and also that substitutions are inter-independent, so we can focus on a single position of the gene. Therefore it is enough to set each node of the tree T to be a random variable with n possible states (we will take $n = 4$ for the nucleotides $\{A, C, G, T\}$). At the root, π at Figure 1.1, the distribution of states is given by (we abuse of he notation) $\pi^T = (\pi_1, \cdots, \pi_n)$.
We will assume that on each edge e of the tree there is a $n \times n$ transition matrix M_e whose entries are indeterminates representing the probabilities of

transition between the states. Since we follow the column-sum-1 convention, in a matrix M^e the element $m^e_{i,j}$ would represent the probability of the father node with state j to become the son node with state i. In Figure 1.1, $e \in \{a, b, c, d\}$. It is important to note that in discrete models **the edge lengths are not meaningful**. The random variables at the leaves are *observed*, while the ones at the interior nodes are *hidden*.

The entries of the matrices M_e and the vector π are the *model parameters*, on which normally many restrictions are imposed, for example forcing some elements to be identical. Moreover, in order to be biologically meaningful, this objects must satisfy some properties, which we define:

Definition 1.1. *A matrix $M = (m_{ij}) \in M_n(\mathbb{R})$ is called a* Markov matrix *if its elements are nonnegative and, for every column of M, the sum of its elements is 1. Differently explained, if for all $i, j \in [n]$ we have $m_{ij} \geq 0$ and*

$$(1, \cdots, 1)M = (1, \cdots, 1)$$

.

Definition 1.2. *A column vector $\pi^T = (\pi_1, \cdots, \pi_n)$ is said to be a* distribution vector *if its elements are nonnegative and their sum is 1. In other words, if for all $i \in [n]$ we have $\pi_i \geq 0$ and*

$$(1, \cdots, 1)\pi = 1.$$

With these definitions, we can say that the properties which the objects π and M^e must satisfy are the following:

- Every M^e must be a Markov matrix.
- The root π must be a distribution vector.

Markov matrices satisfy the following property, which we will repeatedly use:

Proposition 1.3. *Let as $M_1, M_2 \in M_n(\mathbb{R})$ be Markov matrices. Then M_1M_2 is a Markov matrix.*

Proof. It is obvious that matrix M_1M_2 will be nonnegative. Moreover, we easily see that

$$(1, \cdots, 1)M_1M_2 = (1, \cdots, 1)M_2 = (1, \cdots, 1),$$

which finishes the proof.

□

The multiplicative closeness of Markov matrices is important because it allows us to set one matrix which is equivalent to the action of two matrices (and so on). For example, let us look at the tree in Figure 1.2. The random distribution at node 1 will be

$$M_a M_b \pi = (M_a M_b)\pi,$$

hence this process is equivalent to the process in Figure 1.3, in which $M_c = M_a M_b$. This will turn out to be a very relevant discussion along this work.

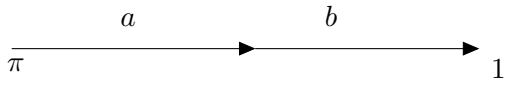

Figure 1.2: Evolution process in two steps.

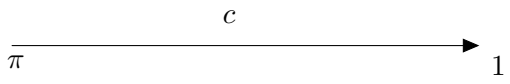

Figure 1.3: Evolution process in one step.

However, discrete modeling has obvious disadvantages, among which the most important one is that it ignores, or at least considers as known, the time which has passed between one node and another. In other words, the variable time (i.e. edge length) is not considered as an unknown we should determine. This posses the necessity of using more complex models for biological research, which we introduce in next section.

1.2 From discrete to continuous-time Markov models

Continuous-time Markov models are the usual approach preferred by biologists for inferring phylogenies. First of all we will justify the election of the modeling and then expose its substitution matrices and their mathematical properties. The interested reader can find more information in [1]. For the sake of clarity we simply state the case where only four states are involved

(one for each of the bases A, C, G, T), although every result we will state applies also for the general case with nearly identical proofs.

Our assumptions when inferring the model will be the following:

- All sites in a nucleotide sequence evolve independently and following the same stochastic process.
- Evolution follows a Markov process (any future state is independent of the past, given the present state).
- The rate of change (which we will soon explain) does not change over time. This property is called *homogeneity*.

When saying we aim to model evolution as a continuous Markov process, we refer to expressing the transition matrix as some smooth, time dependent function. Therefore our first definition will be the following:

Definition 1.4. *The Markov-matrix function $M(t)$ is a* (smooth) transition matrix *if it has the form*

$$M(t) := \begin{pmatrix} p_{A,A}(t) & p_{C,A}(t) & p_{G,A}(t) & p_{T,A}(t) \\ p_{A,C}(t) & p_{C,C}(t) & p_{G,C}(t) & p_{T,C}(t) \\ p_{A,G}(t) & p_{C,G}(t) & p_{G,G}(t) & p_{T,G}(t) \\ p_{A,T}(t) & p_{C,T}(t) & p_{G,T}(t) & p_{T,T}(t) \end{pmatrix},$$

where the smooth function $p_{X,Y}(t) := P(Y|X,t)$ is the probability that nucleotide X changes to nucleotide Y after some time $t \in \mathbb{R}_{\geq 0}$.

The degree of smoothness is not specifically determined, but we want every function to be at least $\mathcal{C}_1$. As we will see, this will naturally be the case. Another point which should be taken into account is that in this definition we are following the convention of making the rows sum 1. However, biologists may prefer the rows-sum-1 convention (i.e. to use the transpose of our $M(t)$). Any of these two methods can be chosen, but one must be careful so no mistake is made.

Our substitution matrix has a disadvantage: attending to its definition, it has no invariant, i.e. a *quantity* which remains unchanged while t increases. It is convenient to restrict our definition in order to create one, which makes computations much simpler. However, we will proceed backwards: we will assume its existence and arrive to the proper definition of our new $M(t)$.

For every two different nucleotides X and Y, we will write $q_{X,Y}$ to refer to the *instantaneous rate of substitutions* of X by Y. In other words, $q_{X,Y}$ is

the speed of X being replaced by Y when time goes to zero (i.e. there is some derivative involved, although the differential equation will be deduced later). We define $q_{X,X} : -\sum_{X \neq Y} q_{X,Y}$. Therefore these $q_{X,Y}$ are our invariants; we are assuming the instantaneous rates of change are constant.

All these rates of change can be condensed in one sole matrix Q. We give a definition of this matrix Q *ab ovo*, i.e. without the need of a previous definition of the rates $q_{X,Y}$:

Definition 1.5. *A* rate-matrix *is a matrix* $Q \in M_4(\mathbb{R})$ *with the form*

$$Q := \begin{pmatrix} q_{A,A} & q_{C,A} & q_{G,A} & q_{T,A} \\ q_{A,C} & q_{C,C} & q_{G,C} & q_{T,C} \\ q_{A,G} & q_{C,G} & q_{G,G} & q_{T,G} \\ q_{A,T} & q_{C,T} & q_{G,T} & q_{T,T} \end{pmatrix},$$

which satisfies the following conditions:

- $q_{X,Y} \geq 0$ *for* $X \neq Y$.
- $q_{X,X} < 0$ *for any* X.
- $\sum_Y q_{X,Y} = 0$, *i.e.* $(1,1,1,1)Q = 0$.

We aim to justify the differential equation we will use to define $M(t)$ depending on Q. If we return now to our interpretation of $q_{X,Y}$ as the instantaneous rate of substitution which is included in our matrix Q, we can infer the following equation:

$$p_{X,Y}(t + \Delta t) = p_{X,Y}(t) - \Big(\sum_{Z \neq Y} q_{Y,Z}\ \Delta t \Big) p_{X,Y}(t) + \sum_{Z \neq Y} q_{Z,Y}\ \Delta t\ p_{X,Z}(t).$$

It can be easily understood helping ourselves with the following figure:

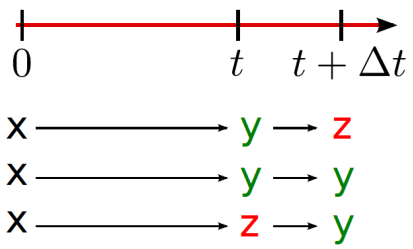

Figure 1.4: Rates of change after t and Δt has passed.

Figure 1.4 represents the following: we consider states X and Y, and an undetermined state $Z \neq Y$. Along a mutation process suffered by X starting

at time $t_0 = 0$ until time t, we know that $p_{X,Y}$ measures the probability of X becoming Y. Provided this was the case, after Δt, either this Y mutated to Z or it remained unchanged. Moreover, at time t there were some X mutated to Z which after time Δt became Y. At time $t + \Delta t$, all possible cases are exhausted in the three scenarios of the figure, each of which has its analogue term in the equation.
This equation can give us our differential equation in terms of $M(t)$ and Q. To that end, first of all we write it with its matrix form, which is the following:

$$M(t + \Delta t) = M(t) + \Delta t Q M(t),$$

which can be rearranged as

$$\frac{M(t + \Delta t) - M(t)}{\Delta t} = QM(t),$$

and after making Δt tend to zero, the differential equation comes up:

$$M'(t) = QM(t). \tag{1.1}$$

We can invoke corollary 4.3 in order to determine the unique solution of this equation. We set $M(0) = \mathbf{1}$, since at time 0 no mutation has occurred, and the aforementioned corollary gives that, for $t \geq 0$,

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

For a brief review of the definition of the exponential of a matrix, *vid.* chapter 4. For some of its most relevant properties, *vid.* proposition 4.2.
Since all this discussion was simply a justification of this equation, it is pertinent, for the sake of formality, to give a new definition of this transition matrix:

Definition 1.6. *Given a continuous-time substitution process associated to a rate-matrix Q, the time-dependent transition matrix $M(t)$ associated to this process is*

$$M(t) := e^{Qt},$$

where $t \geq 0$. Since Q does not depend on t, we will say this is a homogeneous model.

Therefore when we aim to study a stationary model, there are two main ingredients: the matrices Q and the respective duration t of the processes. Roughly speaking, inferring phylogenies can be based on these two objects, as we will explain later in more detail.
Before going forward, we must prove that definition 1.6 satisfies our expectations, i.e. that this $M(t)$ is a Markov matrix (with columns sums equal 1 and nonnegative elements). Actually we have a stronger result, which establishes the connexion between rate matrices and transition matrices, whose proof will use the following lemma:

Lemma 1.7. *Given the column vector v, and any rate matrix Q with associated transition matrix $M(t)$, then for any $t \geq 0$ we have:*

- *If $v^T Q = 0$, then $v^T M(t) = v$.*
- *If $Qv = 0$, then $M(t)v = v$.*

Proof. One simply has to use the definition of exponential. We have

$$v^T M(\frac{t}{m}) = v^T \mathbf{1} + \sum_{k=1}^{\infty} v^T \frac{Q^k}{k!} (\frac{t}{m})^k = v^T$$

where we used $v^T Q^k = v^T Q Q^{k-1} = \vec{0}^T Q^{k-1} = \vec{0}^T$.
The second part of the proposition is proved analogously.

□

Proposition 1.8. *Q is a rate matrix if and only if $M(t) := e^{Qt}$ is a Markov matrix for all $t \geq 0$.*

Proof. ($\Leftarrow$) If $M(t)$ is a Markov matrix for all $t \geq 0$, then for $t \in \mathbb{R}_{\geq 0}$ and any $j \in [4]$ we have

$$\sum_i M_{ij}(t) = 1.$$

We can differentiate this equation with respect to t and get

$$\sum_i M'_{ij}(t) = 0.$$

Moreover, from 1.1, and using that $M(0) = \mathbf{1}$, we infer that

$$M'(0) = QM(0) = Q.$$

Combining these two equations, we conclude that

$$\sum_i Q_{ij}(t) = 0.$$

Moreover, using again $M'(0) = Q$, we can guarantee that $Q_{ij} \geq 0$ provided $i \neq j$. Indeed, since $M(0) = \mathbf{1}$ and $M(t)$ must be a Markov matrix for every t, it cannot happen that for $i \neq j$ we had $M'_{ij}(0) < 0$, because if this was the case, there would be a neighborhood of $t = 0$, let us say $(0, \epsilon)$, in which $M'_{ij}(t) < 0$, hence after integrating we would get

$$M_{ij}(\epsilon) = 0 + \int_0^{\epsilon} M'_{ij}(t) < 0,$$

which contradicts our initial hypothesis of $M(t)$ being a Markov matrix for every t.
All in all, we conclude Q is a rate matrix, as desired.

($\Rightarrow$) Suppose we are given a rate matrix Q and $t \in \mathbb{R}_{\geq 0}$. First of all, we must note that, for any $m \in \mathbb{N}$,

$$\Big(M(\frac{t}{m})\Big)^m = (e^{\frac{t}{m}Q})^m = e^{m\cdot\frac{t}{m}Q} = e^{Qt}.$$

The second equality is not trivial, for exponentiating a matrix does not work as doing so with a complex number. In our case, however, this equality follows from the BCH formula, stated in 4.1, and the commutativity of any matrix with itself (hence $[Q, Q] = 0$).

Now let us suppose we are given a real number $t \in \mathbb{R}_{\geq 0}$. We aim to prove that e^{Qt} is a Markov matrix.
If $t = 0$, the result follow immediately, so let us assume $t > 0$. We claim that there exists an $m \in \mathbb{N}$ big enough such that $M(\frac{t}{m})$ is a Markov matrix. Indeed, the definition of exponential of a matrix yields

$$M(\frac{t}{m}) = \mathbf{1} + \frac{t}{m}Q + \frac{t^2}{m^2}\mathcal{O}(1),$$

where the $\mathcal{O}(1)$ term simply indicates that the rest of the series converges to some matrix, whose matrix norm (we mention this just in case someone is skeptical about this fact) is actually bounded by

$$\delta = ||Q||^2 e^{||Q||t},$$

where $||\cdot||$ is the norm induced by, let us say, the square norm.
We take m big enough, for example

$$m > 2\max\{1,\ |t|^2\delta,\ |t|\cdot||Q||,\ \frac{t\delta}{\min_{i\neq j}|q_{ij}|}\},$$

which forces any diagonal term in $\frac{t}{m}Q$ and $\frac{t^2}{m^2}\mathcal{O}(1)$ to be smaller than $\frac{1}{2}$, hence we get that $M(\frac{t}{m})$ has positive diagonal. Regarding the non diagonal terms of $M(\frac{t}{m})$, it is enough to prove the non negativity of the non diagonal terms of the matrix

$$Q + \frac{t}{m}\mathcal{O}(1),$$

which is actually the case attending to the definition of m.
Putting all this together, we have proved that matrix $M(t/m)$ is nonnegative. Moreover, it is Markov since, by lemma 1.7,

$$(1,1,1,1)M(t) = (1,1,1,1).$$

Finally, attending to proposition 1.3, the product of Markov matrices is Markov, hence $M(t) = M(\frac{t}{m})^m$ must be Markov too, as we wanted to prove.

□

Last proposition could make us think that every transition matrix we can come up with must be the exponential of a rate matrix. On the contrary, we can easily find a counterexample using item 6 in proposition 4.2, which states

$$\det(e^{Qt}) = e^{\mathrm{Tr}(Q)}.$$

Therefore the exponential of a rate matrix has positive discriminant. It is enough to find a Markov matrix with negative discriminant, and this will be out counterexample. Let us say, for example:

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},$$

whose determinant is -1.

1.3 Equilibrium base frequency

Given any time-dependent transition matrix $M(t)$, an interesting feature is the existence of *equilibrium base frequencies.*

Definition 1.9. *The column distribution vector* $\pi^T = \{\pi_A, \pi_C, \pi_G, \pi_T\}$ *is an* equilibrium base frequency *of the transition matrix* $M(t)$ *if, for any* $t \in \mathbb{R}$*,*

$$M(t)\pi = \pi,$$

i.e. π *is a fixed point of* $M(t)$*.*

Studying the fixed points of the smooth function $M(t)$, without any other restriction, is a difficult task which cannot be solved with all generality. However, a time-dependent homogeneous model, determined by a rate matrix Q, is easier to deal with, as the reader may check in a moment.

We know that the rate matrix Q, by definition, satisfies the equality

$$(1, 1, 1, 1)Q = \vec{0},$$

which implies these two straightforward statements:

- The rank of Q is less or equal 3.
- The rate matrix Q has the left eigenvector $(1,1,1,1)$ with associated eigenvalue 0.

We would like to have an appropriate equilibrium base frequency for every Q. Before attacking this problem, we should recall a version of the Perron-Frobenious theorem, as included in [2]:

Theorem 1.10 (Perron-Frobenious theorem)**.** *Let $A = (a_{ij})$ be a $n \times n$ positive matrix, i.e. such that $a_{ij} > 0$ for $i, j \in [n]$. Then there exists a positive real number $r \in \mathbb{R}^+$ such that r is an eigenvalue of A and any other eigenvalue $\lambda \in Spec(A)$ satisfies $|\lambda| < r$. This eigenvalue is simple, hence its associated eigenspace is one-dimensional.*
Moreover, there exists an eigenvector v with eigenvalue r such that all its components are positive, and there are no other positive (or even nonnegative) eigenvectors with the exception of positive multiples of v.

Proposition 1.11. *A homogeneous model associated to a rate matrix Q has a unique equilibrium base frequency π.*

Proof. We consider firstly the strict case when $q_{ij} > 0$ for any $i \neq j$. Let us consider the matrix $P = \alpha\mathbf{1} + Q$, where $\alpha \in \mathbb{R}^+$ is chosen such that $\alpha + q_{ii} > 0$, i.e. matrix P is positive, and so it is P^T. Therefore we can apply the Perron-Frobenious theorem to P, although we also apply this theorem to P^T: this matrix has a unique positive eigenvector, which corresponds to the maximum eigenvalue $r \in \mathbb{R}^+$. However, we have that

$$P^T(1,1,1,1)^T = (\alpha\mathbf{1} + Q^T)(1,1,1,1)^T = \alpha(1,1,1,1),$$

hence this eigenvector must be $(1,1,1,1)^T$ and, more importantly,

$$r = \alpha.$$

Since P and P^T have the same eigenvalues, the Perron-Frobenious theorem implies that there exists a unique positive eigenvector of P. We normalize this eigenvector, so its elements sum 1, and we will call it π. The eigenvector π has eigenvalue α, which implies

$$Q\pi = (P - \alpha\mathbf{1})\pi = 0,$$

hence we are done for this case.

Now we have to prove the non strict case in which $q_{ij} \geq 0$. Let us consider the sequence of matrices $(Q_m)_{m\in\mathbb{N}}$ whose elements are defined as $q^m_{ij} := q_{ij} + \frac{1}{m}$ for $i \neq j$, and $q^m_{ii} = q_{ii} - \frac{n-1}{m}$.

Any of these Q_m belongs to the case we proved below, hence for each of them there exists a positive eigenvector π_m such that $|\pi_m| = 1$ and $Q_m\pi_m = 0$. Using that the unit sphere is a compact set, there must exist a converging sequence

$$\pi_{m_k} \to \pi,$$

where $m_k \to \infty$ as $k \to \infty$, hence $Q_{m_k} \to Q$. All in all,

$$Q\pi = \lim_{k\to\infty} Q_{m_k}\pi_{m_k} \equiv 0.$$

Moreover, this vector π is the limit of a sequence of positive vectors π_{m_k}, hence it has nonnegative elements. This π is the equilibrium base frequency we aimed to find.

□

Chapter 2

Lie Markov models

2.1 Definitions and basic properties

Most of likelihood methods for phylogenetic inference try to fit a single rate-matrix globally across a proposed evolutionary tree history. These rate-matrices are chosen from a *model*, i.e. a set of matrices defined by a certain set of constraints on the elements of a generic rate-matrix. All these objects will be formally defined later. Along this chapter we will need to redefine some of the concepts we introduced before. The justification for this is, first of all, that we aim to follow the papers closely so it is easier to compare their content with our explanations; secondly, that there are good reasons to follow these conventions, being the principal one the mathematical simplicity they bring to our statements.

A *homogeneous* Markov chain is a sequential evolution process in which probability transition rates are constant in time. This is used as an approximation to biological reality, for it is known that this may not be the case. Therefore, as it happens with every model in scientific research, we are simplifying reality for the sake of computability. However, methods have been developed to deal with this fact, among which *Lie Markov models* count themselves as a very interesting one.

From now on, unless the contrary is stated, we will work over the complex field $\mathbb{C}$, since this generalization makes it easier to use well-known mathematical tools. First of all we need some definitions:

Notation 2.1. *We will write θ to refer to the column n-vector with all entries equal to* 1*, i.e. $\theta^T = (1, \cdots, 1)$.*

Definition 2.2. *A matrix $M \in M_n(\mathbb{C})$ is called a* Markov matrix *if, for*

every column of M, the sum of its elements is 1. *Therefore* $\theta^T M = \theta^T$.

Definition 2.3. *The* general Markov model, $\mathcal{M}_{GMM}$, *is the set of all* $n \times n$ *Markov matrices. Differently explained:*

$$\mathcal{M}_{GMM} := \{M \in M_n(\mathbb{C}) : \theta^T M = \theta^T\}.$$

Definition 2.4. *The subset of matrices in* $\mathcal{M}_{GMM}$ *with non-zero determinant is denoted as* $GL_1(n, \mathbb{C})$. *Therefore*

$$GL_1(n, \mathbb{C}) := \{M \in M_n(\mathbb{C}) : \theta^T M = \theta^T \ , \ \det(M) \neq 0\}.$$

Since $M, N \in GL_1(n.\mathbb{C})$ satisfy

$$\theta^T = \theta^T M^{-1};\ \theta^T MN = \theta^T,$$

it is infered that $GL_1(n, \mathbb{C})$ is a subgroup of the general linear group $GL(n, \mathbb{C})$. We are interested in avoiding the degenerated cases which can arise when considering $\mathcal{M}_{GMM}$, hence along this work we will not leave the group $GL_1(n, \mathbb{C})$. This is not a great loss since the set of Markov matrices with zero discriminant is of measure zero in $\mathcal{M}_{GMM}$.

Regarding continuous-time processes, we have the following definitions:

Definition 2.5. *A matrix* $Q \in M_n(\mathbb{C})$ *is called a* rate matrix *if, for every column of* Q, *the sum of its elements is* 0. *Therefore* $\theta^T Q = \mathbf{0}^T$.

Definition 2.6. *The set* $\mathcal{L}_{GMM}$ *is compounded by all* $n \times n$ *rate matrices. Equivalently:*

$$\mathcal{L}_{GMM} := \{Q \in M_n(\mathbb{C}) : \theta^T Q = \mathbf{0}^T\}.$$

Definition 2.7. *The* general rate-matrix model, $e^{\mathcal{L}_{GMM}}$, *is the set of exponentials of all rate matrices. Therefore:*

$$e^{\mathcal{L}_{GMM}} := \{e^Q : Q \in \mathcal{L}_{GMM}\}.$$

As we proved in proposition 1.8, given $Q \in \mathcal{L}_{GMM}$, e^{Qt} is a Markov matrix for every $t \in \mathbb{R}$. (However, one should note that we have redefined the terms involved, although with these new and weaker definitions the statement continues holding; a formal proof can be made by taking the appropriate parts of the proof of proposition 1.8). Moreover, the exponential of a matrix is invertible, attending to item 7 in proposition 4.2. One more property would be interesting for us, its closeness under multiplication. Fortunately, we have the Baker-Campbell-Hausdorff (BCH) formula, stated in equation 4.1:

$$e^X e^Y = \exp\{X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}[X, [X, Y]] + \cdots\},$$

where $[X, Y] := XY - YX$ is the commutator of X and Y. Provided it converges, this series must be an element of $\mathcal{L}_{GMM}$, since $\mathcal{L}_{GMM}$ is a (topologically) closed set which is closed under sum and multiplication, as one can easily check. Therefore, we conclude $e^{\mathcal{L}_{GMM}}$ is a subgroup of $GL_1(n.\mathbb{C})$.

Summarizing, so far we have the following group hierarchy:

$$e^{\mathcal{L}_{GMM}} < GL_1(n, \mathbb{C}) < GL(n, \mathbb{C}).$$

However, one is normally interested in not considering such vast sets, hence more tractable ones must be defined.

Definition 2.8. *A* Markov model $\mathcal{M}$ *is a well defined subset* $\mathcal{M} \subset \mathcal{M}_{GMM}$.

Definition 2.9. *A* rate-matrix model $e^{\mathcal{L}}$ *is the set of exponentials of a well defined subset* $\mathcal{L} \subset \mathcal{L}_{GMM}$. *Differently explained,* $e^{\mathcal{L}} := \{e^Q \mid Q \in \mathcal{L}\}$.

Again by proposition 1.8, it is easy to see that all rate-matrix models are Markov models. Our interest yields on the case $\mathcal{M} = e^{\mathcal{L}}$, hence we will commit an abuse of notation and refer to $\mathcal{L}$ as a *model*.

2.2 Multiplicative closeness

When introducing the sets $GL_1(n, \mathbb{C})$ and $e^{\mathcal{L}_{\mathcal{GMM}}}$, we were interested in proving they are actually groups. As we will see along this section, this property, or more precisely a weaker demand as the following one, is crucial:

Definition 2.10. *A Markov model* $\mathcal{M}$ *is* multiplicatively closed *iff*

$$M_1, M_2 \in \mathcal{M} \Rightarrow M_1 M_2 \in \mathcal{M},$$

i.e. if $\mathcal{M}$ *forms a semigroup under matrix multiplication.*
When $e^{\mathcal{L}}$ *is multiplicatively closed, we will also say that* $\mathcal{L}$ *is multiplicatively closed.*

First of all, we would like to find sufficient conditions for a rate-matrix model $\mathcal{L}$ to be a multiplicatively closed one. This problem was partially commented when introducing the BCH formula in 4.1, which we write again:

$$e^X e^Y = \exp\{X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}[X, [X, Y]] + \cdots\}.$$

Therefore we are looking for conditions on every $X, Y \in \mathcal{L}$ to force

$$X + Y + \frac{1}{2}[X, Y] + \frac{1}{12}[X, [X, Y]] + \cdots \in \mathcal{L}.$$

Let us recall definition 4.14: a closed set $\mathcal{L}$, with the required operators defined, is a Lie algebra if $t_1X + t_2Y \in \mathcal{L}$ and $[X, Y] \in \mathcal{L}$, for all $X, Y \in \mathcal{L}$ and $t_1, t_2 \in \mathbb{C}$. We have the following result:

Proposition 2.11. *If L is a Lie algebra, then $\mathcal{L}$ is multiplicatively closed.*

Proof. From the property $t_1X + t_2Y \in \mathcal{L}$ we infer $\mathcal{L}$ must be an affine space, hence a closed set. Moreover, since $[X, Y] \in \mathcal{L}$, we deduce any of the Lie brackets of the sum belong to $\mathcal{L}$, hence also each of the summands. All in all, the series is a convergent sum of elements in $\mathcal{L}$ (a closed set), hence it converges to an element $Z \in \mathcal{L}$, as we aimed to prove.

□

Therefore the condition we were seeking for $\mathcal{L}$ was "being a Lie algebra". At this point it is pertinent to expose why being multiplicatively closed is a significant property in biological terms. To that end, let us consider a phylogenetic tree in which each of the edges e has an associated rate-matrix Q_e from some model $\mathcal{L}$.
Now let us look at the following figure:

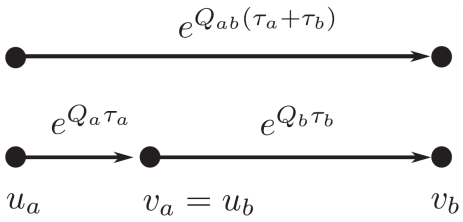

Figure 2.1: Decomposition of the orbits of $\mathcal{G}_4$ into irreducible modules.

We take an edge $e_a = (u_a, v_a)$, whose associated rate-matrix is Q_{e_a} and hence its associated Markov matrix is $e^{Q_a\tau_a}$, where τ_a is the length of the edge e_a. We also consider the edge $e_b = (u_b, v_b)$ with length τ_b and $u_b = v_a$, i.e. it leaves from the tip of e_a. We aim to remove the taxon $v_a = u_b$, and find an equivalent edge $e_{ab} = (u_a, v_b)$ with length $\tau_a + \tau_b$ and an adequate Q_{ab} which satisfies

$$\exp[Q_{ab}(\tau_a + \tau_b)] = \exp[Q_a\tau_a] \exp[Q_b\tau_b].$$

In other words, we have applied a marginalization procedure with respect to the vertex $u_a = v_b$. We have already proved that such a matrix Q_{ab} exists in $\mathcal{L}_{GMM}$. However, we need to ask for more, since we are using a model $\mathcal{L}$ which we want to respect, i.e. we need $Q_{ab} \in \mathcal{L}$. If we write $X := Q_a\tau_a$ and $Y := Q_bt_b$, we can clearly see that, to this end, $\mathcal{L}$ must be multiplicatively closed.

This virtue which multiplicatively closed models have can be taken further. Let us suppose the model $\mathcal{L}$ forms a Lie algebra and consider the rate-matrices $Q_1, \cdots, Q_m \in \mathcal{L}$ and the parameters $\tau_1, \cdots \tau_m \in \mathbb{C}$, as in the following figure:

Figure 2.2: Inhomogeneous process.

This inhomogeneous evolutionary process has a substitution matrix given by

$$e^{\tau_m Q_m} \cdots e^{\tau_1 Q_1} := M(t),$$

where $t = \tau_1 + \cdots + \tau_m$. We can repeatedly apply the multiplicatively closeness of $\mathcal{L}$ and conclude that we can write

$$M(t) = e^{\hat{Q}t},$$

for some matrix $\hat{Q} \in \mathcal{L}$. Hence we have found a matrix $\hat{Q} \in \mathcal{L}$ which acts as a homogeneous average of the inhomogeneous process given, which makes evident the virtues of these models.

2.3 The GTR model is not multiplicatively closed

Among all possible rate-matrix models, the general time reversible model (GTR) counts itself as the most frequently selected for phylogenetic inference. In this section we will show that it actually lacks of multiplicatively closeness, hence alternative ones should be used if one wants to be consistent with the priorities exposed below.

First of all let us introduce the pertinent definitions and the model.

Notation 2.12. *We will write π to refer to any column n-vector with the form*

$$\pi^T = (\pi_1, \cdots \pi_n),\ \pi_i \in \mathbb{R}^+,\ \pi_1 + \cdots \pi_n = 1.$$

This π will be called a distribution vector*.*

Depending on our interest, in some cases it is more comfortable to have $\pi_i \in \mathbb{C}$. However, one should note that the distribution vector π is biologically meaningful only provided $\pi_i \in \mathbb{R}_{\geq 0}$. The case $\pi_i = 0$ is excluded to avoid unnecessary technical issues in the following discussion.

Notation 2.13. *We will write $D(\pi)$ to refer to the diagonal matrix satisfying $D(\pi)_{ii} = \pi_i$, i.e. whose diagonal is the vector π.*

Definition 2.14. *The general time reversible model (GTR) is defined as*

$$\mathcal{L}_{GTR} := \{Q \in \mathcal{L}_{GMM} : \exists \boldsymbol{\pi} \mid QD(\boldsymbol{\pi}) = D(\boldsymbol{\pi})Q^T\}$$

Since we also have $Q^m D(\boldsymbol{\pi}) = D(\pi)(Q^m)^T$, one can use the absolute convergence of the matrix exponential and infer

$$e^{\mathcal{L}_{GTR}} = \{M \in e^{\mathcal{L}_{GMM}} : \exists \boldsymbol{\pi} \mid MD(\boldsymbol{\pi}) = D(\boldsymbol{\pi})M^T\}.$$

Let us go for the aforementioned proof:

Lemma 2.15. *Let $X = (X_{ij}) \in M_n(\mathbb{C})$ satisfy $XD(v) = -D(v)X$ for some vector $v = (v_1, \cdots v_n)^T$. Then, for every $i, j \in [n]$, either $X_{ij} = 0$ or $v_i = -v_j$.*

Proof. This is a simple calculation. We have

$$(XD(v))_{ij} = X_{ij}v_j \text{ and } (D(v)X)_{ij} = v_i X_{ji},$$

hence the given equality implies that, for every i, j,

$$X_{ij}v_j = -v_i X_{ij}.$$

This implies exactly what we aimed to prove.

□

Proposition 2.16. *The GTR model is not a Lie algebra.*

Proof. We only discuss the case $n \geq 3$. Otherwise it does not make much sense to talk about the GTR model.
Let us consider two symmetric matrices $Q_1, Q_2 \in \mathcal{L}_{GMM}$ (they obviously exists, since, for any symmetric matrix in $M_n(\mathbb{C})$, the diagonal terms may be changed in order to make every column sum 0). Let us also consider the distribution vector

$$\pi^T := \frac{1}{n}(1, \cdots, 1) = \frac{1}{n} \Rightarrow D(\pi) = \frac{1}{n}\mathbf{1}_n.$$

Therefore for $i \in [2]$, it obviously holds that

$$Q_i D(\pi) = D(\pi)Q^T,$$

which implies $Q_i \in \mathcal{L}_{GTR}$.

Now let us see what conditions are necessary for these two matrices to make $[Q_1, Q_2] \in \mathcal{L}_{GTR}$. Suppose there existed a distribution vector $\hat{\pi}$ which satisfied

$$[Q_1, Q_2]D(\hat{\pi}) = D(\hat{\pi})[Q_1, Q_2]^T.$$

Since Q_i are symmetric, we have

$$[Q_1, Q_2]^T = (Q_1Q_2 - Q_2Q_1)^T = Q_2Q_1 - Q_1Q_2 = -[Q_1, Q_2].$$

Hence the existence of such a vector as $\hat{\pi}$ implies that

$$[Q_1, Q_2]D(\hat{\pi}) = -D(\hat{\pi})[Q_1, Q_2],$$

which according to previous lemma leads us to the following disjunction: for every i, j, either $\hat{\pi}_i = -\hat{\pi}_j$ or $[Q_1, Q_2]_{ij} = 0$. However, the former equation cannot hold attending to the definition of distribution vector, thus for every i, j we must have $[Q_1, Q_2]_{ij} = 0$, which is nothing but a complicated way to say that $[Q_1, Q_2] = 0$.

Summarizing, so far we have proved that a necessary condition for the GTR model to be a Lie algebra is that, for every two symmetric matrices $Q_1, Q_2 \in \mathcal{L}_{GMM}$, they satisfy $[Q_1, Q_2] = 0$. It is only necessary to find a counterexample of this statement to finish our prove. Indeed, let us consider the following $n \times n$ symmetric matrices:

$$Q_1 = \begin{pmatrix} * & \alpha & \beta & 0 & \cdots & 0 \\ \alpha & * & 0 & 0 & \cdots & 0 \\ \beta & 0 & * & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix}, \; Q_2 = \begin{pmatrix} * & \alpha & \beta' & 0 & \cdots & 0 \\ \alpha & * & 0 & 0 & \cdots & 0 \\ \beta' & 0 & * & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & & 0 \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{pmatrix},$$

where the asterisks are chosen so that they have zero-sum columns. As one can easily calculate,

$$[Q_1, Q_2]_{1,3} = \alpha(\beta - \beta'),$$

hence it is enough to choose $\alpha \neq 0$ and $\beta \neq \beta'$ to have the desired counterexample, which finishes the proof.

□

Since being a Lie algebra is a sufficient but not necessary condition to be multiplicatively closed, last proposition does not finish our discussion about the closeness of the GTR model, and we need another one. For this task, we will need the Perron-Frobenious theorem, which we have stated in chapter 1 as theorem 1.10.

Proposition 2.17. *The GTR model is not multiplicatively closed.*

Proof. We demonstrate this fact only for the case when $n \geq 3$. Given two symmetric matrices $Q_1, Q_2 \in \mathcal{L}_{GMM}$, (we know they must exist), let us consider the two matrices $M_1 = e^{Q_1}$ and $M_2 = e^{Q_2}$. Matrices M_1, M_2 not only belong to $e^{\mathcal{L}_{GMM}}$, but also to $e^{\mathcal{L}_{GTR}}$, with associated distribution vector equal to $\pi = \frac{1}{n}\theta$.

Let us assume that GTR is multiplicatively closed. Then we must have $M_1 M_2 \in e^{\mathcal{L}_{GTR}}$, i.e. there must exist a distribution vector $\hat{\pi}$ such that

$$M_1 M_2 D(\hat{\pi}) = D(\hat{\pi}(M_1 M_2)^T),$$

or differently expressed taking advantage on the symmetry of M_i,

$$(1) \qquad M_1 M_2 D(\hat{\pi}) = D(\hat{\pi}) M_2 M_1.$$

Moreover, since M_i are symmetric and Markov, they satisfy $M_i\theta = \theta$, hence we also have

$$(2) \qquad M_1 M_2 \theta = \theta.$$

But that is not all, because we are still able to find another eigenvector of matrix $M_1 M_2$:

$$(3) \quad M_1 M_2 \hat{\pi} = M_1 M_2 (D(\hat{\pi}\theta)) = D(\hat{\pi}) M_2 M_1 \theta = D(\hat{\pi})\theta = \hat{\pi}.$$

The moment has come to use the Perron-Frobenious theorem. We do know that $M_1 M_2$ is a symmetric matrix, although we cannot assure it is positive. Let us assume this is true for a moment and state an example later in which this is the case.
In equations (2) and (3) we have found two eigenvectors of $M_1 M_2$, each of them strictly positive. From the uniqueness of the positive eigenvector, we infer they must be multiples of each other. Using the sum of their elements, we conclude

$$\hat{\pi} = \frac{1}{n}\theta.$$

When we plug this formula into (1), since $D(\theta) = \mathbf{1}_n$, we obtain

$$M_1 M_2 = M_2 M_1.$$

Therefore if we find Markov, symmetric matrices M_1, M_2 such that do not commute and $M_1 M_2$ is positive, we will have found a counterexample of the multiplicatively closeness of the GTR model and our proposition will be

finished.
Indeed, we can choose the $n \times n$ Markov matrices

$$M_1 = \begin{pmatrix} * & a & b & c & \cdots & c \\ a & * & c & c & \cdots & c \\ b & c & * & c & \cdots & c \\ c & c & c & c & & c \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ c & c & c & c & \cdots & c \end{pmatrix}, \; M_2 = \begin{pmatrix} * & a & b' & c & \cdots & c \\ a & * & c & c & \cdots & c \\ b' & c & * & c & \cdots & c \\ c & c & c & c & & c \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ c & c & c & c & \cdots & c \end{pmatrix},$$

where the parameters must satisfy $a, b, c > 0$ and $a + b + (n-3)c < 1$, and similarly for b', so that the asterisks can be chosen in order to make the matrices Markov and positive. Moreover, the product of positive, Markov matrices is also positive and Markov, hence $M_1 M_2$ satisfies the demanded conditions. It only remains to prove that $M_1 M_2 \neq M_2 M_1$, which can be easily checked to be the case as long as $b \neq b'$.

□

To finish this section, we derive the Lie algebra of the general Markov model $\mathcal{L}_{GMM}$, which in [7] Johnson was the first one stating. The basic definitions and some results of Lie theory are explained in section 4.2.

Let us consider the matrices $\{E_{ij}\}_{i,j\in[n]}$, whose elements are $[E_{ij}]_{kl} = \delta_{ij}\delta_{kl}$. The commutator of any two of them is easy to calculate:

$$[E_{ij}, E_{kl}] = E_{ij}\delta_{jk} - E_{kj}\delta_{il}.$$

Definition 2.18. *The* elementary rate matrices, $\{L_{ij}\}_{i\neq j}$, *are defined as*

$$L_{ij} := E_{ij} - E_{jj}.$$

Note that they are indeed rate matrices since $\theta^T L = 0$, for every $i, j \in [n]$ with $i \neq j$. Moreover, given the generic rate matrix $Q = (q_{ij})_{i,j\in[n]}$, it is clear that

$$Q = \sum_{i\neq j} q_{ij} L_{ij},$$

hence the elementary rate matrices generate $\mathcal{L}_{GMM}$. Actually we can give a more sophisticated result:

Lemma 2.19. *The matrices $\{L_{ij}\}_{i\neq j}$ form a $\mathbb{C}$-basis for the tangent space of $GL_1(n, \mathbb{C})$ (the invertible Markov matrices).*

Proof. It is known that the dimension of a tangent space is equal to the dimension of the Lie group as a manifold. In our case, the dimension of $GL_1(n,\mathbb{C})$ as a manifold is $n(n-1)$. For every $i \neq j$, consider the smooth well-defined path in $GL_1(n,\mathbb{C})$

$$A^{(ij)}(t) := e^{L_{ij}t},$$

which satisfies the well known conditions $A(0) = \mathbf{1}$ and $A'(0) = L_{ij}$. Therefore L_{ij} belongs to the tangent space of $GL_1(n,\mathbb{C})$. Moreover, since there are $n(n-1)$ of these matrices, it only remains to prove they are linearly independent. But this is obvious, for a non trivial linear combination such as

$$\sum_{i\neq j} \alpha_{ij} L_{ij} = 0$$

would imply a non trivial linear combination

$$\sum_{i,j} b_{ij} E_{ij} = 0,$$

which is an absurd.
Therefore we can conclude that the tangent space of $GL_1(n\mathbb{C})$ at the identity is $< \{L_{ij}\}_{i\neq j} >_{\mathbb{C}}$.

□

Proposition 2.20. *The rate matrices $\mathcal{L}_{GMM}$ form a Lie algebra.*

Proof. In previous lemma we have proved that $\mathcal{L}_{GMM}$ is the Lie algebra of the matrix group $GL_1(n,\mathbb{C})$. Moreover, in proposition 4.15 we proved that the Lie algebra of a matrix group is a Lie algebra.

□

The Lie algebra structure (i.e. all he possible multiplications between its generators) is easy to calculate:

$$[L_{ij}, L_{kl}] = (L_{il} - L_{jl})(\delta_{jk} - \delta_{jl}) - (L_{kj} - L_{lj})(\delta_{il} - \delta_{jl}).$$

Provided i, j, k, l are all distinct, some handy products are the following:

$$[L_{ij}, L_{kl}] = 0, \qquad [L_{ij}, L_{il}] = 0, \qquad [L_{ij}, L_{ki}] = L_{ij} - L_{kj}$$

$$[L_{ij}, L_{jl}] = L_{il} - L_{jl}, \ [L_{ij}, L_{kj}] = L_{kj} - L_{ij}, \ [L_{ij}, L_{ji}] = L_{ij} - L_{ji}.$$

With this basis, the biologically meaningful rate matrices can be nicely written. By biologically meaningful we mean the stochastic rate matrices, i.e. those with real and nonnegative off-diagonal entries; by nicely, we mean that a matrix Q can be written as $Q = \sum_{i \neq j} L_{ij}$, where the stochastic condition is satisfied provided the coefficients α_{ij} are real and nonnegative. This particular case suggests the following definition:

Definition 2.21. *A Lie algebra $\mathcal{L} \subset \mathcal{L}_{GMM}$ has a* stochastic basis *if there exists a basis $B_{\mathcal{L}} = \{L_1, \cdots, L_d\}$ of $\mathcal{L}$ such that each $\mathcal{L}_k$ is a convex linear combination of the L_{ij}. In other words, if $L_k = \sum_{i \neq j} \alpha_{ij} L_{ij}$ where $\alpha_{ij} \geq 0$. In this case, we will say that $e^{\mathcal{L}}$ is a* Lie Markov model.

Definition 2.22. *The* dimension *of a Lie Markov model $e^{\mathcal{L}}$ is the vector-space dimension of $\mathcal{L}$.*

Regarding last definitions, in most of cases it is very intuitive: it coincides to the number of free parameters of the model. In general Markov model, we have $n(n-1)$ free parameters, hence the dimension is $n(n-1)$; in the Jukes-Cantor model, we have one free parameter, hence the dimension is 1.

2.4 Permutation symmetries of Markov models

In general terms, looking for permutation symmetries means looking for some invariant through permutations of nucleotides. Finding the correct invariant is the delicate task we must accomplish. Along this section we will label nucleotides A, C, G, T with the integers $1, 2, 3, 4$ respectively.

Now let us imagine we start performing a maximum likelihood inference method, represented as F. Our model is, let us say, a two-dimensional continuous-time Markov model with rate-matrix $Q = \alpha_1 L_1 + \alpha_2 L_2$. Attending to some data $\mathcal{D}$, one numerical matrix of this kind is assigned to every edge of our binary tree $\mathcal{T}$, and also an edge weight θ. In other words, F is a function that returns maximum likelihoods estimates of the free parameters of the model, hence $F(\mathcal{D}) = (\hat{\alpha_1}, \hat{\alpha_2}, \hat{\theta})$.
Let us now proceed analogously, although this time we will permutate the rate parameters in Q, setting our new model to be $Q' = \alpha_2 L_1 + \alpha_1 L_2$; let us name this new process with the function F'. We claim that this new function will return the same maximum likelihood estimates as F, more exactly that the estimated parameters will be $\alpha_2 = \hat{\alpha_1}$ and $\alpha_1 = \hat{\alpha_2}$ i.e. that we will have $F'(\mathcal{D}) = (\hat{\alpha_2}, \hat{\alpha_1}, \hat{\theta})$. The order changes, but we should obtain the same values, for the difference between the two mute parameters α_1 and α_2 is simply the labeling, hence the rate matrix $\hat{Q}$ should the same.

This reasoning can be easily generalized to more free parameters and motivates the following way to characterize the symmetry of a model (section 4.3 explains some results on the symmetric group):

Definition 2.23. *We will say that a Lie Markov model $\mathcal{L}$ has the* symmetry *of the group $G \leq \mathcal{G}_n$ if there is a basis $B_{\mathcal{L}} = \{L_1, \cdots, L_d\}$ of $\mathcal{L}$ such that*

$$\sigma \cdot B_{\mathcal{L}} = \{K_\sigma L_1 K_\sigma^{-1}, \cdots K_\sigma L_d K_\sigma^{-1}\} = B_{\mathcal{L}},\ \forall \sigma \in G$$

and G is the largest subgroup of $\mathcal{G}_n$ with this property.

Therefore the subgroup G acts by permutating the elements of a basis $B_{\mathcal{L}}$. There is a interesting fact about this symmetry: if we fix a basis $B_{\mathcal{L}}$, it induces a group homomorphism $\rho : G \leq \mathcal{G}_n \mapsto \mathcal{G}_d$, where d is the dimension of the model. More explicitly said, if we are given a permutation $\sigma \in G$, we can define a permutation of the d elements of the basis, exactly as

$$(L_1, \cdots, L_d) \mapsto \{K_\sigma L_1 K_\sigma^{-1}, \cdots K_\sigma L_d K_\sigma^{-1}\} = (L_{\rho(\sigma)(1)}, \cdots, L_{\rho(\sigma)(n)}),$$

where each $\rho(\sigma)(i)$ is determined by the respective element of the base assigned to $K_\sigma L_i K_\sigma^{-1}$. Therefore our application is exactly

$$\begin{aligned} \rho\colon G \leq \mathcal{G}_n &\to \mathcal{G}_d \\ \sigma &\mapsto \rho(\sigma). \end{aligned}$$

Moreover, our definition of symmetry implies that $\mathcal{L}$ is also invariant when considered as a vector space: if $\mathcal{L} =< L_1, \cdots, L_d >_{\mathbb{C}}$ then $\sigma \cdot \mathcal{L} = \mathcal{L}$.

Saying about a model that it has $\mathcal{G}_n$ symmetry is a relevant feature: it means that the model does not distinguish any kind of preferred grouping between its nucleotides. Therefore any statistical inference method we use will return the same output no matter which order of the nucleotides we use. We want to give an example of a Lie Markov model with $\mathcal{G}_n$ symmetry, but first we need a lemma:

Lemma 2.24. *If $B_{GMM} = \{L_{ij}\}_{i \neq j}$ is the base of the general Markov model, and we are given $\sigma \in \mathcal{G}_n$, then we have*

$$K_\sigma L_{ij} K_\sigma^{-1} = L_{\sigma(i)\sigma(j)}.$$

Proof. Let us consider the canonical basis vectors $e_1, \cdots, e_n \in \mathbb{C}^n$, which satisfy $K_\sigma e_i = e_{\sigma(i)}$ for any $\sigma \in \mathcal{G}_n$. One can easily check that

$$L_{ij} e_k = \delta_{jk}(e_i - e_k),$$

hence substituting i, j by $\sigma(i), \sigma(j)$ we get

$$L_{\sigma(i)\sigma(j)} e_k = \delta_{\sigma(j)k}(e_i - e_k).$$

Using again the first of these two equalities, together with the definition of K_σ^{-1}, we get

$$K_\sigma L_{ij} K_\sigma^{-1} e_k = K_\sigma L_{ij} e_{\sigma^{-1}(k)} = \delta_{j\sigma^{-1}(k)}(e_{\sigma(i)} - e_k).$$

If we prove that $\delta_{j\sigma^{-1}(k)} = \delta_{\sigma(j)\sigma(k)}$, we will be done, because both matrices of the statement will have the same (ordered) image of the canonical base. But actually this is obvious, because, nearly by definition, $j = \sigma^{-1}(k)$ iff $\sigma(j) = k$. This finishes the proof.

□

Proposition 2.25. *The general n-state Markov Lie algebra $\mathcal{L}_{GMM}$ has $\mathcal{G}_n$ symmetry.*

Proof. Since $\mathcal{L}_{GMM}$ has, as we have repeatedly said, basis $B_{GMM} = \{L_{ij}\}_{ij}$, using the previous lemma we deduce that

$$\sigma \cdot B_{GMM} = \{L_{\sigma(i)\sigma(j)}\}_{i \neq j} = \{L_{ij}\}_{i \neq j} = B_{GMM}$$

for any $\sigma \in \mathcal{G}_n$, which respects our definition of symmetry.

□

2.5 Producing Markov models with $\mathcal{G}_4$ symmetry.

Given the number of states n, we would like to classify every Lie Markov model. This is an ambitious task, which can be reformulated as finding all subalgebras of $\mathcal{L}_{GMM}$. However, it is not only the subalgebra *per se* what interests us, but also finding an appropriate stochastic base as the one given in definition 2.21. This fact, together with the concept of *symmetry* stated in last section, suggests an easier approach to state and study some Lie Markov models: if we restrict ourselves to the models satisfying some kind of symmetry (the most exigent one being $\mathcal{G}_n$) we may come up with an easy expression of the basis.
To this end, it is convenient to take advantage on some group representation theory results, which we state in section 4.3. We will follow the notation used in that section.

2.5.1 Decomposing $\mathcal{L}_{GMM}$

From now on, since we are interested in nucleotide evolution, we fix $n = 4$. The first thing we aim to do is to decompose the general Markov model into

irreducible representations of $\mathcal{G}_4$, which we will do following proposition 4.23. We will use the projection operators Θ_λ to find the integers c^λ of the decomposition

$$\mathcal{L}_{GMM} \cong \oplus_\lambda c_\lambda V^\lambda.$$

From now on, instead of V^λ, we will abuse of the notation and simply write the partition $\lambda = \{\lambda_1^{n_1} \cdots \lambda_s^{n_s}\}$. The partitions of $n = 4$, and therefore the irreducible representations of $\mathcal{G}_4$, are $\{4\}$, $\{31\}$, $\{2^2\}$, $\{21^2\}$. The corresponding character functions are given in the following table:

Table 1 : Characters of $\mathcal{G}_4$

	$\chi^{\{4\}}$	$\chi^{\{31\}}$	$\chi^{\{2^2\}}$	$\chi^{\{21^2\}}$	$\chi^{\{1^4\}}$
e	1	3	2	3	1
[(12)]	1	1	0	-1	-1
[(123)]	1	0	-1	0	1
[(12)(34)]	1	-1	2	-1	1
[(1234)]	1	-1	0	1	-1

The brackets indicate we are referring to all the conjugacy class of the permutation inside. They have, respectively, orders 1, 6, 8, 3 and 6. Recall that the character function only depends on this class, i.e. $\chi(s^{-1}\sigma s) = \chi(\sigma)$ for any $\sigma, s \in \mathcal{G}_4$. It is noticeable that the first row, i.e. $\chi^\lambda(e)$, indicates the dimension of the representation λ. Note also that there are exactly two one-dimensional representations, which are $\{4\}$, the *trivial* representation (every permutation is mapped to the identity), and $\{1^4\}$, the *sign* representation in which each permutation is mapped to either 1 or -1 depending on the sign of the permutation.

Before continuing, we need to study the *defining representation* of $\mathcal{G}_4$, which is defined as follows: for the $\mathbb{C}^4$ vector space generated by e_i, $i \in [4]$, we define the action of $\mathcal{G}$ as $\sigma e_i = e_{\sigma(i)}$. We know that the partitions of $n = 4$ are $\{4\}$, $\{31\}$, $\{2^2\}$, $\{21^2\}$, hence it is only necessary to find the coefficients. Let us consider the operators

$$\Theta_4 = \frac{1}{24}\sum_{\sigma\in\mathcal{G}_4} \chi^{\{4\}}(\sigma)\sigma = \frac{1}{24}\sum_{\sigma\in\mathcal{G}_4} \sigma$$

$$\Theta_{31} = \frac{1}{24}\sum_{\sigma\in\mathcal{G}_4} \chi^{\{31\}}(\sigma)\sigma =$$

$$= \frac{1}{24}\Big(3e + \sum_{\sigma\in[(12)]} \sigma - \sum_{\sigma\in[(12)(34)]} \sigma - \sum_{\sigma\in[(1234)]} \sigma\Big)$$

Noticeably, we have $\Theta_{\{4\}}(e_1) = \frac{1}{4}(e_1 + e_2 + e_3 + e_4)$ (actually for any e_i,

but that is not the point now) and $\Theta_{\{31\}}(e_1) = \frac{1}{24}(6e_1 - 2e_2 - 2e_3 - 2e_4)$. The important thing is that they are not zero, therefore our decomposition of $\mathbb{C}^4$ must contain the irreducibles $\{4\}$ and $\{31\}$. As we saw in the table, the module $\{4\}$ has dimension 1, while the module $\{31\}$ has dimension 3. Since the dimension of $\mathbb{C}^4$ is $4 = 3 + 1$, we conclude that the irreducible decomposition of $\mathbb{C}^4$ is

$$\mathbb{C}^4 = \{4\} \oplus \{31\}.$$

We will use this result in order to decompose in irreducible components the algebra $\mathcal{L}_{GMM} =< \{L_{ij}\}_{1\leq i\neq j\leq 4} >_{\mathbb{C}}$ as a $\mathcal{G}_4$ representation, which, as we proved in lemma 2.24, acts like $\sigma L_{ij} = \mathcal{L}_{\sigma(i)\sigma(j)}$.
Our trick will consist on taking advantage on the similarities between this representation and the one of $\mathcal{G}_4$ acting on the tensor product space

$$\mathbb{C}^4 \otimes \mathbb{C}^4 \cong < \{e_i \otimes e_j\}_{i,j\in[4]} >,$$

defined as

$$\sigma(e_i \otimes e_j) = e_{\sigma(i)} \otimes e_{\sigma(j)}.$$

Indeed, both vector spaces behave similarly, except that the diagonal terms in $\mathcal{L}_{GMM}$ are determined by the rest of elements. Therefore, it is easy to find an isomorphism between the action of $\mathcal{G}_4$ on $\mathcal{L}_{GMM}$ and the one on the subspace of $\{\psi \in \mathbb{C}^4 \otimes \mathbb{C}^4 : \psi_{ii} = 0, i \in [4]\}$. We know that doing the tensorial product means doing the Kronecker product of the matrices involved. Using proposition 4.21 and the table of characters, we see that

$$\mathbb{C}^4 \otimes \mathbb{C}^4 \cong (\{4\} \oplus \{31\}) \otimes (\{4\} \oplus \{31\}) = 2\{4\} \oplus 3\{31\} \oplus \{2^2\} \oplus \{21^2\},$$

and on the other side, the subspace spanned by $e_i \otimes e_i$ is isomorphic to the defining representation spanned by e_i (the isomorphism is obvious), hence

$$< \{e_i \otimes e_i\}_{i\in[4]} > \cong \{4\} \oplus \{31\}.$$

Since this module must be contained in $\mathbb{C}^4 \oplus \mathbb{C}^4$, and since its elements do not satisfy $\psi_{ii} = 0$, we conclude that $\mathcal{L}_{GMM}$ is essentially contained in the rest of moduli, explicitly $\{4\} \oplus 2\{31\} \oplus \{2^2\} \oplus \{21^2\}$. Using the dimension of these vector spaces, namely 12, we see they must be equal. Therefore we can state the following:

Proposition 2.26. *If $n = 4$, the decomposition of the general Markov model $\mathcal{L}_{GMM}$ as a $\mathcal{G}_4$ module is given by*

$$\mathcal{L}_{GMM} \cong \{4\} \oplus 2\{31\} \oplus \{2^2\} \oplus \{21^2\},$$

whose dimension decomposition is $12 = 1 + 2 \times 3 + 2 + 3$.

2.5.2 A convenient basis for $\mathcal{L}_{GMM}$

We aim to state a basis for the decomposition of $\mathcal{L}_{GMM}$ just stated in proposition 2.26, making each submodule being associated to a basis.

To begin with, we come up with the vector

$$L_{id} := \sum_{1\leq i\neq j\leq 4} L_{ij} = \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix},$$

which is the generator of the well known $\mathcal{G}_4$-invariant module $\{4\}$, in which every permutation σ is sent to the identity, i.e. $\sigma L_{id} = L_{id}$. Hence we have the first module of proposition 2.26, $\{4\}$.

Let us define 8 new vectors, namely the row sum vectors and the column sum vectors. For any $i \in [4]$, the *row sum vectors* are

$$R_i := \sum_{j:1\leq i\neq j\leq 4} L_{ij},$$

while the *column sum vectors* are

$$C_i := \sum_{j:1\leq i\neq j\leq 4} L_{ji}.$$

The group $\mathcal{G}_4$ acts on these vectors as follows:

$$\sigma R_i = \sum_{j:1\leq i\neq j\leq 4} L_{\sigma(i)\sigma(j)} = R_{\sigma(i)},$$

$$\sigma C_i = \sum_{i:1\leq i\neq j\leq 4} L_{\sigma(j)\sigma(i)} = C_{\sigma(i)}.$$

Nicely, the sets $\{R_i\}_{i\in[4]}$ and $\{C_i\}_{i\in[4]}$ are invariant under the action of $\mathcal{G}_4$. And not only that, but we clearly see that these actions are isomorphic to the defining representation $\sigma e_i = e_{\sigma(i)}$. Therefore we can write

$$< R_1, R_2, R_3, R_4 >_{\mathbb{C}} \cong < C_1, C_2, C_3, C_4 >_{\mathbb{C}} \cong \{4\} \oplus \{31\}.$$

Attending to their definitions, these vectors satisfy

$$R_1 + R_2 + R_3 + R_4 = C_1 + C_2 + C_3 + C_4 = L_{id},$$

therefore we can say that

$$< L_{id}, \{R_i\}_{i\in[4]}, \{C_i\}_{i\in[4]} >_{\mathbb{C}} \cong \{4\} \oplus 2\{31\},$$

hence we have already accounted the first two moduli in proposition 2.26.

Let us define 3 more vectors. We consider

$$L_\alpha = L_{12} + L_{21} + L_{34} + L_{43},$$
$$L_\beta = L_{13} + L_{31} + L_{24} + L_{42},$$
$$L_\gamma = L_{14} + L_{41} + L_{23} + L_{32}.$$

First of all we must check that these three vectors are $\mathcal{G}_4$-invariant. In order to do this, we consider the set of unordered bipartitions of $\mathcal{G}_4$:

$$S = \{12|34, 13|24, 14|23\},$$

where $ij|kl$ indicates $\{\{i,j\},\{k,l\}\}$. We consider the following map :

$$ij|kl \mapsto L_{ij} + L_{ji} + L_{kl} + L_{lk},$$

which is well defined and establishes a bijection with the set $\{L_\alpha, L_\beta, L_\gamma\}$. Easily we see that the set S is invariant under the following action of $\mathcal{G}_4$:

$$\sigma(ij|kl) = \sigma(i)\sigma(j)|\sigma(k)\sigma(l).$$

Therefore we conclude the same for the set $\{L_\alpha, L_\beta, L_\gamma\}$, which is essentially what we wanted to prove. Moreover, since $L_\alpha + L_\beta + L_\gamma = L_{id}$, and taking into account the dimensions of the decomposition of proposition 2.26, we conclude that the only possibility is that

$$< L_\alpha, L_\beta, L_\gamma >_{\mathbb{C}} \cong \{4\} \oplus \{2^2\},$$

hence with these three vectors we have accounted already the third module of proposition 2.26.

It only remains to find generators for the module $\{21^2\}$. We will do this using the projection operator $\Theta_{\{21^2\}}$. Let us consider the six antisymmetric vectors

$$A_{ij} = L_{ij} - L_{ji},$$

where as always $i, j \in [4]$ and $i \neq j$. Since we are considering all possible combinations i, j, they are invariant under the action of $\mathcal{G}_4$. One can check that

$$P_{ij} := 12 \cdot \Theta_{\{21^2\}} A_{ij} = 2A_{ij} - A_{ik} - A_{il} + A_{jk} + A_{jl},$$

where all indexes are different and $P_{ij} = -P_{ji}$. This equality implies that

$$< \{P_{ij}\}_{1 \leq i \neq j \leq 4} >_{\mathbb{C}} = < \{P_{ij}\}_{1 \leq i < j \leq 4} >_{\mathbb{C}}$$

has at most dimension 6. But now we notice that, for any j, we have

$$\sum_i P_{ij} = 0,$$

(if $i = j$ the summand is ignored) which together with the previous equality bounds the dimension by 4. Now we use the analogous for every j,

$$\sum_i P_{ji} = 0,$$

which bounds it up to 3. Since the projections of the $\mathcal{G}_4$-invariant P_{ij} on the module $\{2^2\}$ are not zero, and since the dimension of $\{21^1\}$ is 3, we conclude that we must have

$$< \{P_{ij}\}_{1\leq i\neq j\leq 4} >_{\mathbb{C}} \cong \{21^2\},$$

as desired.
We can summarize all these results in a sole proposition:

Proposition 2.27. *If $n = 4$, the Lie algebra of the general Markov model $\mathcal{L}_{GMM}$ can be expressed as*

$$\begin{aligned}\mathcal{L}_{GMM} &=< \{L_{ij}\}_{1\leq i\neq j\leq 4} >_{\mathbb{C}} = \\ &\cong < \{L_{id}\} \cup \{L_\alpha, L_\beta, L_\gamma\} \cup \{R_i\}_{i\in[4]} \cup \{C_i\}_{i\in[4]} \cup \{P_{ij}\}_{1\leq i\neq j\leq 4} >_{\mathbb{C}},\end{aligned}$$

with the following linear dependences:

$$\begin{aligned}&L_{id} = L_\alpha + L_\beta + L_\gamma = \sum_i R_i = \sum_i C_i \\ &\sum_i P_{ij} = \sum_i P_{ij} = 0 \text{ for any } j \in [4] \\ &P_{ij} = -P_{ji} \text{ for any } i, j \in [4] \text{ , } i \neq j.\end{aligned}$$

Moreover, the decomposition into modules is

$$\begin{aligned}&< \{L_{id}\} >_{\mathbb{C}} \{4\} >_{\mathbb{C}} \cong \{4\} \\ &< \{L_\alpha, L_\beta, L_\gamma\} >_{\mathbb{C}} \cong \{4\} \oplus \{2^2\} \\ &< \{C_i\}_{i\in[4]} >_{\mathbb{C}} \cong < \{R_i\}_{i\in[4]} >_{\mathbb{C}} \cong \{4\} \oplus \{31\} \\ &< \{P_{ij}\}_{1\leq i\neq j\leq 4} >_{\mathbb{C}} \cong \{21^2\}.\end{aligned}$$

The Lie algebra of the general Markov model $\mathcal{L}_{GMM}$ is stated in result 11 of [4]. The authors of this paper needed to carry out "tedious matrix computations".

2.5.3 The Lie Markov models with $\mathcal{G}_4$ symmetry.

Our strategy will be the following: For every subgroup $H \leq \mathcal{G}_4$, we will consider the quotient group $G = \mathcal{G}_4/H$. As we did at the beginning of

section 4.3.1, we will span it as a vector space, i.e. $< G >_{\mathbb{C}}$, and we can define the action of $\mathcal{G}_4$ on the base of $< G >_{\mathbb{C}}$ as

$$\sigma[s] = [\sigma s],$$

where $\sigma \in \mathcal{G}_4$ and $[s] \in \mathcal{G}_4/H$. It is easy to check that this action is well defined. Therefore what we will have after this process is a representation of $\mathcal{G}_4$. We decompose it and check whether or not it can be contained in our module decomposition of $\mathcal{L}_{GMM}$. If that is the case, we have to see whether it exists or not a $\mathcal{G}_4$-symmetric base $B_{\mathcal{L}}$ such that this representation induces a Lie Markov model , i.e. such that $< \mathcal{G}_4/H >_{\mathbb{C}} \cong \mathcal{L}$ for some Lie Markov model $\mathcal{L}$ (we have to check it is a Lie algebra and a stochastic basis exists). We repeat the process for every $H \leq \mathcal{G}_4$.
The virtue of this reasoning is that it essentially (i.e. up to module isomorphism) finds every Lie Markov model with $\mathcal{G}_4$ symmetry. We briefly justify this claim:

Suppose we have a (this a is very important!) basis of some Lie Markov model $B_{\mathcal{L}} = \{L_1, \cdots, L_d\}$, let us assume it satisfies definition 2.23 with $\mathcal{G}_4$, i.e. $B_{\mathcal{L}}$ has $\mathcal{G}_4$-symmetry. We will also assume that no subsets of B have $\mathcal{G}_4$ symmetry, as later will be conveniently justified. Let us recall that this base induces a group homomorphism $\rho : \mathcal{G}_4 \to \mathcal{G}_d$, depending on how the base permutes. Therefore summarizing, on the one side we have that, for any $\sigma \in \mathcal{G}_4$,

$$\sigma B_{\mathcal{L}} = B_{\mathcal{L}}.$$

This invariance receives the following denomination: $B_{\mathcal{L}}$ is an orbit of $\mathcal{G}_4$. More concretely, $B_{\mathcal{L}}$ is a minimal orbit of $\mathcal{G}_4$, since it has no suborbits, by assumption. On the other side, if for some $i \in [d]$ we write

$$\mathcal{G}_4 L_i := \{\sigma(L_i)\}_{\sigma \in \mathcal{G}_4} = \{L_{\rho(\sigma)(i)}\}_{\sigma \in \mathcal{G}_4},$$

we can see that

$$\mathcal{G}_4 L_i := B_{\mathcal{L}}.$$

Therefore orbits are determined by one of its elements. The *orbit stabilizer theorem* states that there exists a bijection which makes our set $\mathcal{G}_4 L_i = B_{\mathcal{L}}$ correspond to $\mathcal{G}_4/{\mathcal{G}_4}^{L_i}$, where

$$\mathcal{G}_4^{L_i} = \{\sigma \in \mathcal{G}_4 : \sigma(L_i) := L_{\rho(\sigma)(i)} = L_i\}$$

is the *stabilizer* of the element $L_i \in B$. Actually, the element L_i we consider has no importance, because this will only conjugate our set $\mathcal{G}_4^{L_i}$, hence an isomorphic group will be originated. Therefore we will write:

$$\mathcal{G}_4^{B_{\mathcal{L}}} := \mathcal{G}_4^{L_1}.$$

Moreover the orbit stabilizer theorem implies

$$|B_{\mathcal{L}}| \cdot |\mathcal{G}_4/{\mathcal{G}_4}^{B_{\mathcal{L}}}| = |\mathcal{G}_4|,$$

and that is the most important part: the existence of a $\mathcal{G}_4$-symmetric base $B_{\mathcal{L}}$ of size d requires the existence of a subgroup $\mathcal{G}_4^{B_{\mathcal{L}}}$ of size d. If we work backwards, we can consider every $H \leq \mathcal{G}_4$ (which would be our unknown $\mathcal{G}_4^{B_{\mathcal{L}}}$) and look for its associated basis $B_{\mathcal{L}}$. However, if we simply considered some H and kindly tried to look for a base $B_{\mathcal{L}}$ such that all its elements were stabilized by H, this would be a very inefficient procedure. Hence at this point the method we mentioned at the beginning of this subsection comes into play: we will use the decomposition of $< G/H >_{\mathbb{C}}$ to determine whether or not it can be contained in

$$\mathcal{L}_{GMM} \cong \{4\} \oplus 2\{31\} \oplus \{2^2\} \oplus \{21^2\}.$$

It remains only a detail we must take into account. The sum of irreducible components in the decomposition of $\mathcal{L}_{GMM}$ is a consequence of the existence of subsets of $B_{\mathcal{L}}$ which also have $\mathcal{G}_4$-symmetry. Therefore, when our decomposition of $< G/H >_{\mathbb{C}}$ contains more than one irreducible, we will have to construct subalgebras of $\mathcal{L}$ which are consistent with this decomposition.

In order to carry out the mentioned procedure, we copy here the decomposition table which appears in the paper of Sumner et al. :

$H \leq \mathfrak{S}_4$	Copies	Cardinality $= \frac{\lvert\mathfrak{S}_4\rvert}{\lvert H\rvert}$	Decomposition of $\langle \mathfrak{S}_4/H \rangle_{\mathbb{C}}$	Model
$\{e\}$	1	24	$\{4\}\oplus 3\{31\}\oplus 2\{2^2\}\oplus 3\{21^2\}\oplus\{1^4\}$	–
$\mathbb{Z}_2$	6	12	$\{4\}\oplus 2\{31\}\oplus\{2^2\}\oplus\{21^2\}$	GMM
"	3	"	$\{4\}\oplus\{31\}\oplus 2\{2^2\}\oplus\{21^2\}\oplus\{1\}$	–
$\mathbb{Z}_3$	4	8	$\{4\}\oplus\{31\}\oplus\{21^2\}\oplus\{1^4\}$	–
$\mathbb{Z}_4$	3	6	$\{4\}\oplus\{2^2\}\oplus\{21^2\}$	–
$\mathbb{Z}_2 \times \mathbb{Z}_2$	3	6	$\{4\}\oplus 2\{2^2\}\oplus\{1^4\}$	–
"	1	"	$\{4\}\oplus\{31\}\oplus\{2^2\}$	F81+K3ST
$\mathfrak{S}_3$	4	4	$\{4\}\oplus\{31\}$	F81
$\mathbb{Z}_2 \wr \mathbb{Z}_2$	3	3	$\{4\}\oplus\{2^2\}$	K3ST
A_4	1	2	$\{4\}\oplus\{1^4\}$	–
$\mathfrak{S}_4$	1	1	$\{4\}$	Jukes–Cantor

Figure 2.3: Decomposition of the orbits of $\mathcal{G}_4$ into irreducible modules.

In the last column of the table we have the Lie Markov models with $\mathcal{G}_4$ symmetry and isomorphic to $< \mathcal{G}_4/H >$. The only task which remains to do is proving that indeed those and only those are the Lie Markov models with $\mathcal{G}_4$ symmetry.

First of all, let us not that the module $\{4\}$ appears exactly once in every decomposition. This implies that the sum of modules $< \mathcal{G}_4/H >_{\mathbb{C}}$ cannot

compose new submodules of the GMM. Using the orbit stabilizer theorem, this means that every orbit of the possible symmetric bases $B_{\mathcal{L}}$ is minimal (i.e. with no suborbits), as we mentioned before. Moreover, recall that the cardinality of $\frac{|\mathcal{G}_4|}{|H|}$ is the same as $|B_{\mathcal{L}}|$. Since the cardinalities $11, 10, 9, 7, 5$ are not in the table, the only possibility remaining is that a Lie Markov model with this dimension is the composition of other Lie Markov models. However, in the decomposition of $\mathcal{L}_{GMM}$ of proposition 2.26 and in each of the $< G/H >_{\mathbb{C}}$ the trivial module $\{4\}$ appears exactly once, hence a sum of them would contain $2\{4\}$. But in such a case it is impossible that it is contained in $\mathcal{L}_{GMM}$, hence we conclude:

Proposition 2.28. *If $n = 4$, there are no $\mathcal{G}_4$-symmetric Lie Markov models with dimension $11, 10, 9, 7, 5$.*

Actually, we can easily discard two more dimensions with no effort. Cardinalities 2 and 8 only contain decompositions with the module $\{1^4\}$, which is not contained in the decomposition of $\mathcal{L}_{GMM}$. Therefore we can state another proposition:

Proposition 2.29. *If $n = 4$, there are no $\mathcal{G}_4$-symmetric Lie Markov models with dimension $8, 2$.*

Now, since we start the difficult ones, let us work organizedly:

Dimension 1

From the table, we see that there is only one orbit with cardinality one. Hence the associated Lie Markov model must be isomorphic to the trivial $\{4\}$. The base of this model is obviously $\{L_{id}\}$, and it is evident we are referring to the Jukes-Cantor model $\mathcal{L}_{JC}$. A generic rate-matrix is

$$Q = \alpha L_{id} = \alpha \begin{pmatrix} -3 & 1 & 1 & 1 \\ 1 & -3 & 1 & 1 \\ 1 & 1 & -3 & 1 \\ 1 & 1 & 1 & -3 \end{pmatrix}.$$

Therefore:

Proposition 2.30. *If $n = 4$, there exists exactly one Lie Markov model with dimension 1, which is the Jukes-Cantor model.*

Dimension 3

There is only one option with cardinality 3, with decomposition $\{4\} \oplus \{2^2\}$. Using proposition 2.27, we choose

$$< \{L_\alpha, L_\beta, L_\gamma\} >_{\mathbb{C}} \cong \{4\} \oplus \{2^2\},$$

which has abelian Lie algebra, for

$$[L_\alpha, L_\beta] = [L_\alpha, L_\gamma] = [L_\beta, L_\gamma] = 0.$$

A generic rate matrix is

$$Q = \alpha L_a + \beta L_b + \gamma L_\gamma = \begin{pmatrix} * & \alpha & \beta & \gamma \\ \alpha & * & \gamma & \beta \\ \beta & \gamma & * & \alpha \\ \gamma & \beta & \alpha & * \end{pmatrix},$$

where the asterisks are determined by the column sum zero condition. We see it is actually the Kimura 3ST model, hence:

Proposition 2.31. *If $n = 4$, there exists exactly one Lie Markov model with dimension 3, which is the Kimura 3ST model.*

Dimension 4

The sole option with cardinality 4 is $\{4\}\oplus\{31\}$. Attending to proposition 2.27, we have the two options:

$$< \{C_i\}_{i\in[4]} >_{\mathbb{C}} \cong \{4\} \oplus \{31\},$$

$$< \{R_i\}_{i\in[4]} >_{\mathbb{C}} \cong \{4\} \oplus \{31\}.$$

However, $[C_i, C_j] = R_j - R_i - P_{ij}$, hence the first vector space does not form a Lie algebra and only the second option remains. Indeed, it satisfies

$$[R_i, R_j] = R_i - R_j,$$

hence it forms a Lie algebra. A generic matrix of this model is

$$Q = aR_1 + bR_2 + cR_3 + dR_4 = \begin{pmatrix} * & a & a & a \\ b & * & b & b \\ c & c & * & c \\ d & d & d & * \end{pmatrix},$$

which is the Felsenstein 81 model. Therefore:

Proposition 2.32. *If $n = 4$, there exists exactly one Lie Markov model with dimension 4, which is the Felsenstein 81 model.*

Dimension 6

Attending to the decomposition table, the only possibilities are the cases $H = \mathbb{Z}_4$, which gives decomposition $\{4\} \oplus \{2^2\} \oplus \{21^2\}$, and $H = \mathbb{Z}_2 \times \mathbb{Z}_2$,

although only the one with decomposition $\{4\} \oplus \{31 \ \oplus \{2^2\}$ because the other one has the submodule $\{1^4\}$, not contained in $\mathcal{L}_{GMM}$.

If $H = \mathbb{Z}_4$, we have

$$< \mathcal{G}_4/H >_{\mathbb{C}} \cong \{4\} \oplus \{2^2\} \oplus \{21^2\}.$$

Using proposition 2.27, we get

$$< \{P_{ij}\}_{1 \leq i \neq j \leq 4} \cup \{L_\alpha, L_\beta, L_\gamma\} >_{\mathbb{C}} \cong \{4\} \oplus \{2^2\} \oplus \{21^2\}.$$

However, this is not a Lie algebra, for

$$[L_\alpha, P_{12}] = 2L_\alpha + 2L_\beta + 2L_\gamma - 4R_2 - 2R_3 - 2R_4 + 2C_1 - 2C_2,$$

hence we discard it.

If $H = \mathbb{Z}_2 \times \mathbb{Z}_2$, as we said we have

$$< \mathcal{G}_4/H >_{\mathbb{C}} \cong \{4\} \oplus \{31\} \oplus \{2^2\},$$

hence using proposition 2.27 we come up with two possibilities:

$$< \{L_\alpha, L_\beta, L_\gamma\} \cup \{R_i\}_{i \in [4]} >_{\mathbb{C}} \cong \{4\} \oplus \{31\} \oplus \{2^2\},$$

$$< \{L_\alpha, L_\beta, L_\gamma\} \cup \{C_i\}_{i \in [4]} >_{\mathbb{C}} \cong \{4\} \oplus \{31\} \oplus \{2^2\},$$

with the linear dependences

$$L_{id} = \sum_{i \in [4]} R_i = \sum_{i \in [4]} C_i = L_\alpha + L_\beta + L_\gamma.$$

As for the second option, we have $[C_i, C_j] = R_j - R_i - P_{ij}$, hence it does not form a Lie algebra. As for the first one, we have already proved that the Kimura 3ST and the Felsenstein 81 are closed under Lie brackets, hence we nly need to check the crossed ones, which gives

$$[L_{ij|kl}, R_i] = R_j - R_i,$$

hence it forms a Lie algebra. This model will be referred as $K3ST + F81$. Using the linear dependences given above, one can see that model $K3ST + F81$ has indeed dimension 6.
In the article, it is proposed the stochastic base $B_{K3ST+F81}$ with elements

$$\begin{aligned}
W_{12} &:= L_\alpha + (R_1 + R_2) \\
W_{13} &:= L_\beta + (R_1 + R_3) \\
W_{14} &:= L_\gamma + (R_1 + R_4) \\
W_{23} &:= L_\gamma + (R_2 + R_3) \\
W_{24} &:= L_\beta + (R_2 + R_4) \\
W_{34} &:= L_\alpha + (R_3 + R_4),
\end{aligned}$$

which yields that a generic matrix of this model has the form

$$Q = \alpha W_{12} + \overline{\alpha} W_{34} + \beta W_{13} + \overline{\beta} W_{24} + \gamma W_{14} + \overline{\gamma} W_{23}.$$

If $ij|kl$ is a bipartition of [4], the Lie brackets are

$$[W_{ij}, W_{kl}] = 2(W_{ij} - W_{kl}),$$

while if $ij|i'j'$ and $kl|k'l'$ are distinct bipartitions, then

$$[W_{ij}, W_{kl}] = 2(W_{ij} - W_{i'j'}) - 2(W_{kl} - W_{k'l'}).$$

Finally, we can conclude:

Proposition 2.33. *If $n = 4$, there exists exactly one Lie Markov model with dimension 6, which is the K3ST + F81 model.*

The conclusion of our discussion can be summarized in the following theorem, with which we finish this section:

Theorem 2.34. *On four states, there are exactly five Lie Markov models with $\mathcal{G}_4$ symmetry. These are the Jukes Cantor model, with dimension 1; the Kimura 3ST model, with dimension 3; the Felsenstein 81 model, with dimension 4; the K3ST + F81 model, with dimension 6; and the General Markov model, with dimension 12.*

Chapter 3

Implementing Lie Markov models in IQ-TREE

Woodhams et al. (2015), in their article [5], gave a detailed and clear exposition of Lie Markov models sensitive to the grouping of nucleotides into purines (R) and pyrimidines (Y). Compared against the GTR model, they concluded that their performance was satisfying. Moreover, their biological interpretation is consistent with heterogeneity (due to their multiplicative closeness), hence Lie Markov models have proved to deserve being taken into account and implemented in algorithms for phylogenetic inference. Currently, this task is being carried out by the developers of the algorithm IQ-TREE (also with the collaboration of Woodhams himself), exposed in the article [9].
Along this section, we aim to briefly explain the necessary parts of the article [5] and our modest contribution to the implementation of Lie Markov models in IQ-TREE.

3.1 The necessary objects, as taken from [5]

First of all, we should introduce some terminology. When considering the four bases A, C, G, T, attending to their structure, they are typically grouped in purines (R) and pyrimidines (Y). Purines include adenine (A) and guanine (G), while pyrimidines include thymine (T) and cytosine (C). Schematically written, we have:

$$R = \{A, G\}; \quad Y = \{C, T\}.$$

When a mutation occurs from a purine into a purine, or from a pyrimidine into a pyrimidine, we call it a transition. A mutation from R into Y, or

from Y into R, receives the name of transversion. All these classifications are justified because of the biological fact that transitions occur at higher rate than transversions, due to their structure resemblances.
The mathematical consequences of this fact is that, when modeling an evolutionary process, one aims to choose rate-matrices which maintain this RY-grouping. More rigorously speaking, and following the terminology of last section, we want to use models such that its base B has F-symmetry, being F the group generated by the $A \iff G$ and $C \iff T$ permutations. Therefore, from now on **we will forget the lexicographical order and adopt the more convenient one** $\{A, G, C, T\}$, which makes this symmetry more evident. Therefore the rows and columns of our rate matrices Q will be indexed by the DNA bases in order A, G, C, T. As an example of how this convention makes things easier, we can see how the group of permutations F becomes the easily understandable

$$F = \left\{ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \right\}.$$

If we respect this symmetry and have enough number of parameters, it makes sense to talk about transition and transversion rates and, more importantly, they become independent of each other, making it possible that both achieve their optimal (expectedly different) value. For example, between the Lie Markov models with $\mathcal{G}_4$ symmetry we studied in last section, it is clear that all "are" F-symmetric (actually not, since we were forcing the symmwtry to be maximal, but this is actually ignored in the paper). However, the Jukes-Cantor is too poor to make tranversion and transition rates independent, while the GMM is rich enough (and too much, actually). From now on, we will omit the group F and refer to the RY-symmetry.

In the article [5], every and each of the 37 Lie Markov models with RY-symmetry is clearly stated and organized. One of the nicest features of this exposition is that they state a base of the general Markov model such that every RY-Lie Markov model has a (not stochastic) base which is a subset of this one. Note that we are not saying that these basis are stochastic as explained in definition 2.21, because actually they are not. This problem is successfully treated in the article and is not necessary for our work, hence we will not explain it here.

The mentioned base of the GMM is composed by 12 matrices of Figure 3.1.

$$A=\begin{pmatrix}-3&+1&+1&+1\\+1&-3&+1&+1\\+1&+1&-3&+1\\+1&+1&+1&-3\end{pmatrix}\ A_1=\begin{pmatrix}-1&+1&0&0\\+1&-1&0&0\\0&0&-1&+1\\0&0&+1&-1\end{pmatrix}\ C=\begin{pmatrix}0&0&+1&-1\\0&0&-1&+1\\-1&+1&0&0\\+1&-1&0&0\end{pmatrix}$$

$$B=\begin{pmatrix}0&0&+1&-1\\0&0&-1&+1\\+1&-1&0&0\\-1&+1&0&0\end{pmatrix}\ D_1=\begin{pmatrix}-1&+1&0&0\\+1&-1&0&0\\0&0&+1&-1\\0&0&-1&+1\end{pmatrix}\ D=\begin{pmatrix}+1&+1&+1&+1\\+1&+1&+1&+1\\-1&-1&-1&-1\\-1&-1&-1&-1\end{pmatrix}$$

$$E_1=\begin{pmatrix}+1&+1&+1&+1\\-1&-1&-1&-1\\0&0&0&0\\0&0&0&0\end{pmatrix}\ F_1=\begin{pmatrix}+1&+1&-1&-1\\-1&-1&+1&+1\\0&0&0&0\\0&0&0&0\end{pmatrix}\ G_1=\begin{pmatrix}+1&-1&0&0\\+1&-1&0&0\\-1&+1&0&0\\-1&+1&0&0\end{pmatrix}$$

$$E_2=\begin{pmatrix}0&0&0&0\\0&0&0&0\\+1&+1&+1&+1\\-1&-1&-1&-1\end{pmatrix}\ F_2=\begin{pmatrix}0&0&0&0\\0&0&0&0\\+1&+1&-1&-1\\-1&-1&+1&+1\end{pmatrix}\ G_2=\begin{pmatrix}0&0&+1&-1\\0&0&+1&-1\\0&0&-1&+1\\0&0&-1&+1\end{pmatrix}$$

Figure 3.1: The base of the GMM, compounded by 12 matrices.

In some cases it is more convenient to use the matrix $A_2 = 3A_1 - A$ instead of A_1. This matrix A_2 has the form

$$A_2 = \begin{pmatrix}0&+2&-1&-1\\+2&0&-1&-1\\-1&-1&0&2\\-1&-1&+2&0\end{pmatrix}.$$

Once we have this basis of the GMM, stating an RY-Lie Markov model is nothing but selecting an adequate (not all generate one) subset of this basis. This is systematically done in a table given in the article [5], which we copy in Figure 3.2.

In this figure, the name of each model consists on the dimension of the model, followed by the number of parameters necessary to write the matrix as an stochastic matrix, followed by a letter which serves as distinguishable mark. For example, attending to its name, model $RY5.6b$ has a basis formed by 5 elements (namely $A, A_1, D, E_1, E_2,$) and 6 parameters are needed to

Name	Basis matrices	Name	Basis matrices
1.1	A	6.6	A, A_1, B, C, D, D_1
2.2b	A, A_1	6.7a	A, A_1, B, D, E_1, E_2
3.3a	A, A_1, B	6.7b	A, A_1, C, D, E_1, E_2
3.3b	A, A_1, C	6.8a	A, A_1, D, D_1, E_1, E_2
3.3c	A, A_1, D_1	6.8b	A, A_1, D, D_1, G_1, G_2
3.4	A, A_1, D	6.17a	A, A_1, B, D, G_1, G_2
4.4a	A, D, E_1, E_2	6.17b	A, A_1, C, D, G_1, G_2
4.4b	A, A_1, D, D_1	8.8	$A, A_1, D, D_1, E_1, E_2, F_1, F_2$
4.5a	A, A_1, B, D	8.10a	$A, A_1, B, C, D, D_1, E_1, E_2$
4.5b	A, A_1, C, D	8.10b	$A, A_1, B, C, D, D_1, G_1, G_2$
5.6a	A, A_1, B, C, D_1	8.16	$A, A_1, D, D_1, E_1, E_2, G_1, G_2$
5.6b	A, A_1, D, E_1, E_2	8.17	$A, A_1, B, D, E_1, E_2, G_1, G_2$
5.7a	A, A_1, B, E_1, E_2	8.18	$A, A_1, B, D, E_1, E_2, F_1, F_2$
5.7b	A, A_1, B, F_1, F_2	9.20a	$A, A_1, B, C, D_1, E_1, E_2, F_1, F_2$
5.7c	A, A_1, B, G_1, G_2	9.20b	$A, A_1, B, C, D_1, F_1, F_2, G_1, G_2$
5.11a	A, A_1, D_1, E_1, E_2	10.12	$A, A_1, B, C, D, D_1, E_1, E_2, F_1, F_2$
5.11b	A, A_1, D_1, F_1, F_2	10.34	$A, A_1, B, C, D, D_1, E_1, E_2, G_1, G_2$
5.11c	A, A_1, D_1, G_1, G_2	12.12	$A, A_1, B, C, D, D_1,$
5.16	$A, A_1, D, G_1 G_2$		$E_1, E_2, F_1, F_2, G_1, G_2$

Figure 3.2: The RY Lie Markov models.

write its associated matrix $Q_{5.6b}$ as an stochastic matrix. The b indicates the existence of a previous $RY5.6$ model, whose associated letter is a.

This is everything we needed in order to carry out our task.

3.2 How to improve the performance of RY-Lie Markov models

Given a model and a set of species, in order to find the phylogeny with the maximum likelihood, IQ-TREE tries to optimize three kinds of objects: the topology of the tree, the edge lengths and the rate-matrix parameters of the model. Our work consisted in improving the execution time when optimizing the last two of them.

The part of IQ-TREE which interests us can be summarized as follows: We generate a set C of 98 (or some other required number) parsimony trees attending to the given sequences. For each of these trees, we treat the parameters of the rate matrices Q_i as variables, as well as the edge lengths

t_i. We compute the associated transition matrix, i.e. $M_i(t_i) = e^{Q_i t_i}$, then find the values for Q_i and t_i which maximize the likelihood.
We choose a random tree T from C and perturb it using which is called a stochastic NNI. For this new tree T^*, the likelihood is computed. If $l(T^*)$ improves the lowest likelihood in C, then T^* replaces the tree with this likelihood in C. If $l(T^*)$ improves the best likelihood in C, we re-optimize T^*: again we use the parameters of the rate matrices Q_i as variables, as well as the edge lengths t_i, and proceed as with the parsimony trees, i.e. we compute $M_i(t_i) = e^{Q_i t_i}$, then find the values for Q_i and t_i which maximize the likelihood. These values completely determine a new tree T^*. This tree replaces the worst tree in C (i.e. the one with lowest likelihood).
The process is repeated until better trees have not appeared for a while.

The part of this algorithm which we have dealt with is the one of computing $M(t) = e^{Qt}$. There are multiple methods to carry out this computation, among which IQ-TREE uses normally the following two, both approximative:

- Since for fixed t and n big enough we have

$$e^{Q\frac{t}{n}} \simeq \mathrm{Id} + Q\frac{t}{n} + \frac{Q^2}{2!}(\frac{t}{n})^2,$$

  we choose a proper n and do

$$e^{Qt} = (e^{Q\frac{t}{n}})^n \simeq \left(\mathrm{Id} + Q\frac{t}{n} + \frac{Q^2}{2!}(\frac{t}{n})^2\right)^n.$$

  Moreover it is chosen $n = 2^m$, making it possible to compute the n-power with $m-1$ multiplications of matrices. Adding all of them up, we will only do $\log_2(n)$ multiplication of matrices, although we may have precision problems.

- For the given matrix Q, using the EIGEN3 library we numerically compute its Jordan matrix form J (in most of cases, the matrix whose diagonal are the eigenvalues of Q) and its Jordan base S (in most of cases, the columns of S are the eigenvectors of Q.) Then, using proposition 4.10, we have

$$e^{Qt} = Se^{Jt}S^{-1},$$

  and the computation of e^{Jt} is explained in the proposition. Besides the cost of numerically computing the matrices S and J, the computing time will be mainly used for the inverse of S and two multiplications of matrices (the exponential of J should not be too expensive). Although this approach is not too slow, it has the disadvantages of its precision and the reliability of the algorithms for properly computing J and S (in many cases, a non diagonalizable Q poses a problem).

Our aim was to make a more precise, faster and more reliable version of the second of the approaches. The idea of the improvement is straightforward and can be stated in two sequences:

Given an RY-Lie Markov model, we aim to compute e^{Qt} using the Jordan form of Q. If we efficiently write the analytical formulas of J, S and S^{-1}, there is no loss of precision and the executing time may substantially decrease.

The first difficulty of this method which one comes up with is its implementation: there are 37 models, with many parameters, and we aim to give a closed formula (in the friendly case) for their eigenvalues J , eigenvectors S and the inverse of this matrix, S^{-1}, which sums up dozens of kernels which need to be calculated. To overcome this problem we have helped ourselves with the software Mathematica. In our Notebook, we proceeded as follows:

1. Declare the base of Figure 3.1. We have used A_2 instead of A_1. Moreover, we have permutated the elements of the matrices so the natural order of IQ-TREE, $\{A, C, G, T\}$, was respected.
2. Declare each of the models. For example, we would write

   $$Q_{3.4} = aA + a_2A_2 + dD,$$

   choosing always the lower case for the parameter associated to the uppercase base.
3. Compute J the eigenvalues of Q (0 is always one of them).
4. Compute S, the eigenvectors of Q ($(1, \cdots, 1)^T$ is always one of them).
5. Compute S^{-1}. Look for rows and columns with sum 1.

Once we had done this with each of the models, we observed the following:

- With the exception of the models 9.20b and 12.12 (which no not have closed formulas), all of them have the following eigenvalues:

  $$\mathbf{J} = \{0,\ -4(a - a_2),\ 2(-2a - a_2 \pm \sqrt{b^2 - c^2 + d_1})\},$$

  where non used parameters of the models must be taken equal 0. For example, in model 1.1, we have to take $a_2 = 0$, and so on. Attending to the nesting between models, one simply has to compute the eigenvalues of models 10.12 and 10.34 to check this formula.

- When we have consider models with dimension 8 or more, eigenvectors become too complicated (many operations involved). The most interesting case is model 8.18: it is not especially difficult and we know, from [5], that it gives good results when modeling. However, there are denominators depending on the parameters in nearly every formula.
- When Mathematica gives a tractable closed formula for S, it also gives one for S^{-1}. It can happen that one of the elements of a vector is complicated (i.e. it involves many sums and products). In those cases, we try to take advantage on the rows and columns whose sum is 1. Denominators depending on the parameters appear in nearly every formula.

The importance of denominators is the following: on the one side, if they are zero our computations will break; on the orher side, whenever they, in the formulas of both S and S^{-1}, are not zero, we have the guarantee that these two matrix are well defined, i.e. we do not have any singularities. This is a nice feature: we can check whether the denominators are zero or not, and if they are, simply call another exponentiating method, such as the one approximating the power series.

For example, for the model 3.4, with rate matrix $Q_{3.4} = aA + a_2A_2 + dD$, these formulas would look as stated below. We write $\delta = \frac{d}{a-a2}$:

$$\mathbf{J}_{3.4} = \{0,\ -4(a-a_2),\ -2(2a_2+a_2),\ -2(2a+a_2)\}$$

$$S_{3.4} = \begin{pmatrix} 1 & 1-\frac{2d}{a-a_2+d} & 0 & -1 \\ 1 & 1 & -1 & 0 \\ 1 & -1+\frac{2d}{a-a_2+d} & 0 & 1 \\ 1 & -1 & 1 & 0 \end{pmatrix}$$

$$(S_{3.4})^{-1} = \frac{1}{4}\begin{pmatrix} 1+\delta & 1-\delta & 1+\delta & 1-\delta \\ -1-\delta & 1+\delta & -1-\delta & 1+\delta \\ 0 & -2 & 0 & 2 \\ -2 & 0 & 2 & 0 \end{pmatrix}.$$

In this case, the conditions which guarantee that both matrices are well defined and our analytical method can be applied are the following:

$$a - a_2 + d \neq 0;\ \ a - a2 \neq 0.$$

3.3 Results

We have found and programmed these formulas for every model which has between 1 and 5 parameters, with the exception of 5.6a, because we considered its formulas where too complicated to be efficiently computed. In total, these are 18 models.
In order to measure whether a decrease of execution time was achieved or not, we chose five of the models and tested them against the EIGEN3 library. The chosen models where 1.1, 2.2b, 3.3a, 4.4b, 5.11c, while only one input was used, the file example.phy given by IQ-TREE. In order to avoid too much noise due to the machine, we ran the program 5 times for each of the models, find the average of their executing time (exactly, the CPU time) and divide this number by the executing time using the EIGEN3 library (which we also ran 5 times). We would like this ratio to be as close to 0 as possible. The results are given in the following table:

Table 2

	Ratio analytical / EIGEN3 library
1.1	0.9848
2.2b	1.0137
3.3a	0.9807
2.2b	0.9805
1.1	1.0321

Therefore it is easy to infer that not a great difference, regarding the executing time, has been made (never more than 3 per cent). This may not seem intuitive, but actually we ignore the functioning of the EIGEN3 library, which could be very efficient. In any case, this result would be good enough to support our method, for there is no precision lost along the procedure. Moreover, we were able to give simple conditions in the parameters to distinguish the cases in which the Jordan form of the rate matrix Q may poses any computation problem.

Chapter 4

Mathematical tools

This chapter contains all the objects and results which are employed in much broader contexts than phylogenetic reconstruction.

4.1 Exponential of a matrix

Before introducing the definition of exponential of a matrix, it is pertinent to clarify some of the elements which underlie our discussion.

Given a sequence $(a_k)_{k\in\mathbb{N}}$ where $a_k \in \mathbb{C}$, it is well known that its convergence to $L \in \mathbb{C}$ means that it satisfies the $\epsilon - N$ property, i.e. that for any $\epsilon > 0$ we are given, we can find an N such that $|a_k - L| < \epsilon$ for any $k \geq N$. Such an L exists iff (a_k) is a Cauchy sequence.

When dealing with a set partial of sums $S_m = \sum_{k=0}^{k=m} a_k$, defined for all $m \in \mathbb{N}$, the same results apply to the sequence $(S_m)_{m\in N}$. Provided the limit of this sequence exists, we can define

$$\sum_{k=0}^{\infty} a_k := \lim_{m\to\infty} S_m.$$

Generalizing these definitions to sequences of matrices is very easy. Given a sequence $(A_k)_{k\in\mathbb{N}}$ where $A_k = (a_k^{i,j}) \in M_n(\mathbb{C})$, $i, j \in [n]$, we say this sequence converges to a matrix $A = (a^{i,j})$ if each of the n^2 sequences $a_k^{i,j}$ converges to the respective element $a^{i,j}$.
However, this definition is a bit uncomfortable and one prefers to avoid considering the entries of the matrices. To solve this, there is an equivalent

definition: given a matrix norm $|\cdot|$, the sequence $(A_k)_{k\in\mathbb{N}}$ converges to A iff it satisfies the $\epsilon - N$ property as we stated it before.

Regarding partial sums of matrices, we proceed analogously as before and state, for the converging sequence of well-defined partial sums $S_m = \sum_{k=0}^{k=m} A_k$, the following definition

$$\sum_{k=0}^{\infty} A_k := \lim_{m\to\infty} S_m.$$

After this brief clarification, we can state the promised definition:

Definition 4.1. *Given a matrix $A \in M_n(\mathbb{C})$, the exponential of A, e^A, is defined as*

$$e^A := \mathrm{Id} + \frac{A}{1!} + \frac{A^2}{2!}\cdots = \sum_{k=0}^{\infty} \frac{A^k}{k!}.$$

Note that we have set $A^0 := \mathrm{Id}$, including the case $A = 0$.

The exponential of a matrix arises naturally when dealing with linear systems of differential equations, such as

$$\mathbf{x}' = A\mathbf{x},$$

where $A \in M_n(\mathbb{C})$ and $\mathbf{x}(t)$ is a column vector whose elements are unknown derivable functions of t. In this case,the matrix e^{At} comes up. Let us include all these important characteristics, with others, in a complete proposition:

Proposition 4.2. *Given a matrix $A \in M_n(\mathbb{C})$, let us define the function*

$$\begin{aligned} e^{At}\colon \mathbb{R} &\to M_n(\mathbb{C}) \\ t &\mapsto e^{At}. \end{aligned}$$

The function e^{At} satisfies the following properties:

1. *e^{At} is absolutely convergent and uniformly convergent on compact sets in $\mathbb{R}$.*
2. *$D_t(e^{At}) = Ae^{At}$.*
3. *Given the initial value problem $\boldsymbol{x}' = A\boldsymbol{x}$, $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$, its sole solution is $\boldsymbol{x} = e^{(t-t_0)A}\boldsymbol{x}_0$.*
4. *$e^{A(t+s)} = e^{At}e^{As}$.*
5. *If $AB = BA$, then $e^{(A+B)t} = e^{At}e^{Bt} = e^{Bt}e^{At}$.*

6. $\det(e^{At}) = e^{(\operatorname{Tr} A)t}$

7. *The matrix* e^{At} *is invertible for any* $t \in \mathbb{R}$, *and* $(e^{At})^{-1} = e^{-At}$.

8. *If* S *has an inverse and* $A = SJS^{-1}$, *then* $e^{At} = Se^{Jt}S^{-1}$.

Proof. 1. Let $||\cdot||$ be a sub-multiplicative matrix norm. Then

$$\sum_{k=0}^{\infty} ||\frac{A^k t^k}{k!}|| \leq \sum_{k=0}^{\infty} \frac{||A||^k t^k}{k!} = e^{||A||^k t^k} < \infty, \ \forall t \in \mathbb{R}.$$

Therefore we can apply the Weierstrass M-test and conclude that the series $\sum_{k=1}^{\infty} \frac{A^k}{k!}$ is absolutely convergent in $\mathbb{R}$ and uniformly convergent on compact sets in $\mathbb{R}$.

2. A power series which is uniformly convergent on compact sets in $\mathbb{R}$, such as the one we are dealing with, is analytical and can be derived term by term. Therefore

$$D_t(e^{At}) = \sum_{k=1}^{\infty} \frac{A^k t^{k-1}}{(k-1)!} = A\sum_{k=1}^{\infty} \frac{A^{k-1} t^{k-1}}{(k-1)!} = A\sum_{k=0}^{\infty} \frac{A^k t^k}{k!} = Ae^{At}.$$

3. This is a consequence of Picard's theorem for the uniqueness and existence of first order ODEs. One only has to prove that indeed $\mathbf{x} = e^{(t-t_0)A}\mathbf{x}_0$ is a solution.
On one side, $(e^{(t-t_0)A}\mathbf{x}_0)(t_0) = \mathrm{x}_0$. On the other side, $D_t(e^{(t-t_0)A}\mathbf{x}_0) = Ae^{(t-t_0)A}\mathbf{x}_0$. We conclude $\mathbf{x} = e^{(t-t_0)A}\mathbf{x}_0$ is the only solution of the mentioned initial value problem.

4. Let us fix $s \in \mathbb{R}$ and consider the functions $X_1(t) = e^{A(t+s)}$ and $X_2(t) = e^{At}e^{As}$. We want to prove these two functions are equal. We have:

$$X_1' = Ae^{A(t+s)} = AX_1, \ \ X_2' = Ae^{At}e^{As} = AX_2, \ \ X_1(0) = e^{As} = X_2(0).$$

Therefore $X_1(t)$ and $X_2(t)$ are solutions of the same initial value problem, hence using the uniqueness of solutions we conclude

$$X_1(t) = X_2(t) \ \ \forall s \in \mathbb{R},$$

as we wanted to prove.

5. Let us consider the functions $Y_1(t) = e^{(A+B)t}$ and $Y_2(t) = e^{At} \cdot e^{B}$. We have:

$$Y_1' = (A+B)e^{(A+B)t} = (A+B)Y_1, \ \ Y_2' = Ae^{At}e^{Bt} + e^{At}Be^{Bt} = (A+B)Y_2,$$

where one must note that for last equation we used the fact that e^{At} and B commute since A and B commute. Moreover,
$Y_1(0) = \mathrm{Id} = Y_2(0)$. All in all, $Y_1(t)$ and $Y_2(t)$ solve the same initial value problem, hence they are identical, as we wanted to prove.

6. It is a consequence of *Liouville's formula.*

7. Last result implies that the determinant of e^{At} is not zero for any $t \in \mathbb{R}$, hence its inverse exists. Moreover,

$$e^{At}e^{(-A)t} = e^{(A+(-A))t} = e^{0t} = \mathrm{Id}.$$

8. Since $(SJS^{-1})^k = SJ^kS^{-1}$, we only have to apply the definition of exponential:

$$e^{At} = \sum_{k=0}^{\infty} \frac{A^k t^k}{k!} = \sum_{k=0}^{\infty} \frac{(SJS^{-1})^k t^k}{k!} = \sum_{k=0}^{\infty} \frac{SJ^kS^{-1}t^k}{k!} = S\sum_{k=0}^{\infty} \frac{J^k t^k}{k!} S^{-1} = Se^{Jt}S^{-1},$$

as desired.

Corollary 4.3. *Given the matrix initial value problem*

$$M'(t) = AM(t), \;\; M(t_0) = M_0,$$

where $M(t) = (a^{i,j}(t))^{i,j\in[n]}$ is a differentiable matrix function, its sole solution is

$$M(t) = e^{(t-t_0)A}M_0.$$

Proof. It is a consequence of item 3 in last proposition. We can write

$$M(t) = \Big(\mathbf{x}_1(t), \cdots \mathbf{x}_n(t)\Big),$$

where each $\mathbf{x}_i(t)^T = (a_{i1}(t), \cdots, a_{in})$ (i.e. $x_i(t)$ equals the i column of $M(t)$), and solving the matrix initial value problem equals solving n vectorial initial value problems simultaneously.

□

We have seen how useful the exponential of a matrix is. Item 8 in proposition 4.2 suggests it is necessary to expose one of the most convenient methods to calculate it, the diagonalization of a matrix and its generalization, the Jordan canonical form of a matrix.

Definition 4.4. *Given a scalar $\lambda \in \mathbb{C}$ and a natural number $r \in \mathbb{N}$, a* Jordan block $J_r(\lambda)$ *denotes the $r \times r$-matrix whose entries are all zero except from the subdiagonal entries, equal λ, and the diagonal entries, equal* 1*. Therefore:*

$$J_1(\lambda) = \begin{pmatrix}\lambda\end{pmatrix}, \quad J_2(\lambda) = \begin{pmatrix}\lambda & 0\\ 1 & \lambda\end{pmatrix}, \quad J_3(\lambda) = \begin{pmatrix}\lambda & 0 & 0\\ 1 & \lambda & 0\\ 0 & 1 & \lambda\end{pmatrix}$$

and so on.

Definition 4.5. *A block diagonal matrix whose blocks are Jordan blocks is called a* Jordan matrix*.*

Now we can state a very important and well-known theorem, hence we will omit its proof in order to avoid a long digression:

Theorem 4.6. (Jordan canonical form) *Given any square matrix $A \in M_n(\mathbb{C})$, there exists a Jordan matrix*

$$J = \operatorname{diag}(J_1, \cdots, J_l) \in M_n(\mathbb{C})$$

whose Jordan blocks are $J_k = J_r(\lambda_k)$, $1 \leq k \leq l$, and an invertible matrix $S \in M_n(\mathbb{C})$ such that

$$A = SJS^{-1}.$$

The Jordan form J of the matrix A is unique if we do not take into account Jordan matrices whose blocks are permutations of these ones.

Definition 4.7. *The columns of S, let us say $S = \left(v_1, \cdots v_n\right)$, form a* Jordan base *of the matrix A.*

In the previous theorem, when all Jordan blocks have dimension 1×1, J is the diagonal form of A, and the vectors $v_1, \cdots v_n$ compound the base of eigenvectors of A, each of them with eigenvalue equal to the respective diagonal entry. Since we do not dispose of the Jordan form when we are given the matrix A, the method we employ to distinguish whether the matrix A is diagonalizable or not is the comparison of *algebraic multiplicity* and *geometric multiplicity*for every eigenvalue λ_i.

Definition 4.8. *Given a matrix A whose characteristic polynomial is $p_A(x) = \det(A - \lambda \mathrm{Id})$, and such that λ_i is an eigenvalue of A (hence $p_A(\lambda_i) = 0$), the* algebraic multiplicity of λ_i*, denoted as $r(\lambda_i)$, is the biggest natural number r such that $(x - \lambda_i)^r \Big| p_A(x)$.*

Definition 4.9. *Given a matrix A and one of its eigenvalues λ_i, the* geometric multiplicity of λ_i, *which we will denote as $g(\lambda_i)$, is the dimension of the kernel of $A - \lambda_i \mathrm{Id}$, i.e.*

$$g(\lambda_i) = \dim \ker(A - \lambda_i \mathrm{Id}).$$

It is easy to prove that, for any eigenvalue λ_i, $1 \leq g(\lambda_i) \leq r(\lambda_i)$. It is only possible to diagonalize a matrix if, for every eigenvalue λ_i, we have $g(\lambda_i) = r(\lambda_i)$, which is the test we aimed to find.

The following proposition summarizes the technique one uses when aiming to compute the exponential e^{At}:

Proposition 4.10. *Let $J = \mathrm{diag}(J_1, \cdots, J_l) \in M_n(\mathbb{C})$, where $J_k = J_r(\lambda_k)$, $1 \leq k \leq l$, be the Jordan canonical form of $A \in M_n(\mathbb{C})$ and S an invertible matrix such that $A = SJS^{-1}$. Then:*

1. $e^{At} = Se^{Jt}S^{-1}$
2. $e^{Jt} = \mathrm{diag}\left(e^{J_1 t}, \cdots, e^{J_l t}\right)$
3. *Given a Jordan block $J_r(\lambda)$, it holds that*

$$e^{J_r(\lambda)t} = \exp\left(J_r(\lambda)t\right) = \begin{pmatrix} 1 & & & & \\ t & 1 & & & \\ \frac{t^2}{2} & t & 1 & & \\ \vdots & \ddots & \ddots & \ddots & \\ \frac{t^{r-1}}{(r-1)!} & \cdots & \frac{t^2}{2} & t & 1 \end{pmatrix} e^{\lambda t}$$

Proof. 1. It has been already proved in proposition 4.2.

2. It can be easily seen by computing the multiplication of J by itself:

$$J^2 = \mathrm{diag}(J_1^2, \cdots, J_l^2),$$

and so on if one continues multiplying. Therefore the infinite sum of e^{Jt} yields the mentioned result.

3. Let us introduce some notation so this proof is easier to write and read. Given the dimension $r \in \mathbb{N}$, we will denote by $\mathrm{diag}_r^i(\lambda)$ the $r \times r$-matrix whose entries are all zeros except for the i-th subdiagonal, whose entries are all λ. If $i \geq r$, this is the null matrix. For example, $\mathrm{Id} = \mathrm{diag}_r^0(1)$, and more interestingly for our purposes, $J_r(\lambda) = \lambda \mathrm{Id} + \mathrm{diag}_r^1(1)$. It is very easy to prove that, for any $\alpha \in \mathbb{C}$,

$$(\mathrm{diag}_r^1(\alpha))^k = \mathrm{diag}_r^k(\alpha^k),$$

which implies that

$$(J_r(\lambda))^k = \Big(\lambda \mathrm{Id} + \mathrm{diag}_r^1(1)\Big)^k = \sum_{l=0}^{k} \binom{k}{l} \lambda^{k-l}\, \mathrm{diag}_r^l(\lambda^l),$$

hence, if we write $J := J_r(\lambda)$, we have

$$\begin{aligned} e^{Jt} =& \sum_{k=0}^{\infty} \frac{J^k}{k!} t^k = \sum_{k=0}^{\infty} \frac{\sum_{l=0}^{k} \binom{k}{l} \lambda^{k-l}\, \mathrm{diag}_r^l(\lambda^l)}{k!} t^k \\ =& \sum_{l=0}^{\infty} \sum_{k=l}^{\infty} \frac{\lambda^{k-l}\, \mathrm{diag}_r^l(\lambda^l)}{(k-l)! \cdot l!} t^k \\ =& \sum_{l=0}^{\infty} \frac{\mathrm{diag}_r^l(\lambda^l)}{l!} t^l \sum_{k=l}^{\infty} \frac{(\lambda t)^{k-l}}{(k-l)!} \\ =& \sum_{l=0}^{\infty} \frac{\mathrm{diag}_r^l(\lambda^l)}{l!} t^l e^{\lambda t} \end{aligned}$$

Since, by definition, $\mathrm{diag}_l(\alpha) := 0$ when $l \geq r$, we infer that

$$e^{J_r(\lambda)t} = e^{\lambda t} \sum_{l=0}^{r-1} \mathrm{diag}_r^l\Big(\frac{\lambda^l t^l}{l!}\Big),$$

which is exactly what we wanted to prove.

□

When we are dealing with real or complex numbers, one of the most useful properties of the exponential function is the fact that, for let us say $x, y \in \mathbb{C}$,

$$e^{x+y} = e^x e^y.$$

Unfortunately, such an elegant formula does not exist for the exponential of a matrix. However, it is possible to generalize it, which is known as the Baker-Campbell-Hausdorff (BCH) formula:

$$e^X e^Y = \exp\{X + Y + \frac{1}{2}[X,Y] + \frac{1}{12}[X,[X,Y]] + \cdots\}, \tag{4.1}$$

where $[X,Y] := XY - YX$ is called the commutator of X and Y, and the rest of terms in the sum are also commutators of commutators of X and Y. The convergence of this sum is not a trivial matter, although we will take it for granted since this would distract us from our topic (for a detail description, see [3]). As we do in other parts of this work, the BCH formula has interesting applications when one aims to decide whether a given set of exponential of matrices is a closed group under multiplications or not.

4.2 Elementary Lie theory

We will introduce the elementary vocabulary used in Lie theory, as mentioned in section 2 of [4]. These are very general tools which we are using with well-behaving objects, hence apparently difficult definitions will end up being friendly.

Definition 4.11. *A* matrix group $\mathcal{G}$ *is a topologically closed subgroup of* $GL(n,\mathbb{C})$*. In other words, if a sequence of elements of* $\mathcal{G}$ *converges to an invertible matrix* M*, then* $M \in \mathcal{G}$*.*

Definition 4.12. *If we regard* $\mathcal{G}$ *as a manifold, the* tangent space *of* $\mathcal{G}$ *at a matrix* M*,* $T_M(\mathcal{G})$*, is the set of all* $X \in M(n,\mathbb{C})$ *such that there exists a smooth path* $A : [0,1] \mapsto \mathcal{G}$ *satisfying*

$$A(0) = M \text{ and } A'(0) := \frac{dA(t)}{dt}\Big|_{t=0} = X.$$

Definition 4.13. *The Lie algebra of a matrix group* $\mathcal{G}$ *is* $T_{\mathbf{1}}(\mathcal{G})$*, i.e. the tangent space of* $\mathcal{G}$ *at the identity.*

The previous definition has a reciprocal one, which is the following:

Definition 4.14. *A closed set* $\mathcal{L} \subset M_n(\mathbb{C})$ *is a* Lie algebra *if*

$$t_1X + t_2Y \in \mathcal{L} \text{ and } [X,Y] \in \mathcal{L},$$

for all $X, Y \in \mathcal{L}$ *and* $t_1, t_2 \in \mathbb{R}$*.*

It is important to note that our definition of Lie algebra makes it a $\mathbb{R}$-vector space, not a complex one. The two last definitions seem to use ambiguous or maybe contradictory terminology, hence it is convenient to prove that the reciprocity of these definitions works well, which is our objective in following proposition:

Proposition 4.15. *The Lie algebra of a matrix group* $\mathcal{G}$ *is a Lie algebra. Reciprocally, for every Lie algebra* $\mathcal{L}$*, there exists a matrix group* $\mathcal{G}$ *such that* $\mathcal{L}$ *is the Lie algebra of* $\mathcal{G}$*.*

Proof. As for the first part of the statement, firstly we should prove that $\mathcal{G}$ is a vector space, i.e. that for any two matrices $X, Y \in T_{\mathbf{1}}(\mathcal{G})$ and $\alpha \in \mathbb{R}$, the matrix $X + \alpha Y$ belongs to $T_{\mathbf{1}}(\mathcal{G})$. We take the paths A and B, respectively associated to X and Y. Let us consider the path

$$C(t) := A(t)B(\alpha t).$$

Obviously, $C(t) \in \mathcal{G}$, and we have that

$$C(0) = A(0)B(0) = \mathbf{1} \text{ and } C'(t)|_0 = [A'B + \alpha AB']_{t=0} = X + \alpha Y,$$

as desired. Note that this implies that the Lie algebra of a matrix group is a closed set. Now we should prove that, for any $X, Y \in \mathcal{G}$, the commutator $[X, Y]$ belongs to $\mathcal{G}$. Let us take again the paths A and B, associated to X and Y, and consider the smooth function

$$f(s) := A(s)B'(0)A(s)^{-1},$$

which satisfies $f(s) \in T_{\mathbf{1}}(\mathcal{G})$ for any $s \in [0, 1]$. Indeed, for any fixed $s \in [0, 1]$, we can consider the well defined path in $\mathcal{G}$:

$$D(t) := A(s)B(t)A(s)^{-1},$$

which fulfills the conditions $D(0) = \mathbf{1}$ and $D'(0) = f(s)$.
Hence we have a function $f(s)$ whose image is contained in $T_{\mathbf{1}}(\mathcal{G})$. Therefore, for $s \in (0, 1)$ also the image of the smooth function

$$g(s) := \frac{f(s) - f(0)}{s}$$

is contained in $T_{\mathbf{1}}(\mathcal{G})$. Since f is smooth, the limit $\lim_{s\to 0+} g(s)$ must exist and be equal to $f'(0)$, and using the fact that $T_{\mathbf{1}}(\mathcal{G})$ is a closed set, the existing limit of elements in $T_{\mathbf{1}}(\mathcal{G})$ must be contained in $T_{\mathbf{1}}(\mathcal{G})$, thus $f'(0) \in T_{\mathbf{1}}(\mathcal{G})$. It only remains to compute $f'(0)$. Regarding $f'(s)$, we have

$$f'(s) = A'(s)YA(s)^{-1} - A(s)YA(s)^{-1}A'(s)A(s)^{-1}.$$

Substituting $s = 0$, we arrive to the desired $f'(0) = XY - YX$, which was our purpose and finishes this part of the demonstration.

Regarding the second part of the proof, for a given Lie algebra $\mathcal{L}$, let us consider the set

$$\mathcal{G} := e^{\mathcal{L}} = \{e^Q : Q \in \mathcal{L}\}.$$

Obviously, we aim to prove that $\mathcal{G}$ is the sought matrix group. We need to prove that $T_{\mathbf{1}}(\mathcal{G}) = \mathcal{L}$.
First, let us go for the inclusion $\subset$. Consider $X \in T_{\mathbf{1}}(\mathcal{G})$. By definition, there exists a path $A : [0, 1] \mapsto \mathcal{G}$ such that $A(0) = \mathbf{1}$ and $A'(0) = X$. Using the smoothness of the exponential and the inverse-function theorem, there must exist at least one local inverse of the exponential of a matrix, $\log(\cdot)$, which we set to be defined in a neighborhood of $A(0)$, let us say $H \subset e^{\mathcal{L}}$, choosing the determination making $log : H \mapsto \mathcal{L}$. For ϵ small enough, and $t \in [0, \epsilon]$ we define

$$B(t) := log(A(t)),$$

whose image belongs to $\mathcal{L}$. Therefore for $t \in [0, \epsilon]$, we have

$$A(t) = e^{B(t)},$$

which implies $B(0) = 0$. Moreover, since

$$A'(t) = B'(t)e^{B(t)},$$

we substitute $t = 0$ and arrive to

$$B'(0) = X.$$

But now we can use the same argument we used when proving the first part of the proof: $\mathcal{L}$ is a closed set, and $B(t) \in \mathcal{L}$ for any $t \in [0, \epsilon]$, hence the existing limit when $s \to 0$ of the function $\frac{B(s)-B(0}{s}$ must also be contained in $\mathcal{L}$, which takes us to $B'(0) = X \in \mathcal{L}$, as desired.
Now let us prove the inclusion $\supset$. For any $X \in G = e^{\mathcal{L}}$, let us consider the path from $[0, 1]$ to $\mathcal{G}$ defined as

$$A(t) = e^{Xt}.$$

It is a well-defined, smooth path, which satisfies $A(0) = \mathbf{1}$, and also $A'(t) = Xe^{Xt}$, hence $A'(0) = X$, which is nothing but the definition of X belonging to $T_{\mathbf{1}}(\mathcal{G})$, i.e. the Lie algebra of $\mathcal{G}$. This finishes the proof.

□

Proposition 4.16. *The Lie algebra of the general linear group, $GL(n, \mathbb{C})$ is $M_n(\mathbb{C})$.*

Proof. We have to prove that $T_{\mathbf{1}}(GL(n\mathbb{C})) = M_n(\mathbb{C})$. The inclusion $\subset$ is obvious. We only have to prove the inclusion $\supset$.
Assume we are given any $X \in M_n(\mathbb{C})$. Consider the smooth path $A(t) := e^{Xt}$. It satisfies $A(0) = \mathbf{1}$ and $A'(0) = X$, which is the definition of $X \in T_{\mathbf{1}}(GL(n\mathbb{C}))$

□

Last proposition hides an ambiguity. Attending to definition 4.13, the Lie algebra of a matrix group, $T_{\mathbf{1}}(GL(n\mathbb{C}))$, is a vector space over $\mathbb{R}$, and not over $\mathbb{C}$, while one would usually consider $M_n(\mathbb{C})$ as a complex vector space. This difference is relevant, because the dimension, generators etc. of any subspace will depend on this feature. Let us see this explicitly:
As for $M_n(\mathbb{C})$, considered as a complex vector space, we know that it

is generated by the elementary matrices $\{E_{ij}\}_{i,j\in[n]}$, whose elements are $[E_{ij}]_{kl} = \delta_{ik}\delta_{jl}$. In a briefer sentence,

$$M_n(\mathbb{C}) =< \{E_{ij}\}_{i,j\in[n]} >_{\mathbb{C}} .$$

Regarding $T_{\mathbf{1}}(GL(n\mathbb{C}))$, we have that it is a real vector space which has, for example, the base $\{E_{ij}\}_{i,j\in[n]} \cup \{\mathbf{i}E_{kl}\}_{k,l\in[n]}$, where as usual $\mathbf{i} = \sqrt{-1}$. Summarizing,

$$T_{\mathbf{1}}(GL(n\mathbb{C})) =< \{E_{ij}\}_{i,j\in[n]} \cup \{\mathbf{i}E_{kl}\}_{k,l\in[n]} >_{\mathbb{R}} .$$

Once we noted this fact, since it is not comfortable to work with a space containing complex numbers considering it as a real vector space, and also for the sake of homogeneity, we will use the following tool:

Definition 4.17. *The* complexification *of a (real) Lie algebra $T_{\mathbf{1}}(\mathcal{G})$, designed as $T_{\mathbf{1}}(\mathcal{G})^{\mathbb{C}}$, is the complex vector space $T_{\mathbf{1}}(\mathcal{G})$ spanned by all linear combinations $c_1X_1 + c_2X_2$ with $X_1, X_2 \in T_{\mathbf{1}}(\mathcal{G})$ and $c_1, c_2 \in \mathbb{C}$.*
When $T_{\mathbf{1}}(\mathcal{G}) = T_{\mathbf{1}}(\mathcal{G})^{\mathbb{C}}$ as sets and $T_{\mathbf{1}}(\mathcal{G}) = T_{\mathbf{1}}(\mathcal{G})^{\mathbb{C}} =< X_1, \cdots, X_k >_{\mathbb{C}}$, where $\{X_i\}_{i\in[k]}$ is a set of linearly independent tangent vectors over $\mathbb{C}$, we say that $\{X_i\}_{i\in[k]}$ forms a $\mathbb{C}$-basis of $T_{\mathbf{1}}(\mathcal{G})$.

Regarding last definition, one only has $T_{\mathbf{1}}(\mathcal{G}) = T_{\mathbf{1}}(\mathcal{G})^{\mathbb{C}}$ when $\mathcal{G}$ is a manifold over $\mathbb{C}$. Along this work, when considering the matrix groups we are interested in, that will be the case, hence we will be able to implicitly assume that complexification has been performed. Therefore every Lie algebra we deal with will be considered to be a vector space over $\mathbb{C}$.

4.3 The symmetric group $\mathcal{G}_n$.

For the references of the results of this section, *vid.* [8].
We will frequently need to permutate the elements of a matrix, which is nicely done by using the symmetric group. For example, let us suppose we are given the states $[n]$ and consider the matrix $M = (m_{ij})_{i,j\in[n]}$, such that m_{ij} indicates hte probability of passing from one state to another one. If now we relabel the states, we would like to know how the new matrix M is. We can formulate this problem more formally. Given a permutation $\sigma \in \mathcal{G}_n$, it acts on the set $[n] = \{1, \cdots, n\}$ as

$$\sigma[n] = \{\sigma(1), \cdots, \sigma(n)\}.$$

Now if we want to consider the transition matrix whose possible states are $\sigma[n]$ and not $[n]$, we claim that this matrix is

$$K_\sigma M K_\sigma^{-1},$$

where K_σ, the $n \times n$ permutation matrix representing $\sigma \in \mathcal{G}_n$, is defined as $[K_\sigma] := \delta_{i,\sigma(j)}$. This is easily explained as follows: given a vector v written in the order $(1, \cdots, n)$, the matrix associated to $\sigma[n]$ will act not on v, but on the vector σv. Therefore we will do

$$K_\sigma M K_\sigma^{-1} \sigma v = K_\sigma M v,$$

which will give as output the permutated version of Mv, as desired.

4.3.1 Representation theory

We aim to state some important results of representation theory. Giving a detailed explanation of this theory as we did with the rest of objects in our work would enlarge it too much, hence we will be forced to omit it. The reader should consult the very well explained book "The Symmetric Group", written by Bruce E. Sagan, cited in [8].

Assume we have a multiplicative group G, and also that we have either a $\mathbb{C}$-algebra A or $\mathbb{C}$-vector space V. For the sake of simplicity, we will only refer to the less restrictive V, and our results will obviously apply also for A since algebras are vector spaces. To begin with, we will limit ourselves for the case in which the canonical basis can be indexed by elements of G. For example, if $|G| = n$, we could consider the elements of G, let us say $\{g_1, \cdots, g_n\}$ as the canonical basis. Hence we would have

$$V :=< g_1, \cdots, g_n >_{\mathbb{C}} = \{\alpha_1 g_1 + \cdots + \alpha_n g_n\} \cong \mathbb{C}^n,$$

where one should note that the elements of G are like the coordinate we are moving in, as if we had written $\mathbf{e_i}$ instead. However, they allow us to define the action of an element $g \in G$ on an element $v = \alpha_1 g_1 + \cdots + \alpha_n g_n \in V$ as follows:

$$gv = g \sum_{i=0}^{n} \alpha_i g_i = \sum_{i=0}^{n} \alpha_i (g g_i) \in V.$$

Or more generally, we have defined an action of G on the module V, i.e. a function ρ defined as

$$\begin{aligned} \rho : G \times V &\to V \\ (g, v) &\mapsto gv \end{aligned}$$

which behaves very well, for $g_1(g_2v) = (g_1g_2)v$ and $g(v_1 + v_2) = gv_1 + gv_2$. This ρ is **a** *representation* of G (note that the representation is not only the function, but also the domain where it is defined, hence includes the vector space V.) Therefore, representations and G-modules are two faces of the same coin, and actually these terms are used interchangeably.
Now we should take into account that talking about a group G is nothing but talking about a subgroup of the symmetric group, $\mathcal{G}_n$. Therefore it is convenient to have an example with this specific group.
Let us consider the $n!$-dimensional vector space $V =< \mathcal{G}_n >_{\mathbb{C}}$. We can define the one-dimensional subspace

$$W :=< \sum_{\sigma \in \mathcal{G}_n} \sigma >_{\mathbb{C}},$$

i.e. the multiples of the vector $v = \sum_{\sigma \in \mathcal{G}_n} \sigma$. If we take any permutation $s \in \mathcal{G}_n$, we can see that

$$sv = \sum_{\sigma \in \mathcal{G}_n} s\sigma = \sum_{\sigma \in \mathcal{G}_n} \sigma = v,$$

hence we can infer that the subspace W is invariant under the action of $\mathcal{G}_n$, i.e. W is a submodule of the $G-$module V. On the other side, W is one-dimensional, hence it cannot contain any $\mathcal{G}_n$-submodule. Since $W \cong \mathbb{C}$, actually we can redefine the restriction of ρ to W as :

$$\begin{aligned} \rho : \mathcal{G}_n \times \mathbb{C} &\to \mathbb{C} \\ (g, \alpha) &\mapsto \alpha. \end{aligned}$$

In general, any ρ is determined by the action of G on the basis of V. In last case, the base of $\mathbb{C}$ is $\{1\}$, and we have $\sigma\ 1 = 1$ for any $\sigma \in \mathcal{G}_n$. We will say this is a *trivial* representation.

This is a very good opportunity to introduce more vocabulary:

Definition 4.18. *If a representation $\rho : G \times V \to V$ satisfies that V contains a non-trivial G-submodule, we will say that ρ is a* reducible *representation. Otherwise, we will say it is an* irreducible *representation.*

For example, our representation $\rho : \mathcal{G}_n \times W \to W$ is not only trivial, but also irreducible. Moreover, we know that a module can be decomposed in submoduli, hence we can write $\mathcal{G}_n = W \oplus U$ for some submodule $U \leq V$. We will say that the representation $\rho : \mathcal{G}_n \times V \to V$ has been decomposed as $\rho : \mathcal{G}_n \times W \oplus U \to W \oplus U$, or simply abuse of the notation and say that the representation decomposes as $V = W \oplus U$.

Now let us forget our initial definition $V :=< g_1, \cdots, g_n >_{\mathbb{C}}$ and consider a vector space $V =< e_1, \cdots, e_m >_{\mathbb{C}}$ such that G, with $|G| = n$, acts on V

through the representation ρ. For every $g \in G$, the application of g on the base element e_i determines a vector v_i, i.e. $ge_i = v_i$. We will informally write

$$g \sim M_g = \Big(v_1 | \cdots | v_m\Big),$$

where the matrix M_g has the vectors v_i as columns. Therefore the representation ρ is completely determined by this n matrices of size $m \times m$. Somehow, what we are doing is respectfully introducing the elements of G in $GL(m, \mathbb{C})$ (the inverse is necessary since it is a group). Note that we are not saying nothing about one-to-one relationships, because actually we already saw for the case $m = 1$ that every element of G was going to the identity. What we do respect is the group operator, i.e. $M_{g_1 g_2} = M_{g_1} M_{g_2}$. This is why a representation can also be defined as an homomorphism $\phi : G \to GL(\mathbb{C}, m)$.

In addition, this idea of the elements of G as matrices allow us to define a new tool: the *character function.*

Definition 4.19. *If we write $\phi(g)$ as the matrix associated to g as previously explained, the* character function *is defined as*

$$\begin{aligned} \chi : G &\to \mathbb{C} \\ g &\mapsto \mathrm{Tr}(\phi(g)). \end{aligned}$$

If a character is defined for a submodule W of V, we will write χ^W, and the trace must only be computed for the elements concerning W.

Attending to last definition, if we consider χ^W, then for any $g \notin W$, we will obviously have $\chi^W(g) = 0$. We can even define an scalar operator $< \cdot, \cdot >$, for any two characters of different or equal representations of the group G, as

$$< \chi_1, \chi_2 >= \frac{1}{|G|} \sum_{g \in G} \chi_1(g) \chi_2(g).$$

Before continuing, recall that a *partition* of n is a non increasing ordered set of non-negative integers $\lambda = (\lambda_1, \cdots \lambda_r)$ such that $\sum_{i=1}^{r} \lambda_i = n$. We will write $\lambda = \{\lambda_1^{n_1} \cdots \lambda_s^{n_s}\}$ to denote the partition which has n_i copies of integer λ_i. For example $\lambda = (3, 3, 1) = \{3^2 1\}$ is a partition of 7.

Now we are prepared to state, without demonstrating them, some important results:

Theorem 4.20. (Maschke's theorem) *Given a group G acting on V through the representation ρ, we can write*

$$V = W_1 \oplus \cdots \oplus W_k,$$

where W_i are irreducible representations.

Proposition 4.21. (Cor. 1.9.4 of [8]) *Let us consider the representation V of G, which can be decomposed as*

$$V = m_1 W_1 \oplus \cdots \oplus m_k W_k,$$

where the W_i are irreducible and pairwise inequivalent, and the m_i indicates the number of copies which appear in the decomposition. Moreover, let χ^i be the character of module W_i. Then:

1. $\chi = m_1\chi^1 + \cdots + m_k\chi^k$.
2. $< \chi, \chi^j >= m_j$ *for all* j.
3. $< \chi^i, \chi^j >= 0$ *if* $i \neq j$.
4. $< \chi, \chi >= m_1^2 + \cdots m_k^2$.
5. V *is irreducible iff* $< \chi, \chi >= 1$.

Proposition 4.22. *The irreducible representations of $\mathcal{G}_n$ are in one-to-one correspondence with the partitions of n. Differently explained, for each partition λ of n, there exists essentially only one m_λ-dimensional vector space V^λ and a representation (using the homomorphism definition)*

$$\phi_\lambda : \mathcal{G}_n \to GL(V^\lambda) \cong GL(m_\lambda, \mathbb{C})$$

such that ϕ_λ is irreducible.

Proposition 4.23. *Using the terms of last proposition, given a representation $\phi : \mathcal{G}_n \to GL(V)$, we can decompose it into irreducible representations as*

$$V \cong \oplus_\lambda c_\lambda V^\lambda,$$

where the sum is taken over all partitions λ of n and the c_λ are nonnegative integers. Moreover, if we define the projection operator *of $\mathcal{G}_n$ in the irreducible representation λ as*

$$\Theta_\lambda := \frac{1}{|\mathcal{G}_n|} \sum_{\sigma \in \mathcal{G}_n} \chi^\lambda(\sigma)\sigma,$$

then we have

$$\Theta_\lambda V = c_\lambda V^\lambda.$$

Regarding the projection operator, it will be handy when we want to compute the integers c_λ.

Bibliography

[1] Pachter and Sturmfels, *Algebraic Statistics for Computational Biology* section 4.5.1
Cambridge University,2005.

[2] Meyer, chapter 8 *Matrix analysis and applied linear algebra*,
Siam, 2000

[3] Blanes and Casas *On the convergence and optimization of the BCH formula.*
Linear Algebra Appl., 2004

[4] J.G. Sumner, J. Fernández-Sánchez, P.D. Jarvis *Lie Markov models*
Journal of Theoretical Biology, 2012

[5] J.G. Sumner, J. Fernández-Sánchez, Michael D. Woodhams *A new hierarchy of phylogenetics models consistent with heterogeneous substitution rates*
Systems Biology, 2015

[6] J.G. Sumner, J. Fernández-Sánchez, Michael D. Woodhams et al. *Is the General Time-Reversible model bad for molecular phylogenetics?*
Systems Biology, 2012

[7] Johnson *Markov-type Lie groups in* $GL(n, \mathbb{R})$
Journal of maths and physics, 1985

[8] Sagan *The symmetric group.*
Graduate texts in mathematics, Springer 2001

[9] Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler and Bui Quang Minh *IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.*
Molecular biology and evolution, 2015