<table>
<thead>
<tr>
<th>Title:</th>
<th>Application of algebraic techniques to phylogenetic reconstruction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author:</td>
<td>Laura Cifuentes Fontanals</td>
</tr>
</tbody>
</table>
| Advisors:      | Marta Casanellas Rius  
                 Jesus Fernandez Sanchez                              |
| Department:    | Matemàtica Aplicada I                                                                           |
Application of algebraic techniques to phylogenetic reconstruction

Laura Cifuentes Fontanals

Advisors: Marta Casanellas Rius and Jesus Fernandez Sanchez

Departament de Matemàtica Aplicada I
## Contents

1 Introduction 1

2 Background 3
   2.1 Alignments of DNA sequences 3
   2.2 Phylogenetic trees 4
   2.3 Markov models of nucleotide substitution 6
   2.4 Flattening 10

3 A new phylogenetic reconstruction model 12
   3.1 Fourier coordinates 12
   3.2 Flattening in Fourier coordinates 20
   3.3 Flattening adjustment 21
   3.4 Phylogenetic reconstruction 27

4 Conclusions 29


\section{Introduction}

According to Darwin’s theory of natural selection, all current species have evolved from some previous ones. This evolution can be represented using phylogenetic trees, which are tree diagrams that relate every species with its ancestor. The study of these relations among the different species through the analysis of sequences of biological data such as DNA, RNA or proteins is called phylogenetics. Its main goal is to hypothesize about the shape and distribution of these trees. The reconstruction of these trees is important not only for determining the evolutionary history of the species but also for the study of their genes and physiology.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{phylogenetic_tree.png}
\caption{Example of phylogenetic tree}
\end{figure}

In order to construct these trees, it is usual to look at the DNA molecules in the genome of the species living on the planet nowadays. Although the DNA sequences corresponding to the same gene can be very different in two different species and even be located in different parts of the genome, phylogenetic models work with \textit{alignments}: sets of DNA sequences of different species that are supposed to have evolved from the same DNA sequence at the common ancestor.

Phylogenetic methods try to determine which tree corresponds to a given alignment by using either distances among species or evolutionary models. All these methods have advantages and disadvantages and a new area that uses algebraic statistics is emerging in phylogenetics ([AR07], [PS05]).

In this line, the main objectives of this project are:

- Understanding the problem of phylogenetic reconstruction and the use of Markov models for nucleotide substitution.
• Studying the application of the discrete Fourier transform to evolutionary group-based models.

• Conceiving a new method for phylogenetic reconstruction, using the properties of the Fourier transform.

• Implementing it in c++.

To do so, we are going to focus on trees with four leaves (quartets), which means using information of alignments of four species. Working only with four species is not a wasted effort because there are quartet-based methods that use quartet trees in order to reconstruct larger trees (see [RG01]).

We shall consider the Kimura 3-parameter model of evolution (see [Kim81]). This model was introduced by Kimura in 1981 and it is still widely used due to its interesting properties. Mathematically speaking, this model is well suited for being studied under the discrete Fourier transform.

Using this change of coordinates we are able to propose a new phylogenetic reconstruction method that is independent of the outer branches of the quartet tree. Thus the method becomes unaffected by certain phenomena such as the long branch attraction, hence gaining in robustness. We have implemented it in c++ and future work includes testing it for simulated DNA sequences and real data.
2 Background

2.1 Alignments of DNA sequences

As it has been said before, nowadays it is usual to look at the DNA molecules in the genome of the species to reconstruct phylogenetic trees. The DNA is a nucleic acid, formed by smaller molecules called nucleotides. A nucleotide is composed of a sugar, a phosphate group and a nitrogen-containing base, which determines the type of the nucleotide. There are four different nitrogenous bases: adenine (A), cytosine (C), guanine (G) or thymine (T). The nucleotides in DNA are arranged in two polymer strands coiled one around the other forming a double helix. However, one strand is determined by the other: the two polynucleotide strands are bonded together by hydrogen bonds between the nitrogenous bases of each strand, following the base pairing rules (A-T and C-G). Consequently, we will think about DNA as an ordered sequence of nucleotides (representing one of the strands).

Figure 2: Double strand DNA chain

Due to the existence of mutation processes, the DNA sequences corresponding to the same gene are different for different species. A mutation is a change of a nucleotide in the DNA chain. There are three types of mutations: suppression (one nucleotide is deleted from the DNA chain), substitution (one nucleotide takes the place of another) or insertion (an extra nucleotide is added to the chain). Therefore, two sequences that have evolved from the same DNA sequence can look very different and even have a different number of nucleotides. Furthermore, they may not be in the same place in the genome because not all
the species have the same number of nucleotides, genes nor chromosomes. This is why, in order to study the relation between two species, it is important to know which parts of the genome have to be compared. An alignment is a table that contains sequences of nucleotides of different species (one in each row), where all the nucleotides in the same column are supposed to come from the same nucleotide of the common ancestor. Finding the optimal alignment for a set of DNA sequences is a complex problem. As it is usually done in phylogenetics, we shall assume that the alignment is given and we are only going to consider mutations of substitution.

<table>
<thead>
<tr>
<th></th>
<th>Homo Sapiens</th>
<th>Gorilla gorilla</th>
<th>Pan troglodytes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AACTTCGAGGCTTACCGCTG</td>
<td>AACGTCTATGCTCACCGATG</td>
<td>AAGGTCGATGCTCACCGATG</td>
</tr>
</tbody>
</table>

Example of alignment

Once we have an alignment and a probabilistic evolutionary model (see section 2.3), we can compute the relative frequencies of each combination of nucleotides. Since each phylogenetic tree has its own evolutionary characteristics, the probability of obtaining a certain alignment is different for each tree. There are different methods that use these probabilities to determine the tree that has originated the alignment.

2.2 Phylogenetic trees

A phylogenetic tree is a branching diagram that represents the evolution of different species. Each leaf is related with a current species and the internal nodes are associated to common ancestors. The branches between two nodes represent the evolutionary processes between them. The length of these branches represents the evolutionary distance between the two species it connects (namely, the amount of elapsed substitutions between them).

Mathematically speaking, a tree is a connected acyclic graph.

**Definition 2.1.** A phylogenetic tree on a set of species $S$ is a tree $T$ together with a bijection $\Psi$ from the set of leaves of $T$ to $S$.

**Definition 2.2.** The degree of a node in a graph is the number of edges adjacent to it. The leaves of a tree are the nodes whose degree is one.
The bijection $\Psi$ labels the leaves of the tree with species in $S$. The topology of a phylogenetic tree is the topology that corresponds to the labeled graph (without taking into account the length of the branches). This is, two trees are topologically equivalent if there exists a graph isomorphism between both that preserves the labeling of the leaves.

Phylogenetic trees may have a distinguished interior vertex, called root. A root represents a common ancestor for all the current species of the tree. Rooted trees can be assumed to be binary, which means that every branch is divided in two other branches until we arrive to the leaves. However, it is impossible to reconstruct the exact position of this root using evolutionary models and, therefore, most of phylogenetic methods consider trees without root. These unrooted trees are assumed to be trivalent: each inner node has degree 3.
shows the three possible different topologies for this kind of trees which are usually denoted by AB|CD, AC|BD and AD|BC respectively:

![Figure 6: The three different topologies for unrooted trees with 4 leaves](image)

### 2.3 Markov models of nucleotide substitution

Let T be a phylogenetic tree on a set of n species, either rooted or unrooted. If the tree is rooted then the root naturally defines an orientation on the edges of the tree. If it is not rooted, then we choose an internal node and direct the edges out of it.

Apart from that, in order to define a statistical model for the evolutionary process, one assumes certain conditions:

a) Substitutions in a DNA sequence are completely random.

b) Different positions in the DNA sequence evolve independently and with the same substitution probabilities.

c) The evolutionary process follows a Markov process, that is, the substitutions on a branch only depend on the parent node.

Then, we fix a phylogenetic tree T with n leaves and we assign a discrete random variable $X_i$ to each node $i$ of the tree (as it is shown in figure 7) whose possible values are the different nucleotides we have: A, C, G, T. The variables related to the leaves of the tree are observed because they correspond to current species and we have DNA sequences of them. On the other hand, the inner nodes are related to the ”hidden” variables. Each observation at the leaves of the tree can be expressed as an observation of the vector $X = (X_1, ..., X_n)$.

Now, since substitutions only depend on the present stage and not on the previous steps (Markov process), we can associate a transition matrix for each branch $e$. This matrix will represent the probability for each nucleotide to change during the evolutionary process along the branch:
Figure 7: Left: phylogenetic tree corresponding to the species *Gorilla Gorilla*, *Pan troglodytes* and *Homo Sapiens*. Right: statistical model of a rooted 3-leaved tree.

\[
M^e = \begin{pmatrix}
A & C & G & T \\
C & P(A|C, e) & P(C|C, e) & P(G|C, e) & P(T|C, e) \\
T & P(A|T, e) & P(C|T, e) & P(G|T, e) & P(T|T, e)
\end{pmatrix},
\]

where \( P(x|y, e) \) is the conditional probability for the nucleotide \( y \) to be substituted by a nucleotide \( x \) through the evolutionary process along the branch \( e \). Note that each row sums to 1 and this will be used throughout the project.

All these probabilities together with the nucleotide distribution of the root node, expressed as a vector \( \pi = (\pi_A, \pi_C, \pi_G, \pi_T) \) are the (unknown) parameters of the model.

Once we have defined these matrices, depending on the restrictions that we impose to the probabilities we can obtain different evolutionary models: (see [Cas12], [CFS10a], [PS05])

- **General Markov model.** No restrictions in the probabilities.

\[
M^e = \begin{pmatrix}
a_e & b_e & c_e & d_e \\
e_e & f_e & g_e & h_e \\
i_e & j_e & k_e & l_e \\
m_e & n_e & o_e & p_e
\end{pmatrix}
\]

- **Strand symmetric.** To reflect the symmetry of the DNA molecule, it assumes \( \pi_A = \pi_T, \pi_C = \pi_G \) and transition matrices of the form:
\[
M^e = \begin{pmatrix}
a_e & b_e & c_e & d_e \\
e_e & f_e & g_e & h_e \\
h_e & g_e & f_e & c_e \\
d_e & c_e & b_e & a_e
\end{pmatrix}
\]

- **Kimura 3-parameter.** It assumes all nucleotides are uniformly distributed \( \pi_A = \pi_C = \pi_G = \pi_T = 1/4 \) and, to distinguish between the types of substitution between classes of nucleotides (purines and pyrimidines), the transition matrices are of the form:

\[
M^e = \begin{pmatrix}
a_e & b_e & c_e & d_e \\
b_e & a_e & d_e & c_e \\
c_e & d_e & a_e & b_e \\
d_e & c_e & b_e & a_e
\end{pmatrix}
\]

- **Kimura 2-parameter.** Same assumptions as Kimura 3-parameter but assuming also that \( d_e = b_e \).

\[
M^e = \begin{pmatrix}
a_e & b_e & c_e & b_e \\
b_e & a_e & b_e & c_e \\
c_e & b_e & a_e & b_e \\
b_e & c_e & b_e & a_e
\end{pmatrix}
\]

- **Jukes-Cantor.** Uniform distribution of nucleotides at the root and same probability of substitution for all nucleotides.

\[
M^e = \begin{pmatrix}
a_e & b_e & b_e & b_e \\
b_e & a_e & b_e & b_e \\
b_e & b_e & a_e & b_e \\
b_e & b_e & b_e & a_e
\end{pmatrix}
\]

The model used throughout this project is the Kimura 3-parameter, but our results also hold for its submodels Kimura 2-parameter and Jukes-Cantor.

Now that we have described the transition matrices, we are able to compute the probability of having an alignment from a given tree. Consider a rooted tree \( T \) of three species \( (X_1, X_2, X_3) \), like the one in figure 7. The probability of having an observation \( x_1 x_2 x_3 \) at the leaves can be expressed as

\[
p_{x_1,x_2,x_3}^T = P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | T)
\]
This probability can be written in terms of the entries of the transition matrices of each branch as follows:

\[ p_{x_1x_2x_3}^T = \sum_{x_r \in \{A,C,G,T\}} \pi_{x_r} M^1(x_r, x_1) M^2(x_r, x_2) M^3(x_r, x_3). \]

**Remark 2.3.** The root of a phylogenetic tree is not identifiable. That is, different root placements can lead to the same vector of probabilities at the leaves, if we allow different transition matrices.

Therefore, we cannot aim at inferring the location of the root from an alignment. However, to write the joint distribution at the leaves in terms of the transition matrices, we need to orient the tree from a chosen interior node. For example, in case of having 4 different species (see figure 8), we orient the tree out of \( r \). Then, the joint distribution at the leaves would be given by

\[ p_{x_1x_2x_3x_4}^T = \sum_{x_r, x_s \in \{A,C,G,T\}} \pi_{x_r} M^1(x_r, x_1) M^2(x_r, x_2) M^e(x_r, x_s) M^3(x_s, x_3) M^4(x_s, x_4) \]

Figure 8: Phylogenetic 4-leaved tree (denoted as 1234) with transition matrices \( M^1, M^2, M^3, M^4, M^e \).

Given an alignment of \( n \) DNA sequences of a set of species \( S \), the relative frequencies of column patterns \( x_1...x_n \) are estimations \( \hat{p}_{x_1...x_n} \) of the probabilities \( p_{x_1...x_n}^T \) of observing \( x_1...x_n \) at the leaves of the tree \( T \) which lead to the alignment.

Thus, the relative frequencies give a vector \( \hat{p} \) that can be compared to the vector of joint distribution at the leaves of the different possible trees \( T \).

As it has been said before, there are different methods to use these relative frequencies to determine from which tree the alignment comes from. One of these methods is the
maximum likelihood. It selects the tree that maximizes the probability of obtaining a certain alignment, estimating the parameters involved in the process. Another method is the neighbor-joining. It consists in finding the two species of the tree that minimize a function, which tries to represent the evolutionary distance between two species. Once they are found, they are joined together to an inner node and they are substituted by a new species represented by that node. The method keeps going until all the species are represented in the tree.

Both methods have advantages and disadvantages and a new area that uses algebraic statistics in phylogenetics is emerging. The method we propose here lies within this new area.

2.4 Flattening

In order to facilitate the representation of the probability vectors, it is usual to rewrite them as matrices.

**Definition 2.4.** Let $p$ be a probability vector for a certain alignment of $4$ species,

$$p = (p_{AAAA}; p_{AAAC}; p_{AAAG}; p_{AAAT}; \ldots; p_{TTTT})$$

Then, the flattening of $p$ for the bipartition $12|34$ of the leaves is

$$\text{flatt}_{12|34}(p) = \begin{pmatrix}
  p_{AAAA} & p_{AAAC} & \cdots & p_{AAAT} \\
  p_{ACAA} & p_{ACAC} & \cdots & p_{ACTT} \\
  \vdots & \vdots & \ddots & \vdots \\
  p_{TTAA} & p_{TTAC} & \cdots & p_{TTTT}
\end{pmatrix}$$

We can extend this definition to the other possible bipartitions of leaves: $\text{flatt}_{13|24}(p)$, $\text{flatt}_{14|23}(p)$.

Therefore, the flattening of a vector is one of its representations as a matrix, where the rows represent the states at two leaves (or species) and the columns the other two.

Moreover, these flattening matrices present different characteristics depending on which tree the probability vector comes from.

**Theorem 2.5.** (See [AR07]) Under the general Markov model, if $p$ is the joint distribution at the leaves of the tree $12|34$, the matrix $\text{flatt}_{12|34}(p)$ has rank less than or equal to $4$ while $\text{flatt}_{13|24}(p)$ or $\text{flatt}_{14|23}(p)$ have rank $16$ in general.
Therefore, the flattening of the probability vector is a useful tool to get more information about the original tree.

If the Markov model is another of the ones introduced in the previous section, then there is also a characterization of the flattening in terms of the rank of certain blocks, see [CFS10b] and section 3.2 for the Kimura 3-parameter model.
3 A new phylogenetic reconstruction model

After the previous introduction to phylogenetics, now we focus on the Kimura 3-parameter model. Furthermore, we shall use a new coordinate system, called the Fourier coordinates and see how probability parameters change and how to use the relations between the new parameters to determine the tree topology that originated a certain alignment.

3.1 Fourier coordinates

The ideas underlying this section can be found in [CGS05], [Ste94], [SSE93].

First of all, before defining the new system, we consider the bijection between the set of nucleotides Σ = {A, C, G, T} and the group G = Z₂ × Z₂ given by

\[
\begin{align*}
{A, C, G, T} & \leftrightarrow (Z₂ \times Z₂, +) \\
A & \leftrightarrow (0, 0) \\
C & \leftrightarrow (0, 1) \\
G & \leftrightarrow (1, 0) \\
T & \leftrightarrow (1, 1)
\end{align*}
\]

This bijection makes Σ into a group where A is the neutral element.

Now, we consider \( \chi_i \) as the character function \( \chi_i : G \rightarrow (C^*, \cdot) \) defined by the following table

\[
\begin{array}{c|cccc}
& A & C & G & T \\
\hline
\chi_A & 1 & 1 & 1 & 1 \\
\chi_C & 1 & 1 & -1 & -1 \\
\chi_G & 1 & -1 & 1 & -1 \\
\chi_T & 1 & -1 & -1 & 1 \\
\end{array}
\]

that is, each entry of the table is the value \( \chi_i(j) \) of the character \( \chi_i \) applied to the element \( j \).

Now, from the fact that \( \chi \) is a group homomorphism and the symmetry of the table, we can deduce some properties.

**Proposition 3.1.** With the above notation, we have

(i) \( \chi_i(j) = \chi_j(i) \);

(ii) \( \chi_i(j) = \chi_i(-j) \);
(iii) $\chi_{i}^{\prime}(j) = \chi_{i}(j)$;

(iv) $\chi_{i+j}(k) = \chi_{i}(k) \cdot \chi_{j}(k)$;

(v) $\chi_{i}(j + k) = \chi_{i}(j) \cdot \chi_{i}(k)$;

(vi) $\sum_{j} \chi_{i}(j) = \left\{ \begin{array}{ll}
4 & \text{if } i = A \\
0 & \text{if } i \neq A
\end{array} \right.$

As it has been said before, we are going to use a new coordinate system: the Fourier coordinates. This change will affect the probability matrices as well as the probability coordinates themselves.

**Definition 3.2.** Let $\Sigma = \{A, C, G, T\}$ be a coordinate basis. We define $\hat{\Sigma} = \{\bar{A}, \bar{C}, \bar{G}, \bar{T}\}$ as the Fourier basis, where

\[
\bar{A} = A + C + G + T \quad \bar{C} = A - C + G - T \\
\bar{G} = A + C - G - T \quad \bar{T} = A - C - G + T
\]

Note that the matrix of base change from $\hat{\Sigma}$ to $\Sigma$ is given by

\[
Q = \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 \\
1 & -1 & -1 & 1
\end{pmatrix}
\]

Note also that $Q^{-1} = \frac{1}{4}Q$.

**Definition 3.3.** Given a vector $p = (p_{x})_{x \in \Sigma}$, we define the Fourier transform of $p$ as

$\hat{p} = (q_{x})_{x \in \Sigma}$

such that

\[
\sum_{x \in \Sigma} p_{x}x = \sum_{x \in \Sigma} q_{x}\bar{x}.
\]

In other words, the Fourier transform of $p$ is given by the Fourier coordinates of $p$:

\[
p = \sum_{i \in \{A, C, G, T\}} p_{i} \cdot i \quad \Rightarrow \quad p = \sum_{i} q_{i} \cdot \bar{i}
\]

Similarly, we can define the Fourier transform of a tensor $p$:
**Definition 3.4.** Given a vector \( p = (p_{x_1...x_n})_{x_1,...,x_n \in \Sigma} \), we define the Fourier transform of \( p \) as

\[
\hat{p} = (q_{x_1...x_n})_{x_1,...,x_n \in \Sigma}
\]

such that

\[
\sum_{x_1...x_n \in \Sigma} p_{x_1...x_n} x_1 \otimes ... \otimes x_n = \sum_{x_1...x_n \in \Sigma} q_{x_1...x_n} \bar{x}_1 \otimes ... \otimes \bar{x}_n.
\]

The following proposition clarifies the dependence between the probability coordinates and the Fourier coordinates.

**Proposition 3.5.** We have

\[
p_{i_1...i_n} = \sum_{j_1...j_n} \chi_{i_1}(j_1) ... \chi_{i_n}(j_n) q_{j_1...j_n};
\]

\[
q_{i_1...i_n} = \frac{1}{4^n} \sum_{j_1...j_n} \chi_{i_1}(j_1) ... \chi_{i_n}(j_n) p_{j_1...j_n}.
\]

We can observe that, in fact, this transformation, for \( n = 4 \), is equivalent to

\[
q = \frac{1}{4^n} Q \otimes Q \otimes Q \otimes Q \cdot p.
\]

Now that we have seen how probability coordinates are transformed, it is time for the probability matrices. Given \( M \) a transition matrix, which corresponds to a linear map \( \mathbb{C}^4 \to \mathbb{C}^4 \) with matrix \( M \) in the canonical basis \( \{A, C, G, T\} \), and \( \hat{M} \) the matrix corresponding to the same map but on the basis \( \{ar{A}, \bar{C}, \bar{G}, \bar{T}\} \), then

\[
\hat{M} = Q^{-1} \cdot M \cdot Q.
\]

Let us see now how matrices from the Kimura 3-parameter model are transformed.

**Lemma 3.6.** If \( M \) is a Kimura 3-parameter matrix,

\[
M = \begin{pmatrix}
    a & b & c & d \\
    b & a & d & c \\
    c & d & a & b \\
    d & c & b & a
\end{pmatrix}
\]

then its Fourier transform is given by
\[
\hat{M} = \begin{pmatrix}
(a+b+c+d) & 0 & 0 & 0 \\
0 & (a+b-c-d) & 0 & 0 \\
0 & 0 & (a-b+c-d) & 0 \\
0 & 0 & 0 & (a-b-c+d)
\end{pmatrix}
\]

Moreover, the uniform distribution vector for the nucleotides in the root \(\pi_r = (\pi_A, \pi_C, \pi_G, \pi_T) = (1/4, 1/4, 1/4, 1/4)\) becomes \(\hat{\pi}_r = (1/4, 0, 0, 0)\).

**Proof.** It follows by direct computation:

\[
\hat{M} = Q^{-1}MQ = \frac{1}{4}QMQ = \frac{1}{4} \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 \\
1 & -1 & -1 & 1
\end{pmatrix}
\begin{pmatrix}
a & b & c & d \\
b & a & d & c \\
c & d & a & b \\
d & c & b & a
\end{pmatrix}
\begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 \\
1 & -1 & -1 & 1
\end{pmatrix}
\]

\[
= \frac{1}{4}
\begin{pmatrix}
4(a+b+c+d) & 0 & 0 & 0 \\
0 & 4(a+b-c-d) & 0 & 0 \\
0 & 0 & 4(a-b+c-d) & 0 \\
0 & 0 & 0 & 4(a-b-c+d)
\end{pmatrix}
\]

\[
\Rightarrow \hat{M} = \begin{pmatrix}
(a+b+c+d) & 0 & 0 & 0 \\
0 & (a+b-c-d) & 0 & 0 \\
0 & 0 & (a-b+c-d) & 0 \\
0 & 0 & 0 & (a-b-c+d)
\end{pmatrix}
\]

and

\[
\hat{\pi}_r = Q^{-1} \cdot \pi_r = \frac{1}{4} \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 \\
1 & -1 & -1 & 1
\end{pmatrix}
\begin{pmatrix}
1/4 \\
1/4 \\
1/4 \\
1/4
\end{pmatrix}
= \begin{pmatrix}
1/4 \\
0 \\
0 \\
0
\end{pmatrix}
\]

**Remark 3.7.** Note that the transformed matrices are diagonal. This is the reason why we have picked this particular new system.
We introduce new notation for the elements in the diagonal of the transformed matrix \( \hat{M} \), which are the eigenvalues of the original transition matrix \( M \):

\[
\begin{align*}
m_A &= a + b + c + d, \\
m_C &= a + b - c - d, \\
m_G &= a - b + c - d, \\
m_T &= a - b - c + d.
\end{align*}
\]

As \( M \) is a transition matrix, its rows sum to one and hence \( m_A = 1 \).

In this project, we are considering the model Kimura 3-parameters (so our transformed matrices will be diagonal) and trees with 4 leaves (\( n = 4 \)).

From now on and unless we say otherwise, we assume that the vector \( p = (p_{x_1 x_2 x_3 x_4})_{x_i \in \Sigma} \) is a pattern distribution obtained from the tree 12\mid34 (see figure 8) with transition matrices \( M^1, M^2, M^3, M^4, M^e \).

The expression for \( p \) in terms of the transition matrices for each branch in this case is

\[
p_{x_1 x_2 x_3 x_4} = \sum_{x_r, x_s} (\pi_r)_{x_r} M^1_{x_r x_1} M^2_{x_r x_2} M^e_{x_r x_3} M^3_{x_r x_4} M^4_{x_r x_4}.
\]

We need to keep in mind that the Fourier coordinates cannot be directly related to any biological concept anymore. From now on, we deal with them as just parameters.

First of all, we state an easy lemma that will be useful in the following proofs.

**Lemma 3.8.** Let \( M \) be a transition matrix and write \( \hat{M} \) for its Fourier transform. Then,

\[
\sum_j \chi_N(j)M_{i,j} = \chi_N(i)m_N
\]

where \( m_N \) is the eigenvalue of the matrix \( M \) corresponding to the nucleotide \( N \).

In particular, if \( N = A \), this is equal to 1.

**Proof.** We can express the eigenvalues of the matrix \( M \) as scalar products of its four different entries \((a, b, c, d)\) and the column of the matrix \( Q \) corresponding to the nucleotide \( N \) as follows:
\(m_A = (a, b, c, d) \cdot Q_A = (a, b, c, d) \cdot (1, 1, 1, 1)^t = a + b + c + d\)

\(m_C = (a, b, c, d) \cdot Q_C = (a, b, c, d) \cdot (1, 1, -1, -1)^t = a + b - c - d\)

\(m_G = (a, b, c, d) \cdot Q_G = (a, b, c, d) \cdot (1, -1, 1, -1)^t = a - b + c - d\)

\(m_T = (a, b, c, d) \cdot Q_T = (a, b, c, d) \cdot (1, -1, -1, 1)^t = a - b - c + d\)

and, in general,

\(m_N = (a, b, c, d) \cdot Q_N = \chi_N(A)a + \chi_N(C)b + \chi_N(G)c + \chi_N(T)d\)

Therefore, we have that

\(m_N = \sum_j \chi_N(j)M_{A,j}\)

On the other hand, each entry \(M_{i,j}\) only depends on the difference \(j - i\). Thus we have \(M_{i,j} = M_{i,i+j-i} = M_{h,j-i}\) Using this property together with Proposition 3.1, we get that

\[\sum_j \chi_N(j)M_{i,j} = \sum_j \chi_N(j)M_{h,j-i} = \sum_j \chi_N(i)\chi_N(-i)\chi_N(j)M_{h,j-i}\]

\[= \chi_N(i)\sum_j \chi_N(j-i)M_{h,j-i} = \chi_N(i)m_N\]

In particular, if \(N = A\) then \(\chi_A(i) = 1\) and

\[\sum_j \chi_A(j)M_{i,j} = \sum_j M_{i,j} = 1\]

using that the sum of each row in \(M\) is 1.

Now, we can formulate the following proposition.

**Proposition 3.9.**

\[q_{x_1,x_2,x_3,x_4} = \begin{cases} m_{x_1}^1 m_{x_2}^2 m_{x_1+x_2}^3 m_{x_3}^4 m_{x_4}^4, & \text{if } x_1 + x_2 = x_3 + x_4; \\ 0, & \text{otherwise.} \end{cases}\]

where \(m_x^i\) is the eigenvalue of the Fourier transform \(\hat{M}^i\) of \(M^i\).

**Proof.** It follows by direct computation:
\[ q_{x_1x_2x_3x_4} = \sum_{j_1j_2j_3j_4} \chi_{x_1}(j_1) \chi_{x_2}(j_2) \chi_{x_3}(j_3) \chi_{x_4}(j_4) p_{j_1j_2j_3j_4} \]

\[ = \sum_{j_1j_2j_3j_4} \chi_{x_1}(j_1) \chi_{x_2}(j_2) \chi_{x_3}(j_3) \chi_{x_4}(j_4) \sum_{j_r} (\pi_r)_{j_r} M_{j_rj_1}^1 M_{j_rj_2}^2 M_{j_rj_3}^3 M_{j_rj_4}^4 \]

\[ = \frac{1}{4} \sum_{j_1j_2j_3j_4j_r} \left\{ \chi_{x_1}(j_1) \chi_{x_2}(j_2) \chi_{x_3}(j_3) \chi_{x_4}(j_4) \right\}

\cdot M_{h,j_1-j_r}^1 M_{h,j_2-j_r}^2 M_{h,j_3-j_r}^3 M_{h,j_4-j_r}^4 \}

where we have used the fact that \( M_{i,j} = M_{h,j-i} \).

\[ q_{x_1x_2x_3x_4} = \frac{1}{4} \sum_{j_1j_2j_3j_4j_r} \left\{ \chi_{x_1}(j_1 + j_r) \chi_{x_2}(j_2 + j_r) \chi_{x_3}(j_3 + j_s + j_r) \chi_{x_4}(j_4 + j_s + j_r) \right\}

\cdot M_{h,j_1}^1 M_{h,j_2}^2 M_{h,j_3}^3 M_{h,j_4}^4 \}

where \( \tilde{j}_1 = j_1 - j_r, \tilde{j}_2 = j_2 - j_r, \tilde{j}_3 = j_3 - j_s, \tilde{j}_4 = j_4 - j_s, \tilde{j}_s = j_s - j_r \).

Now we focus on the first part of the summation. Using the properties described in Proposition 3.1, we have

\[ \chi_{x_1}(\tilde{j}_1 + j_r) \chi_{x_2}(\tilde{j}_2 + j_r) \chi_{x_3}(\tilde{j}_3 + j_s) \chi_{x_4}(\tilde{j}_4 + j_s) \]

\[ = \chi_{x_1}(\tilde{j}_1) \chi_{x_1}(j_r) \chi_{x_2}(\tilde{j}_2) \chi_{x_2}(j_r) \chi_{x_3}(\tilde{j}_3 + j_s + j_r) \chi_{x_4}(\tilde{j}_4 + j_s + j_r) \]

\[ = \chi_{x_1}(\tilde{j}_1) \chi_{x_1}(j_r) \chi_{x_2}(\tilde{j}_2) \chi_{x_2}(j_r) \chi_{x_3}(\tilde{j}_3) \chi_{x_3}(j_r) \chi_{x_4}(\tilde{j}_4) \chi_{x_4}(j_r) \]

\[ = \chi_{x_1}(\tilde{j}_1) \chi_{x_2}(\tilde{j}_2) \chi_{x_3}(\tilde{j}_3) \chi_{x_3}(\tilde{j}_4) \chi_{x_4}(\tilde{j}_4) \chi_{x_4}(j_r) \]

\[ = \chi_{x_1}(\tilde{j}_1) \chi_{x_2}(\tilde{j}_2) \chi_{x_3}(\tilde{j}_3) \chi_{x_3}(\tilde{j}_4) \chi_{x_4}(\tilde{j}_4) \chi_{x_4}(j_r) \chi_{x_1}(j_r) + \chi_{x_2}(j_r) + \chi_{x_3}(j_r) + \chi_{x_4}(j_r) \]

Returning to the whole expression:
Therefore, using Lemma 3.8 in the last step.

Using property (vi) from Proposition 3.1, we know that

\[ q_{x_1 x_2 x_3 x_4} = \frac{1}{4} \sum_{j_1, j_2, j_3, j_4, j_r} \{ x_1 (\tilde{j}_1) x_2 (\tilde{j}_2) x_3 (\tilde{j}_3) x_4 (\tilde{j}_4) \} \cdot \left[ M^1_{\tilde{j}_1} M^2_{\tilde{j}_2} M^4_{\tilde{j}_3} M^4_{\tilde{j}_4} \right] \]

\[ = \frac{1}{4} \left( \sum_{j_1, j_2, j_3, j_4} \chi_{x_1 + x_2 + x_3 + x_4} (\tilde{j}_r) \chi_{x_1} (\tilde{j}_1) \chi_{x_2} (\tilde{j}_2) \chi_{x_3} (\tilde{j}_3) \chi_{x_4} (\tilde{j}_4) \right) \cdot \left( \sum_{j_r} \chi_{x_1 + x_2 + x_3 + x_4} (\tilde{j}_r) \right) \]

Using property (vi) from Proposition 3.1, we know that

\[ \sum_{j_r} \chi_{x_1 + x_2 + x_3 + x_4} (\tilde{j}_r) = \begin{cases} 4 & \text{if } x_1 + x_2 + x_3 + x_4 = A; \\ 0 & \text{if } x_1 + x_2 + x_3 + x_4 \neq A. \end{cases} \]

This means that \( q_{x_1 x_2 x_3 x_4} = 0 \) if \( x_1 + x_2 + x_3 + x_4 \neq A \) so we are going to consider only the terms where \( x_1 + x_2 + x_3 + x_4 = A \) or, in other words, where \( x_1 + x_2 = x_3 + x_4 \). Then, if \( x_1 + x_2 = x_3 + x_4 \),

\[ q_{x_1 x_2 x_3 x_4} = \frac{1}{4} \cdot 4 \left( \sum_{j_1} \chi_{x_1} (\tilde{j}_1) M^1_{\tilde{j}_1} \right) \left( \sum_{j_2} \chi_{x_2} (\tilde{j}_2) M^2_{\tilde{j}_2} \right) \left( \sum_{j_3} \chi_{x_3} (\tilde{j}_3) M^3_{\tilde{j}_3} \right) \left( \sum_{j_4} \chi_{x_4} (\tilde{j}_4) M^4_{\tilde{j}_4} \right) \cdot \left( \sum_{j_3} \chi_{x_3} (\tilde{j}_3) \right) \left( \sum_{j_4} \chi_{x_4} (\tilde{j}_4) \right) \]

\[ = m_{x_1}^1 m_{x_2}^2 m_{x_3 + x_4}^e m_{x_3}^3 m_{x_4}^4 \]

using Lemma 3.8 in the last step.

Therefore,

\[ q_{x_1 x_2 x_3 x_4} = m_{x_1}^1 m_{x_2}^2 m_{x_1 + x_2}^e m_{x_3}^3 m_{x_4}^4 \]

\[ \square \]
3.2 Flattening in Fourier coordinates

Once we have the vector expressed in the Fourier coordinates, we consider the flattening of \( \hat{p} \) relative to the bipartition 12|34 in order to derive new properties. To this aim, \( \hat{p} \) is represented as a matrix, whose rows and columns will be indexed by the states of the leaves 1, 2 and 3, 4 respectively (see Definition 2.4).

However, since most of the coordinates of this vector are 0 (Proposition 3.9), we can rearrange the entries of this matrix to get a block diagonal matrix. Since the non-zero coordinates are those \( q_{x_1 x_2 x_3 x_4} \) where \( x_1 + x_2 = x_3 + x_4 \), we order rows and columns to form blocks in terms of the value of this sum:

- first block: \( x_1 + x_2 = x_3 + x_4 = A \), corresponding to the pairs of nucleotides: A+A, C+C, G+G, T+T.
- second block: \( x_1 + x_2 = x_3 + x_4 = C \), corresponding to the pairs of nucleotides: A+C, C+A, G+T, T+G.
- third block: \( x_1 + x_2 = x_3 + x_4 = G \), corresponding to the pairs of nucleotides: A+G, C+T, G+A, T+C.
- fourth block: \( x_1 + x_2 = x_3 + x_4 = T \), corresponding to the pairs of nucleotides: A+T, C+G, G+C, T+A.

And the matrix will have the form

\[
\overrightarrow{\text{flatt}}_{12|34}(p) = \begin{pmatrix}
B_A & 0 & 0 & 0 \\
0 & B_C & 0 & 0 \\
0 & 0 & B_G & 0 \\
0 & 0 & 0 & B_T
\end{pmatrix}
\]

where each block is

\[
B_A = \begin{pmatrix}
q_{AAAA} & q_{AACC} & q_{AAGG} & q_{AATT} \\
q_{CCAA} & q_{CCCC} & q_{CCGG} & q_{CCCT} \\
q_{GGAA} & q_{GGCC} & q_{GGGG} & q_{GGTT} \\
q_{TTAA} & q_{TTCC} & q_{TTGG} & q_{TTTT}
\end{pmatrix},
\]

\[
B_C = \begin{pmatrix}
q_{ACAC} & q_{ACCA} & q_{ACGT} & q_{ACTG} \\
q_{CAAC} & q_{CACA} & q_{CAGT} & q_{CATG} \\
q_{GTAC} & q_{GTCA} & q_{GTGT} & q_{GTGG} \\
q_{TGAC} & q_{TGCA} & q_{TGTG} & q_{TGGT}
\end{pmatrix},
\]

\[
B_G = \begin{pmatrix}
q_{AGAG} & q_{AGCT} & q_{AGGA} & q_{AGTC} \\
q_{CTAG} & q_{CTCT} & q_{CTGA} & q_{CTTC} \\
q_{GAGA} & q_{GACT} & q_{GAAG} & q_{GATC} \\
q_{TCAG} & q_{TCCT} & q_{TCGA} & q_{TCTC}
\end{pmatrix},
\]

\[
B_T = \begin{pmatrix}
q_{ATAT} & q_{ATCG} & q_{ATGC} & q_{ATTA} \\
q_{CTAT} & q_{CTCG} & q_{CTGC} & q_{CTTA} \\
q_{GATG} & q_{GCCT} & q_{GCAG} & q_{GCCT} \\
q_{TAGA} & q_{TAGT} & q_{TAGG} & q_{TATA}
\end{pmatrix}.
\]
Definition 3.10. This block diagonal matrix will be called the Fourier flattening of \( p \) relative to 12|34.

We observe that, by Proposition 3.9, each block of the Fourier flattening relative to 12|34 has rank 1 if \( p \) is a distribution on the 12|34 topology and \( \neq 1 \) otherwise (see [CFS10b]).

### 3.3 Flattening adjustment

As it has been said above, the length of the branches of the tree gives us information about the number of mutations between two species. Although this is interesting for biologists, our goal is to determine the topology of the phylogenetic tree, so we focus on the topology instead of the branch length. To do so, we "normalize" each row of the Fourier flattening of \( p \) relative to 12|34 by dividing it by \( m_{x_1}^1 m_{x_2}^2 \) and similarly for columns (see Proposition 3.9).

First of all we proceed with block \( A \) and define this normalization term as the Fourier sum of the probabilities in the same row.

**Definition 3.11.** We define the marginal of leaves 1,2 as

\[
p_{x_1x_2++} = \sum_{x_3x_4} p_{x_1x_2x_3x_4}.
\]

Analogously, we consider the marginal of leaves 3,4 as \( p_{+++x_3x_4} \).

These marginals can be computed easily from the transition matrices of the tree.

**Proposition 3.12.** The marginal of leaves 1,2 is equal to

\[
p_{x_1x_2++} = \frac{1}{4} (M^1 \cdot M^2)_{x_1,x_2}.
\]

Therefore, the Fourier transform of \( (p_{x_1x_2++})_{x_1x_2 \in \Sigma} \) is given by \( (\hat{p}_{x_1x_2++})_{x_1x_2 \in \Sigma} \), where

\[
\hat{p}_{x_1x_2++} = \begin{cases} 
\frac{1}{4} m_{x_1}^1 m_{x_2}^2, & \text{if } x_1 = x_2; \\
0, & \text{if } x_1 \neq x_2.
\end{cases}
\]
Proof.

\[ p_{x_1 x_2 ++} = \sum_{x_3 x_4} p_{x_1 x_2 x_3 x_4} = \sum_{x_3 x_4} \sum_{x_s x_r} \pi_r M^1_{x_r x_1} M^2_{x_r x_2} M^e_{x_s x, r} M^3_{x_s x_3} M^4_{x_s x_4} = \]

\[ = \frac{1}{4} \sum_{x_r} M^1_{x_r x_1} M^2_{x_r x_2} \left( \sum_{x_s} M^e_{x_r x_s} \right) \left( \sum_{x_3} M^3_{x_3 x_3} \right) \left( \sum_{x_4} M^4_{x_4 x_4} \right) = \frac{1}{4} \sum_{x_r} M^1_{x_r x_1} M^2_{x_r x_2} \]

where we have used that the sum of all the elements of a row in a transition matrix is 1. Therefore, we get that

\[ p_{x_1 x_2 ++} = \frac{1}{4} (M^{1t} \cdot M^2)_{x_1, x_2} = \frac{1}{4} (M^1 \cdot M^2)_{x_1, x_2}, \]

or as a matrix:

\[ p_{++} = \frac{1}{4} M^1 \cdot M^2 \]

We compute now the Fourier transform of this expression.

\[ \hat{p}_{++} = (\hat{p}_{x_1 x_2 ++}) = Q^{-1} (p_{x_1 x_2 ++}) Q = \frac{1}{4} Q^{-1} M^1 M^2 Q \]

\[ = \frac{1}{4} (Q^{-1} M^1 Q) (Q^{-1} M^2 Q) = \frac{1}{4} \hat{M}^1 \hat{M}^2 \]

Therefore,

\[ \hat{p}_{x_1 x_2 ++} = \frac{1}{4} (\hat{M}^1 \hat{M}^2)_{x_1, x_2} = \begin{cases} \frac{1}{4} m^1_{x_1} m^2_{x_2}, & \text{if } x_1 = x_2; \\ 0, & \text{if } x_1 \neq x_2. \end{cases} \]

\[ \square \]

**Remark 3.13.** Notice that the Fourier transform of the marginal does not coincide with the marginal of the Fourier coefficients \( q_{x_1 x_2 ++} : \)

\[ q_{x_1 x_2 ++} = \sum_{x_3 x_4} q_{x_1 x_2 x_3 x_4} = \sum_{x_3 x_4} m^1_{x_3} m^2_{x_4} m^e_{x_1 x_2 x_3} m^3_{x_3 x_4} m^4_{x_3 x_4} \]

\[ = m^1_{x_1} m^2_{x_2} m^e_{x_1 x_2} \sum_{x_3} m^3_{x_3} m_{x_1 x_2 x_3} m^4_{x_3} \]

the last equality because of Proposition 3.9. This sum depends on the probabilities of the 3rd, 4th and central branches. Therefore, it cannot be the same, in general, as the previous one.
We have seen before (Lemma 3.6) that the transformed matrices obtained from Kimura 3-parameter transition matrices are diagonal. Therefore, the resulting matrix $\hat{M}^1 \cdot \hat{M}^2$ is also diagonal. According to this, $\hat{p}_{x_1x_2++}$ is different from 0 only when $x_1 = x_2$. We want to define a marginal for all the possible combinations of nucleotides in the different branches, so we need to adapt the concept of marginal.

In our generalization, the marginals will be weighted by the $\chi_N$ characters, where $N \in \{A,C,G,T\}$.

**Notation 3.14.** Let $p = (p_{x_1x_2x_3x_4})_{x_1,x_2,x_3,x_4 \in \Sigma}$ be a probability distribution arising from the tree topology $12|34$. We write

$$p_{x_1x_2\theta_N+} = \sum_{x_3} \chi_N(x_3) \sum_{x_4} p_{x_1x_2x_3x_4}$$

where $N \in \{A,C,G,T\}$.

Similarly, we define $p_{x_1x_2+\theta_N}$, $p_{\theta_N+x_3x_4}$ and $p_{+\theta_Nx_3x_4}$.

Now, we express this tensor in terms of the transition matrices of each branch and compute its Fourier transform.

**Proposition 3.15.**

a) For each pair $(x_1, x_2)$, we have

$$p_{x_1x_2\theta_N+} = \frac{1}{4} (\hat{M}^1 \cdot \Phi_N \cdot \hat{M}^2)_{x_1,x_2}$$

where

$$\Phi_N = \begin{pmatrix}
\chi_N(A) & 0 & 0 & 0 \\
0 & \chi_N(C) & 0 & 0 \\
0 & 0 & \chi_N(G) & 0 \\
0 & 0 & 0 & \chi_N(T)
\end{pmatrix}.$$

It follows that

$$p_{\cdot \chi_N+} = \frac{1}{4} M^1 \Phi_N M^2$$
b) The Fourier coordinates of $(p_{x_1 x_2 \theta_N^{+}})_{x_1,x_2}$ are given by

$$\hat{p}_{x_1 x_2 \chi_N^{+}} = m^3_N m^2_{N} m^1_{x_1} m^2_{x_2} \hat{\Pi}_{x_1 + x_2 + N}$$

where

$$\hat{\Pi}_{i+j+N} = \begin{cases} 0, & \text{if } i+j \neq N; \\ 1/4, & \text{if } i+j = N. \end{cases}$$

This proposition can be extended to the rest of combinations of the leaves in the tree.

**Proof.**

$$p_{x_1 x_2 \theta_N^{+}} = \sum_{x_3} \chi_N(x_3) \sum_{x_4} p_{x_1 x_2 x_3 x_4}$$

$$= \sum_{x_3} \chi_N(x_3) \sum_{x_4} \sum_{r,s} \pi_r M^1_{r,x_3} M^2_{r,s,x_3} M^3_{s,x_4}$$

$$= \frac{1}{4} \sum_{r,s,x_3} \chi_N(x_3) M^1_{r,x_3} M^2_{r,s} M^3_{s,x_3} \sum_{x_4} M^4_{s,x_4}$$

$$= \frac{1}{4} \sum_{r,s} M^1_{r,x_3} M^2_{r,s} M^3_{s,x_3} \chi_N(x_3) M^4_{s,x_3}$$

where we have used that the sum of all the elements of the same row in a transition matrix is 1. By Lemma 3.8, we know that

$$\sum_{x_3} \chi_N(x_3) M^3_{s,x_3} = \chi_N(s)m^3_N$$

Therefore, continuing with the previous expression

$$p_{x_1 x_2 \theta_N} = \frac{1}{4} \sum_{r,s} M^1_{r,x_1} M^2_{r,x_2} M^3_{s,x_3} \chi_N(s)m^3_N = \frac{m^3_N}{4} \sum_{r} M^1_{r,x_1} M^2_{r,x_2} \sum_{s} M^e_{r,s} \chi_N(s)$$

$$= \frac{m^3_N}{4} \sum_{r} M^1_{r,x_1} M^2_{r,x_2} m^e_{N} \chi_N(r) M^1_{r,x_1} M^2_{r,x_2}$$

$$= \frac{m^3_N}{4} m^e_{N} (M^{1t} \Phi_N M^2)_{x_1,x_2}$$

where $\Phi_N = \text{diag}(\chi_N(A), \chi_N(C), \chi_N(G), \chi_N(T))$. 

24
Applying the Fourier transform to the previous expression, we have that

\[
\hat{p}_{x_1 x_2 \theta_N^+} = Q^{-1} p_{x_1 x_2 \theta_N^+} + Q = \frac{1}{4} m_N^2 m_N^c \left( Q^{-1} M^1 Q \right) \left( Q^{-1} \Phi_N Q \right) \left( Q^{-1} M^2 Q \right)
\]

\[
= \frac{1}{4} m_N^2 m_N^c M^1 \left( Q^{-1} \Phi_N Q \right) M^2
\]

Now we analyze the part \( Q^{-1} \Phi_N Q \):

\[
Q^{-1} \Phi_N Q = \frac{1}{4} \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 \\
1 & -1 & -1 & 1
\end{pmatrix}
\begin{pmatrix}
\chi_N(A) & 0 & 0 & 0 \\
0 & \chi_N(C) & 0 & 0 \\
0 & 0 & \chi_N(G) & 0 \\
0 & 0 & 0 & \chi_N(T)
\end{pmatrix}
\begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 \\
1 & -1 & -1 & 1
\end{pmatrix}
\]

\[
= \frac{1}{4} \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 \\
1 & -1 & 1 & -1 \\
1 & -1 & -1 & 1
\end{pmatrix}
\begin{pmatrix}
\chi_N(A) & \chi_N(A) & \chi_N(A) & \chi_N(A) \\
\chi_N(C) & \chi_N(C) & -\chi_N(C) & -\chi_N(C) \\
\chi_N(G) & -\chi_N(G) & \chi_N(G) & -\chi_N(G) \\
\chi_N(T) & -\chi_N(T) & -\chi_N(T) & \chi_N(T)
\end{pmatrix}
\]

\[
= \frac{1}{4} \begin{pmatrix}
\sum_t \chi_{A+A}(t) \chi_N(t) & \sum_t \chi_{A+C}(t) \chi_N(t) & \sum_t \chi_{A+G}(t) \chi_N(t) & \sum_t \chi_{A+T}(t) \chi_N(t) \\
\sum_t \chi_{C+A}(t) \chi_N(t) & \sum_t \chi_{C+C}(t) \chi_N(t) & \sum_t \chi_{C+G}(t) \chi_N(t) & \sum_t \chi_{C+T}(t) \chi_N(t) \\
\sum_t \chi_{G+A}(t) \chi_N(t) & \sum_t \chi_{G+C}(t) \chi_N(t) & \sum_t \chi_{G+G}(t) \chi_N(t) & \sum_t \chi_{G+T}(t) \chi_N(t) \\
\sum_t \chi_{T+A}(t) \chi_N(t) & \sum_t \chi_{T+C}(t) \chi_N(t) & \sum_t \chi_{T+G}(t) \chi_N(t) & \sum_t \chi_{T+T}(t) \chi_N(t)
\end{pmatrix}
\]

\[
= \frac{1}{4} \begin{pmatrix}
\sum_i \chi_{i+j+N}(t) \\
\sum_i \chi_{i+j+N}(t) \\
\sum_i \chi_{i+j+N}(t) \\
\sum_i \chi_{i+j+N}(t)
\end{pmatrix}
\]

\[
= \begin{cases}
4/4 = 1 & \text{if } i + j + N = 0 \iff i + j = N \\
0 & \text{if } i + j + N \neq 0 \iff i + j \neq N
\end{cases}
\]

Therefore,

\[
\hat{p}_{x_1 x_2 \chi_N} = m_N^2 m_N^c m_{x_1} m_{x_2} \hat{\Pi}_{x_1+x_2+N}
\]

where

\[
\hat{\Pi}_{x_1+x_2+N} = \begin{cases}
0 & \text{if } x_1 + x_2 \neq N \\
1/4 & \text{if } x_1 + x_2 = N
\end{cases}
\]

\[
\square
\]
Since the goal in this section is to find a term to normalize the entries of the matrix, now
that we have defined a marginal for each block, we will use these results to simplify our
block matrix.

**Theorem 3.16.** Let $F$ be Fourier flattening relative to $1234$ of a joint distribution $p$ on
4 leaves. Then, using the generalized marginals, we can rescale the rows and columns of
$F$ obtaining a new matrix $\tilde{F}$ such that:

(i) $\text{rank } \tilde{F} = \text{rank } F$

(ii) if $p$ comes from the 4-leaved tree $1234$, then all block entries of $\tilde{F}$ are equal to 1.

**Proof.** Consider a non-zero element of the matrix $F$, $q_{x_1x_2x_3x_4}$ with $x_1 + x_2 + x_3 + x_4 = 0$.
We have seen that it has the following form:

$$q_{x_1x_2x_3x_4} = m_{x_1}^1m_{x_2}^2m_{x_1+x_2}^3m_{x_3}^4m_{x_4}^4.$$

According to the Proposition 3.15, we know that if $x_1 + x_2 = x_3 + x_4 = N$

$$\hat{p}_{x_1x_2}\theta_{N+} = \frac{1}{4}m_N^3m_N^e m_{x_1}^1m_{x_2}^2$$

$$\hat{p}_{x_3x_4}\theta_{N+} = \frac{1}{4}m_N^1m_N^e m_{x_3}^4m_{x_4}^4$$

Thus, we divide each row $(x_1, x_2)$ of $F$ by $\hat{p}_{x_1x_2}\theta_{N+}$, where $N = x_1 + x_2$, and each column
$(x_3, x_4)$ of $F$ by $\hat{p}_{\theta_{M+x_3x_4}}$, where $M = x_3 + x_4$.

$$\frac{q_{x_1x_2x_3x_4}}{\frac{1}{4}m_N^3m_N^e m_{x_1}^1m_{x_2}^2} = \frac{\frac{1}{16}m_{x_1}^1m_{x_2}^2m_N^3m_{x_3}^4m_{x_4}^4}{\frac{1}{16}m_N^3m_N^e m_{x_1}^1m_{x_2}^2m_N^1m_N^e m_{x_3}^4m_{x_4}^4} = \frac{16}{m_N^3m_N^1m_N^e}$$

Now, multiplying again the rows by $\frac{1}{4}(p_{x_1+x_3+})_{N,N} = \frac{1}{16}m_N^1m_N^3m_N^3$, we get

$$\frac{16}{m_N^3m_N^1m_N^e} \cdot \frac{1}{16}m_N^1m_N^3m_N^3 = 1$$

Therefore, we have obtained a block matrix whose elements from each block are all equal
to 1.
This proves (ii) and (i) is trivial.

\[ \square \]

**Remark 3.17.** *Note that this is a stronger condition than having rank = 1.*

### 3.4 Phylogenetic reconstruction

In order to decide which tree best fits a given alignment, we use the properties of the flattening matrix obtained by doing the Fourier transform and the adjusting process described in the previous section.

More precisely, we have studied the case when the tree has the topology $12|34$ and we have obtained a block diagonal matrix where all the elements in the same block have the same value. If the probability vector was not generated by the $12|34$ topology, then the flattening would not have rank 1 blocks (see [CFS10b]).

In addition, if the tree configuration is $13|24$ or $14|23$ we can do the Fourier transform, the flattening of the vector and the simplification process with the leaves (1,3) or (1,4) respectively instead of (1,2) for the rows, and (2,4) or (2,3) respectively instead of (3,4) for the columns, and we will also get a block diagonal matrix with that property.

In order to determine which of the three possible topologies fits the data better, we define a distance to the subspace generated by this theoretical matrix:

**Definition 3.18.** Let $B$ be a matrix in $\mathcal{M}_{4\times 4}$ and $V$ the subspace

\[
V = \left\{ \begin{pmatrix} n & n & n & n \\ n & n & n & n \\ n & n & n & n \\ n & n & n & n \end{pmatrix} \text{ such that } n \in \mathbb{R} \right\} \subseteq \mathcal{M}_{4\times 4}. 
\]

Then, we define the distance to the subspace as

\[
D_F(B, V) = \sqrt{\sum_{i,j} (B_{i,j} - m_a)^2}
\]

where $m_a$ is the arithmetic mean of the values of $B$.

However, our ideal matrix has all the entries equal to 1. Therefore, we can be more precise and define the distance to this matrix as follows:
**Definition 3.19.** Let $B$ be a matrix in $\mathcal{M}_{4 \times 4}$ and $H$ the matrix

$$H = \begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1
\end{pmatrix}$$

Then, we define the distance to $H$ as

$$D_F(B, H) = \sqrt{\sum_{i,j} (B_{i,j} - 1)^2}.$$ 

**Definition 3.20.** Let $B$ be the block diagonal matrix obtained after adjusting the flattening of the Fourier transform vector and $B_N$ for $N \in \{A, C, G, T\}$ each of the blocks. Then, we define the score of $B$ as

$$S(B) = \frac{D_F(B_A, V) + D_F(B_C, V) + D_F(B_G, V) + D_F(B_T, V)}{4}.$$ 

We define also the alternative score as

$$S_1(B) = \frac{D_F(B_A, H) + D_F(B_C, H) + D_F(B_G, H) + D_F(B_T, H)}{4}.$$ 

In order to determine the tree topology corresponding to a given alignment, we calculate the Fourier transform and arrange the flattening in terms of the coordinates for the three different topologies. After that, we adjust the matrix dividing by the Fourier transformation of the marginal at the corresponding leaves and compute the distance to the subspace (using Definition 3.18) or to the expected matrix (using Definition 3.19). The configuration with the lowest score will give us the topology of the tree corresponding with the alignment. It will be important in future work to decide which score, $S$ or $S_1$ is more suitable.
4 Conclusions

Throughout this project, we have achieved all the objectives we had proposed at the beginning. We have conceived a new method for phylogenetic reconstruction using the properties of the Fourier transform such that it is independent of the length of the branches at the leaves. Moreover, we have implemented this method in c++. More precisely,

- We have checked that the Fourier transform allows to write the flattening entries as monomials in the transformed parameters.

- We have checked that the Fourier transform decomposes the flattening as a block diagonal matrix with blocks of rank 1.

- We have proved that the marginals and certain generalizations of marginals allow the recovery of products of transformed parameters.

- We have proposed an adjustment of the flattening matrix using these marginal generalizations, so that the flattening becomes parameter independent.

- We have used this adjustment to propose a new phylogenetic reconstruction method.

However, there are still many things to continue working on. Our future work will focus on:

- Test this method on data simulated under Kimura 3-parameter model.

- Study possible generalization of our method to the general Markov model.

- Test it on data simulated under the general Markov model.

- Test it as an input method of quartet-based methods.

- Test it on real data.

- Write a research paper with the findings of this work.

We will continue working on this field to achieve the previous objectives with the support of the fellowship for collaboration in research given by the Ministerio de Educación, Cultura y Deporte.
References


