© REVISTA DE MATEMÁTICA: TEORÍA Y APLICACIONES 2023 **30**(1) : 71–87
CIMPA – UCR   ISSN: 1409-2433 (PRINT), 2215-3373 (ONLINE)
DOI: https://doi.org/10.15517/rmta.v30i1.50357

ADAPTATION OF THE CLOSEST TREE METHOD FOR A TWO STATE QUARTET TO JUKES-CANTOR TRIPOD TREES

ADAPTACIÓN DEL MÉTODO DEL ÁRBOL MÁS CERCANO SOBRE UN CUARTETO DE DOS ESTADOS A TRÍPODES TIPO JUKES-CANTOR

ERNESTO ÁLVAREZ GONZÁLEZ[*]

Received: 17/Mar./2022; Revised: 06/Sep./2022;
Accepted: 18/Oct./2022

[*] Universidad Complutense de Madrid, Departamento de Geometría, Álgebra y Topología, Madrid, España. E-Mail: eralva01@ucm.es

Abstract

We aim to fit 4-state sequences of DNA characters from three species to a tripod tree, whose evolutionary model is Jukes-Cantor. For this purpose, we adapt the closest tree method used in the fit of 2-state sequences coming from four species to a quartet, where the states are purines and pyrimidines and the evolutionary model is CFN. The adaptation requires a multi stage methodology called 'reduction process'. We take the frequencies of 2-state character patterns on the quartet as parameters and search for solutions to the fit.

Keywords: closest tree method; Hadamard conjugation; phylogenetic reconstruction; observed data fitting; reduction process; the Gröbner cover algorithm.

Resumen

Nuestro objetivo es adaptar secuencias de caracteres de ADN con cuatro estados provenientes de tres especies a un árbol filogenético tipo trípode, cuyo modelo evolutivo es Jukes-Cantor. Para ello, adaptamos el método del árbol más cercano utilizado en el ajuste de secuencias de 2 estados provenientes de cuatro especies a un cuarteto, donde los estados son purinas y pirimidinas y el modelo evolutivo es CFN. La adaptación requiere una metodología de múltiples etapas llamada 'proceso de reducción'. Tomamos las frecuencias de los patrones de caracteres de 2 estados en el cuarteto como parámetros y buscamos soluciones para el ajuste.

Palabras clave: método del árbol más cercano; conjugación de Hadamard; ajuste de datos observados; reconstrucción filogenética; proceso de reducción; algoritmo Gröbner cover.

Mathematics Subject Classification: 92D15.

1 Introduction

Phylogenetic reconstruction aims at finding the evolutionary history of a group of species. The ancestral relationships that give rise to the current species [9] can be tracked through diagrams called *phylogenetic trees* [4], whose nodes have a discrete random variable for observing the state (or character) of a genetic entity. Between nodes there can also be given a substitution model for estimating the probability of a state change.

A typical data set of departure in phylogenetic reconstruction is an alignment of characters. If we fix the phylogenetic tree, the next step is data fitting.

The main goal of [2] is to set conditions on the parameter space to fit the observed sequences of DNA characters from three species to the phylogenetic tree of figure 1 (*rooted tripod tree*), whose evolutionary model is Jukes-Cantor (JC model), under the *molecular clock condition*. Substitution models on the edges provide the parameters. However, their work limits to set just boundary conditions on the parameter space, leaving for future research the question of solutions in the interior, and if they exist, the uniqueness. The main goal of this article is to return to these two questions applied to the tripod trees in figures 1 and 2, whose molecular substitution model is Jukes-Cantor with and without the molecular clock condition, respectively.

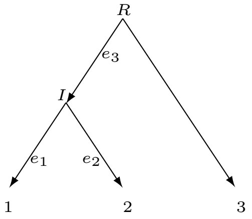

Figure 1: Example of a rooted tripod tree with edges e_1, e_2, e_3.

The data fitting technique that the authors in [2] follow is *maximum likelihood estimation*, but we use the version of the *closest tree method* that M.D. Hendy shows in [6] after the intermediate process that is explained in section 3.

Our data fitting method is complementary to that in [8], where the authors provide a version of the closest tree method for fitting the observed sequences of DNA characters from a set of species to an unrooted phylogenetic tree whose substitution model on the edges is a generalized version of Kimura Three Parameters.

2 Preliminaries

2.1 Spectral sequence spectrum on Jukes-Cantor tripod trees

To each node on a tripod tree, we associate a 4-state discrete random variable for observing adenines (A), guanines (G), cytosines (C) and thymines (T).

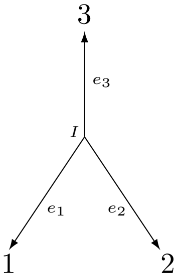

Figure 2: Example of a tripod tree with edges e_1, e_2, e_3.

We classify three kind of substitutions:

1. transitions ($A \leftrightarrow G, T \leftrightarrow C$);
2. type I transversions ($A \leftrightarrow T$, $G \leftrightarrow C$);
3. type II transversions ($A \leftrightarrow C$, $G \leftrightarrow T$).

To each edge e_j on the tripod tree, we associate the expected number of transitions, type I transversions and type II transversions as in [7]: let α_j, β_j and γ_j be the rates of transitions, type I transversions and type II transversions on e_j, respectively. If t_j is the time span, then $q_j(\alpha_j) = \alpha_j t_j$, $q_j(\beta_j) = \beta_j t_j$ and $q_j(\gamma_j) = \gamma_j t_j$ are the expected number of transitions, type I transversions and type II transversions, respectively. We call *q-parameters* to these expected numbers of substitutions. If $\alpha_j = \beta_j = \gamma_j$ and the probabilities of state change between nodes on e_j is determined by the matrix M_j below, then the tripod tree holds the Jukes-Cantor substitution model and is called a Jukes-Cantor tripod tree (JC tripod tree). One Jukes-Cantor tripod tree having a unique, common rate of substitution $\alpha = \beta = \gamma$ on the edges, shows the molecular clock condition and is called a MC Jukes-Cantor tripod tree.

$$M_j = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 1+3\exp(-4\alpha_j t_j) & 1-\exp(-4\alpha_j t_j) & 1-\exp(-4\alpha_j t_j) & 1-\exp(-4\alpha_j t_j) \\ 1-\exp(-4\alpha_j t_j) & 1+3\exp(-4\alpha_j t_j) & 1-\exp(-4\alpha_j t_j) & 1-\exp(-4\alpha_j t_j) \\ 1-\exp(-4\alpha_j t_j) & 1-\exp(-4\alpha_j t_j) & 1+3\exp(-4\alpha_j t_j) & 1-\exp(-4\alpha_j t_j) \\ 1-\exp(-4\alpha_j t_j) & 1-\exp(-4\alpha_j t_j) & 1-\exp(-4\alpha_j t_j) & 1+3\exp(-4\alpha_j t_j) \end{pmatrix} \end{array}.$$

Let $\chi(j) = i$ mean that the j-th node a the tripod tree shows DNA character $i \in \{A, G, C, T\}$. Then $\chi(j) \to \chi(k)$ stands for a substitution. In table 2, we use the integers 0 through 3 to mean the kind of substitution: 0 for no substitution; 1 for transitions; 2 for type I transversions; 3 for type II transversions.

An alignment of 4-state sequences 1, 2 and 3 is a 3-row list of DNA characters A, C, G and T (see table 1 as an example). Each column observes

Table 1: Alignment of three nucleotide sequences with ten sites.

Character	1	2	3	4	5	6	7	8	9	10
$\chi(1)$	C	C	A	T	T	G	A	A	G	A
$\chi(2)$	A	C	A	G	T	A	G	T	G	T
$\chi(3)$	A	C	A	G	C	A	A	T	G	T

Table 2: Substitutions from the third row in table 1.

Substitution	1	2	3	4	5	6	7	8	9	10
$\chi(3) \to \chi(1)$	3	0	0	3	1	1	0	$T \to A$ =2	0	2
$\chi(3) \to \chi(2)$	0	0	0	0	1	0	1	$T \to T$ =0	0	0

DNA characters on leaves in the tripod tree at a site. These triplets are called *character patterns*. Given the alignment, there is a frequency distribution of character patterns. Characters in rows 1 and 2 in table 1 can be obtained as substitutions from the character at the third row. Then pairs in columns of table 2 are called *substitution patterns*. Therefore there is also a frequency distribution of substitution patterns.

Following [2], substitution patterns can be represented as pairs of subsets of $\{1, 2\}$:

The ordered pair (A, B) is the substitution pattern such that

1. $A \backslash B$: set of species obtained by transitions from the third row character.

2. $B \backslash A$: set of species obtained by type I transversions from the third row character.

3. $A \cap B$: set of species obtained by type II transversions from the third row character.

4. $\{1, 2\} \backslash (A \cup B)$: set of species sharing the same third row character.

For example, substitution patterns in sites 1, 7 and 8 in table 2 are represented as $(\{1\}, \{1\})$, $(\{2\}, \emptyset)$ and $(\emptyset, \{1\})$, respectively.

For the alignment of three species, there can be sixteen substitution patterns at most. Their frequencies can be located within a 4×4 matrix, where index row indicates the left entry of the ordered pair for the corresponding substitution pattern whereas column index indicates the right entry of the ordered pair, according to the notation used in [2]. This matrix is called the *Spectral*

Sequence Spectrum. The virtue of this matrix is that it can be computed in terms of the q-parameters via Hadamard conjugation [7].

We show next the Spectral Sequence Spectrum for Jukes-Cantor tripod trees and for MC Jukes-Cantor tripod trees as in [2]:

Spectral Sequence Spectrum, P_{JC}, for a Jukes-Cantor tripod tree

$$P_{JC} = \begin{array}{c} \\ \emptyset \\ \{1\} \\ \{2\} \\ \{1,2\} \end{array} \begin{array}{c} \begin{array}{cccc} \emptyset & \{1\} & \{2\} & \{1,2\} \end{array} \\ \begin{pmatrix} a_0 & a_1 & a_2 & a_3 \\ a_1 & a_1 & a_4 & a_4 \\ a_2 & a_4 & a_2 & a_4 \\ a_3 & a_4 & a_4 & a_3 \end{pmatrix} \end{array}, \tag{1}$$

where,

$$\begin{aligned}
a_0 &= \frac{1}{16} + \frac{3}{16}yz + \frac{3}{16}xz + \frac{3}{16}xy + \frac{6}{16}xyz, \\
a_1 &= \frac{1}{16} + \frac{3}{16}yz - \frac{1}{16}xz - \frac{1}{16}xy - \frac{2}{16}xyz, \\
a_2 &= \frac{1}{16} - \frac{1}{16}yz + \frac{3}{16}xz - \frac{1}{16}xy - \frac{2}{16}xyz, \\
a_3 &= \frac{1}{16} - \frac{1}{16}yz - \frac{1}{16}xz + \frac{3}{16}xy - \frac{2}{16}xyz, \\
a_4 &= \frac{1}{16} - \frac{1}{16}yz - \frac{1}{16}xz - \frac{1}{16}xy + \frac{2}{16}xyz,
\end{aligned} \tag{2}$$

with $x = e^{-4q_1}$, $y = e^{-4q_2}$ and $z = e^{-4q_3}$ being the *pathset variables* as in [2].

Spectral Sequence Spectrum, P_{MCJC}, for an MC Jukes-Cantor rooted tripod tree

$$P_{MCJC} = \begin{array}{c} \\ \emptyset \\ \{1\} \\ \{2\} \\ \{1,2\} \end{array} \begin{array}{c} \begin{array}{cccc} \emptyset & \{1\} & \{2\} & \{1,2\} \end{array} \\ \begin{pmatrix} a_0 & a_1 & a_1 & a_3 \\ a_1 & a_1 & a_4 & a_4 \\ a_1 & a_4 & a_1 & a_4 \\ a_3 & a_4 & a_4 & a_3 \end{pmatrix} \end{array}, \tag{3}$$

where,

$$\begin{aligned}
a_0 &= \frac{1}{16} + \frac{3}{16}y^2 + \frac{3}{8}yz + \frac{3}{8}y^2z,\\
a_1 &= \frac{1}{16} - \frac{1}{16}y^2 + \frac{1}{8}yz - \frac{1}{8}y^2z,\\
a_3 &= \frac{1}{16} - \frac{1}{16}y^2 + \frac{1}{8}yz - \frac{1}{8}y^2z,\\
a_4 &= \frac{1}{16} - \frac{1}{16}y^2 - \frac{1}{8}yz + \frac{1}{8}y^2z,
\end{aligned} \tag{4}$$

with $y = e^{-4q_2}$ and $z = e^{-4q_3}$ being the pathset variables.

2.2 Closest tree method on a two state quartet

In 1989, M. D. Hendy introduced the *closest tree method* [6]. This is a method for fitting a finite set of 2-state sequences (where purines and pyrimidines are the states) to an unrooted phylogenetic tree under the CFN model. We restrict our attention to the *quartet* of figure 3. To each node on the quartet, we associate a two-state discrete random variable for observing purines (X) and pyrimidines (Y). We put emphasis only on the elements that we use in section 3.

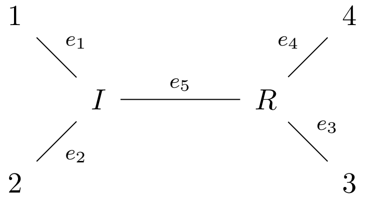

Figure 3: An example of a quartet with edges e_1 through e_5.

To each edge e_i on the quartet, we associate the expected number of substitutions as in [6]: let λ_j be the rate of substitutions on e_i. If t_i is the time span, then $q_i' = \lambda t_i$ is the number of substitutions (or edge length).

We fix a unique rate of substitutions, λ, for the edges on the quartet. CFN model on the quartet implies stochastic matrices N_j, as below, for computing the probabilities of state change between nodes on the corresponding edges.

$$N_j = \begin{array}{c c} & \begin{array}{cc} X & Y \end{array} \\ \begin{array}{c} X \\ Y \end{array} & \begin{pmatrix} \frac{1+\exp(-2q'_j)}{2} & \frac{1-\exp(-2q'_j)}{2} \\ \frac{1-\exp(-2q'_j)}{2} & \frac{1+\exp(-2q'_j)}{2} \end{pmatrix} \end{array}.$$

An alignment of 2-state sequences 1, 2, 3 and 4 is a 4-row list of characters X and Y. Each column is a sequence of purines and pyrimidines distributed at leaves on the quartet. Each column indicates a different site. These sequences are classyfied into 8 *bipartitions*: $A_1 = \{1\}$, $A_2 = \{1, 2\}$, $A_3 = \{1, 3\}$, $A_4 = \{1, 2, 3\}$, $A_5 = \{1, 4\}$, $A_6 = \{1, 2, 4\}$, $A_7 = \{1, 3, 4\}$ and $A_8 = \{1, 2, 3, 4\}$. Each of these bipartitions indicates which leaves on the quartet share the same character on leaf 1. Given the alignmnet, there is a frequency distribution of bipartitions that can be kept into the vector $\vec{s} = (s_1, s_2, \ldots s_8)$. Let ρ be the vector (5) where $H = (h_{m,i})$ is the Hadamard matrix (6) and the superscript t stands for the transpose operation.

$$\rho = H^t \vec{s}^{\,t}, \tag{5}$$

where

$$H = \begin{pmatrix} 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}. \tag{6}$$

For j in $\{1, 2, \ldots 8\}$ do

$$r_j := -\frac{1}{2}\log(\rho_j). \tag{7}$$

Formula (7) sums the edge lengths over non overlapping, connected paths in the quartet of figure 3. For example, $r_6 = -\frac{1}{2}\log(\rho_6) = q'_2 + q'_5 + q'_4$ and $r_8 = -\frac{1}{2}\log(\rho_8) = q'_1 + q'_2 + q'_3 + q'_4$.

Equations (5) and (7) establish a one to one correspondence between vectors $\vec{s}$ and $\vec{r}$.

A cut on e_j produces the bipartition of those leaves to its side that includes leaf 1. For example, a cut in e_5 produces A_2, a cut in e_3 produces A_6.

Let $e_j \leftrightarrow A_m$ mean this correspondence. Let K be the 8×5 matrix (8), with entries $k_{i,j} = \frac{1-h_{m,i}}{2}$, where $e_j \leftrightarrow A_m$.

$$K = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}. \tag{8}$$

Let $\vec{q'} = (q'_1, q'_2, q'_3, q'_4, q'_5)$ be the vector of edge lengths on the quartet. The Moore-Penrose inverse K^+ to K is:

$$K^+ = \begin{pmatrix} 0 & \frac{1}{3} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{6} & \frac{1}{6} \\ 0 & \frac{1}{3} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{6} & \frac{1}{6} \\ 0 & -\frac{1}{6} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{3} & \frac{1}{6} \\ 0 & -\frac{1}{6} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{3} & \frac{1}{6} \\ 0 & -\frac{1}{6} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{6} & -\frac{1}{3} \end{pmatrix}. \tag{9}$$

According to [6], it follows:

$$\vec{r} = K\vec{q}^t; \tag{10}$$

$$\vec{q} = K^+\vec{r}^t. \tag{11}$$

Given an observed vector of frequencies $\vec{s_o}$ of bipartitions A_1 through A_8 for the alignment of 2-state sequences 1, 2, 3 and 4, M. D. Hendy provides in [6] a criterion for selecting the quartet of figure 3 as the ‘best fit’ tree: let $\vec{r_o}$ be in corresponce to $\vec{s_o}$ according to the Equations (5) and (7). Let $\mathcal{R}$ be the set of possible vectors $\vec{r}$ that can be derived from the quartet in agreement to the Equations (5) and (7). Then $KK^+\vec{r_o}$ is the closest point to $\vec{r_o}$ in $\mathcal{R}$. Said differently, the distance from $\vec{r_o}$ to $\mathcal{R}$ is $\|KK^+\vec{r_o} - \vec{r_o}\|$ (cf. [1]). In section 4, we do this and take the entries of $\vec{s_o}$ as parameters.

2.3 The Gröbner cover algorithm

A. Montes published in 2010 the Gröbner cover algorithm for analyzing polynomial systems with parameters. A good reference for introducing Gröbner bases is [3].

We next give a short view of the Gröbner cover algorithm.

Let $\bar{a} = \{a_1, a_2, \cdots, a_m\}$ be a set of parameters and let $Z = \{z_1, z_2, \cdots, z_n\}$ be a set of variables. We define $\mathbb{R}[\bar{a}]\mathbb{R}[Z]$ as the ring of polynomials with variables in Z and coefficients in $\mathbb{R}[\bar{a}]$.

Let $\{p_1(\bar{a}, Z), p_2(\bar{a}, Z), \ldots, p_r(\bar{a}, Z)\}$ be a set of polynomials in $\mathbb{R}[\bar{a}]\mathbb{R}[Z]$. For each assignment of real values to the parameters, $\zeta_{\bar{a}} : \bar{a} \to \mathbb{R}$, the goal is describing the complex algebraic variety $\mathscr{V}(I_{\zeta_{\bar{a}}}) \subset \mathbb{C}^n$ of the ideal $I_{\zeta_{\bar{a}}} = < p_1(\zeta_{\bar{a}}, Z), p_2(\zeta_{\bar{a}}, Z), \ldots p_r(\zeta_{\bar{a}}, Z) > \subset \mathbb{R}[Z]$.

Let $\succ_Z$ be a monomial order for Z as in [3]. For $lpp(I_{\zeta_{\bar{a}}})$, we mean the set of leading power products associated to each polinomial in $I_{\zeta_{\bar{a}}}$ with respect to $\succ_Z$.

The Gröbner cover of $\mathbb{C}^m$ with respect to $(\succ_Z, I_{\bar{a}})$ is a set of pairs $\{(S_1, B_1), (S_2, B_2), \cdots, (S_r, B_r)\}$ having the following properties:

- The sets S_i are locally closed with respecto to the Zariski topology (differences of closed sets), pairwise disjoint, whose union is $\mathbb{C}^m$. These sets are called *segments*.

- For any two distinct assignments $\zeta_{\bar{a}}, \zeta'_{\bar{a}}$ to the parameters, the sets of leading power products $lpp(I_{\zeta_{\bar{a}}})$ and $lpp(I_{\zeta'_{\bar{a}}})$ agree on each intersection $\mathbb{R}^m \cap S_i$ for every $i \in \{1, 2, \ldots r\}$.

- For each $i \in \{1, 2, \ldots r\}$, $B_i \subset \mathscr{O}(S_i)[X]$ is a finite, reduced Gröbner bases over S_i, where $\mathscr{O}(S_i)$ is the ring of regular functions on the segment.

The Gröbner cover algorithm is implemented in Singular and is freely available in `https://www.singular.uni-kl.de/`. Gert's book in [5] provides an accessible, full guide to the use of Singular in the context of commutative algebra.

3 The reduction process on tripod trees

For phylogenetic trees whose nodes have a 4-state discrete random variable, the *reduction process* consists in reading purines and pyrimidines instead of adenines, citosines, guanines and thymines.

In order to apply the closest tree method to a rooted tripod tree as that of figure 1, whose nodes have a 4-state discrete variable, we insert a new edge from its root R. In case of a tripod tree as that of figure 2, we could insert a new edge starting from a middle point on the third edge e_3. In either case, we call the resulting phylogenetic tree as the *tripod's quartet extension*. Then we read purines (X) from the states A and G, and pyrimidines (Y) from the states C and T.

Lemma 1 *Consider an alignment of 2-state sequences 1, 2, 3 and 4, whose phylogenetic tree is the quartet of figure 3. By marginalizing the distribution frequencies of character patterns over the fourth leaf, we get the distribution* $s' : \{X, Y\}^3 \to [0, 1]$ *of character patterns on the tripod component with leaves 1, 2 and 3. Let* $\vec{s} = (s_1, s_2, \ldots s_8)$ *be the vector of frequencies of bipartitions associated to the given alignment. Then the following relations hold:*

$$s_4 + s_8 = s'_{XXX} + s'_{YYY}, \tag{12}$$

$$s_2 + s_6 = s'_{XXY} + s'_{YYX}, \tag{13}$$

$$s_1 + s_5 = s'_{XYY} + s'_{YXX}, \tag{14}$$

$$s_3 + s_7 = s'_{XYX} + s'_{YXY}. \tag{15}$$

Proof. We explain just the equality (12). The rest of these follow similarly.

Bipartitions $A_2 = \{1, 2\}$ and $A_6 = \{1, 2, 4\}$ differ in the fourth leaf. If $\chi(1) = X$, then A_2 observes $XXYY$ and A_6 observes $XXYX$. These two events are considered by s'_{XXY}. If $\chi(1) = Y$, then A_2 observes $YYXX$ and A_6 observes $YYXY$. These two events are considered by s'_{YYX}. ■

4 Data fitting

Proposition 1 *Let* $\vec{s_o}$ *be the observed vector of frequencies of bipartitions for an alignment of 2-state sequences as it was considered at the end of section 2.2. Take these frequencies as parameters. Let* $\vec{r_o}$ *be in correspondence to* $\vec{s_o}$ *according to the Equations (5) and (7). Let* $\mathcal{R}$ *be the set of vectors* $\vec{r}$ *derived from the quartet of figure 3. Express* $\vec{r_o}$ *in terms of the parameters according to the Equations (5) and (7). Then there exist non-extreme conditions on the parameters guarantying a best fit to the quartet.*

Proof. We do $\rho = H^t \vec{s_o}^t$ as in (5) and get linear combinations of the parameters. We take logarithms as in (7) for each index j in ρ and get the vector $\vec{r_o}$. We equate to zero each term in $KK^+\vec{r} - \vec{r}$ to establish conditions on the parameters for selecting the quartet of figure 3 as the best fit. We reduce each equation involving more than one logarithm to an equation with just one logarithm. We exponentiate every resulting equation and look for an appropriate solution to the given system in which there is no extreme assumptions as when two or more different parameters equal to each other or as when some parameters equal to zero. The conditioning equations on the parameters are the following:

$$\begin{aligned}-(s_3+s_4+s_5+s_6)(s_4-s_6)s_1 = {} & s_3^2s_4+s_3^2s_8+s_3s_4^2+s_3s_4s_5+\\ & s_3s_4s_6+s_3s_4s_8-s_3s_5s_6+s_3s_6s_8-s_4s_5s_6-s_4s_5s_8-\\ & s_5^2s_6-s_5^2s_8-s_5s_6^2-s_5s_6s_8-s_3s_4+s_5s_6,\end{aligned} \tag{16}$$

$$\begin{aligned}-(s_3+s_4+s_5+s_6)s_2 = {} & s_3s_4+s_3s_6+s_3s_8+s_4^2+s_4s_5+2s_4s_6+\\ & s_4s_8+s_5s_6+s_5s_8+s_6^2+s_6s_8-s_4-s_6,\end{aligned} \tag{17}$$

$$\begin{aligned}(s_3+s_4+s_5+s_6)(s_4-s_6)s_7 = {} & s_3^2s_6+s_3^2s_8-s_3s_4s_5+s_3s_4s_6+\\ & s_3s_4s_8+s_3s_5s_6+s_3s_6^2+s_3s_6s_8-s_4^2s_5-s_4s_5^2-s_4s_5s_6-\\ & s_4s_5s_8-s_5^2s_8-s_5s_6s_8-s_3s_6+s_4s_5.\end{aligned} \tag{18}$$

■

Proposition 2 *Let Ω be an alignment of 4-state sequences 1, 2, 3 and 4; and let Ω_Q be the 2-state alignment after the reduction process. Assume for Ω_Q the quartet of figure 3 as a model of evolution. Let s' be the distribution of frequencies for the quartet's tripod component as in Lemma 1. Let $\vec{s} = (s_1, s_2, \dots s_8)$ be the vector of frequencies of bipartitions on Ω_Q. Assume the conclusions of Lemma 1. Let P be the Spectral Sequence Spectrum for the given tripod component. Then the following conditions are true:*

- $s'_{XXX} + s'_{YYY} =$ *sum of $\emptyset$-column entries in P;*
- $s'_{XYY} + s'_{YXX} =$ *sum of $\{1\}$-column entries in P;*

- $s'_{XYX} + s'_{YXY} =$ *sum of* $\{2\}$*-column entries in* P*;*
- $s'_{XXY} + s'_{YYX} =$ *sum of* $\{1,2\}$*-column entries in* P*.*

Proof. We just prove the first equality as the others follow similarly.

Character patterns $[AAA]^t$, $[AAG]^t$, $[AGA]^t$, $[AGG]^t$, $[GAA]^t$, $[GAG]^t$, $[GGA]^t$, $[GGG]^t$ reduce to $[XXX]^t$; character patterns $[CCC]^t$, $[CCT]^t$, $[CTC]^t$, $[CTT]^t$, $[TCC]^t$, $[TCT]^t$, $[TTC]^t$, $[TTT]^t$ reduce to $[YYY]^t$.

All these character patterns distribute on the $\emptyset$-column of matrix P:

- $(\emptyset, \emptyset) = \{[AAA]^t, [CCC]^t, [GGG]^t, [TTT]^t\}$;
- $(\{1\}, \emptyset) = \{[AGG]^t, [GAA]^t, [CTT]^t, [TCC]^t\}$;
- $(\{2\}, \emptyset) = \{[AGA]^t, [GAG]^t, [CTC]^t, [TCT]^t\}$;
- $(\{1,2\}, \emptyset) = \{[AAG]^t, [GGA]^t, [CCT]^t, [TTC]^t\}$.

Then, the probability of occurrence of the event $\{[XXX]^t, [YYY]^t\}$ at leaves on the tripod component is the sum of $\emptyset$-column entries in P. ■

4.1 Jukes-Cantor tripod tree

Corolary 1 *In view of Proposition 2 and the entries in matrix (1), the equalities in Lemma 1 become*

$$\begin{aligned}
&\frac{1}{4} + \frac{1}{4}yz + \frac{1}{4}xz + \frac{1}{4}xy - (s_4 + s_8) = 0;\\
&\frac{1}{4} - \frac{1}{4}yz - \frac{1}{4}xz + \frac{1}{4}xy - (s_2 + s_6) = 0;\\
&\frac{1}{4} + \frac{1}{4}yz - \frac{1}{4}xz - \frac{1}{4}xy - (s_1 + s_5) = 0;\\
&\frac{1}{4} - \frac{1}{4}yz + \frac{1}{4}xz - \frac{1}{4}xy - (s_3 + s_7) = 0.
\end{aligned} \tag{19}$$

Theorem 3 *The Gröbner cover algorithm applied to the System (19), where* $s_1, s_2, \ldots, s_8$ *are parameters satisfying relations (16) through (18) and* x, y, z *are the pathset variables, decomposes the parameter space* $\mathbb{R}^8$ *into segments such that just one of them is biologically meaningful, for which the corresponding canonical Gröbner bases has at most a unique solution* (x, y, z) *with* $0 < x < 1, 0 < y < 1, 0 < z < 1$.

Proof. The Gröbner cover algorithm produces the unique meaningful segment $\mathscr{C}_{11} = \mathbb{R}^8 \setminus \{\mathscr{V}(2s_3s_4 + 2s_3s_8 + 2s_4^2 + 2s_4s_5 + 2s_4s_6 + 2s_4s_8 - s_4 + 2s_5s_8 + 2s_6s_8 - s_6) \cup \mathscr{V}(s_4 - s_6) \cup \mathscr{V}(s_3 - s_4 - s_5 + s_6) \cup \mathscr{V}(s_3 + s_4 - s_5 - s_6) \cup \mathscr{V}(s_3 - s_4 + s_5 - s_6) \cup \mathscr{V}(s_3 + s_4 + s_5 + s_6)\}$, whose Canonical gröbner bases has the following generators:

1. $g_{11} = A_1A_2y + B_1B_2z$,

2. $g_{12} = A_2C_1x - B_1B_2z$,

3. $g_{13} = {B_1}^2B_2E_1(z)^2 - A_1{A_2}^2C_1$,

where $A_1 = -s_6 + s_3 + s_4 - s_5$, $A_2 = 2s_3s_4 + 2s_3s_8 + 2s_4^2 + 2s_4s_5 + 2s_4s_6 + 2s_4s_8 + 2s_5s_8 + 2s_6s_8 - s_4 - s_6$, $B_1 = s_4 - s_6$, $B_2 = -s_4 + s_3 + s_5 - s_6$, $C_1 = s_6 + s_3 - s_4 - s_5$ and $E_1 = s_3 + s_4 + s_5 + s_6$.

As in section 2.1, $x = \exp(-4q_1)$, $y = \exp(-4q2)$ and $z = \exp(-4q3)$ are the pathset variables, where q_1, q_2, q_3 are the q-parameters for the Jukes-Cantor tripod tree of figure 2. It makes sense for them to archive the restrictions $0 < x < 1, 0 < y < 1, 0 < z < 1$.

From the generator g_{13}, it is clear that condition $0 < z^2 < 1$ occurs in four cases:

1. $A_1 > 0, B_2 > 0, C_1 > 0$ and ${B_1}^2B_2E_1 - A_1{A_2}^2C_1 > 0$;

2. $A_1 > 0, B_2 < 0, C_1 < 0$ and $-{B_1}^2B_2E_1 + A_1{A_2}^2C_1 > 0$;

3. $A_1 < 0, B_2 > 0, C_1 < 0$ and ${B_1}^2B_2E_1 - A_1{A_2}^2C_1 > 0$;

4. $A_1 < 0, B_2 < 0, C_1 > 0$ and $-{B_1}^2B_2E_1 + A_1{A_2}^2C_1 > 0$.

From the generator g_{12}, $0 < x^2 = \frac{A_1B_2}{C_1E_1} < 1$ if $C_1E_1 - A_1B_2 > 0$ in correspondence to the previous cases first and fourth or if $-C_1E_1 + A_1B_2 > 0$ in correspondence to the cases second and third, respectively.

From the generator g_{11}, $0 < y^2 = \frac{B_2C_1}{A_1E_1} < 1$ if $A_1E_1 - B_2C_1 > 0$ in correspondence to the previous cases first and second or if $-A_1E_1 + B_2C_1 > 0$ in correspondence to the cases third and fourth, respectively. ■

4.2 MC Jukes-Cantor rooted tripod tree

Corolary 2 *In view of Proposition 2 and the entries in matrix (3), equalities in Lemma 1 become*

$$\begin{aligned} \frac{1}{4}+\frac{1}{2}yz+\frac{1}{4}y^2-(s_4+s_8)=0; \\ \frac{1}{4}-\frac{1}{2}yz+\frac{1}{4}y^2-(s_2+s_6)=0; \\ \frac{1}{4}-\frac{1}{4}y^2-(s_1+s_5)=0; \\ \frac{1}{4}-\frac{1}{4}y^2-(s_3+s_7)=0. \end{aligned} \tag{20}$$

Theorem 4 *The Gröbner cover algorithm applied to the System (20), where* $s_1, s_2, \ldots, s_8$ *are parameters satisfying all relations (16) through (18) and* y, z *are variables, decomposes the parameter space* $\mathbb{R}^8$ *into segments such that just one of them is biologically meaningful, for which the corresponding canonical Gröbner bases has at most a unique solution* (y, z) *with* $0 < z < y < 1$.

Proof. The Gröbner cover algorithm produces the segment $\mathscr{C}_{21} = \mathscr{V}(s_3 - s_5) \setminus \{[\mathscr{V}(2s_4^2 + 4s_4s_5 + 2s_4s_6 + 2s_4s_8 - s_4 + 4s_5s_8 + 2s_6s_8 - s_6) \cap \mathscr{V}(s_3 - s_5)] \cup [\mathscr{V}(s_4 - 2s_5 + s_6) \cap \mathscr{V}(s_3 - s_5)] \cup [\mathscr{V}(s_4 + 2s_5 + s_6) \cap \mathscr{V}(s_3 - s_5)]\}$, whose canonical Gröbner basis has the following generators:

1. $g_{21} = Ay + Bz$,

2. $g_{22} = -BCz^2 - D^2$,

where $A = 2s_4^2 + 4s_4s_5 + 2s_4s_6 + 2s_4s_8 - s_4 + 4s_5s_8 + 2s_6s_8 - s_6$, $B = -s_4 + 2s_5 - s_6$, $C = 2s_5 + s_4 + s_6$ and $D = 2s_4^2 + 4s_4s_5 + 2s_4s_6 + 2s_4s_8 + 4s_5s_8 + 2s_6s_8 - s_4 - s_6$.

As in section 2.1, $y = \exp(-4q2)$ and $z = \exp(-4q3)$ are the pathset variables, where q_2, q_3 are the q-parameters for the MC Jukes-Cantor rooted tripod tree of figure 1.

From the generator g_{22} it is clear that condition $0 < z^2$ occurs if $B < 0$. To guarantee $z^2 < 1$ it has to be $-BC - D^2 > 0$. Similarly, $A > 0$ and $B < 0$ imply $0 < y$. Condition $BD^2 + A^2C > 0$ implies $y < 1$.

Finally, condition $0 < z < y < 1$ is important to guarantee $q_1 = q_2 < q_3$. This last condition holds when $-B - A > 0$. ■

5 Conclusions

Theorems 3 and 4 show that, when the fitting procedure is applied sucessfully to the observed sequences of DNA characters from 3 species to tripod trees as those in figures 2 and 1, respectively, the solutions are unique. More over, these solutions are obtained for non extreme values of $s_1, s_2, \ldots s_8$ coming from the tripod's quartet extension. This way we deal successfully with the main goal of this article, but partially, because we made use of a different data technique as the one in [2], and we tracked other parameters.

Acknowledgements

We thank Antonio Montes (from Universidad Politécnica de Cataluña) and Mariemi Alonso (from Universidad Complutense de Madrid) for their advice in solving parametric systems of equations. We also thank Marta Casanellas (from Universidad Politécnica de Cataluña) for her advice in checking coherence and relevance of covered topics in the manuscript.

Financial support

This work was partially supported by the Spanish's "Ministerio de Ciencia e Innovación", ref. PID2020-114750GB-C32. It was also partially supported by the researching group "Entrepreneurs" UABJO-CA-46, where I am a collaborator.

References

[1] R. B. Bapat. *Linear algebra and linear models*. Third. Universitext. Springer, London; Hindustan Book Agency, New Delhi, 2012, viii+167. DOI: 10.1007/978-1-4471-2739-0.

[2] B. Chor, M. Hendy, and S. Snir. *Maximum likelihood Jukes-Cantor triplets: analytic solutions*. Molecular biology and evolution **23**(2006), no. 3, 626–632. DOI: 10.1093/molbev/msj069.

[3] D. A. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms*. Fourth. Undergraduate Texts in Mathematics. An introduction to computational algebraic geometry and commutative algebra. Springer, Cham, 2015, xvi+646. DOI: 10.1007/978-3-319-16721-3.

[4] S. N. Evans. *Fourier analysis and phylogenetic trees*. Modern signal processing. Vol. 46. Math. Sci. Res. Inst. Publ. Cambridge Univ. Press, Cambridge, 2004, 117–136. eprint: `http://library.msri.org/books/Book46/index.html`.

[5] G.-M. Greuel and G. Pfister. *A Singular introduction to commutative algebra*. Springer-Verlag, Berlin, 2002, xviii+588. DOI: `10.1007/978-3-662-04963-1`.

[6] M. Hendy. *The Relationship Between Simple Evolutionary Tree Models and Observable Sequence Data*. Systematic Biology **38**(Dec. 1989), no. 4, 310–321. DOI: `10.2307/2992397`.

[7] M. Hendy and S. Snir. *Hadamard conjugation for the Kimura 3st model: combinatorial proof using path sets*. IEEE/ACM Transactions on Computational Biology and Bioinformatics **5**(2008), no. 3, 461–471. DOI: `10.1109/TCBB.2007.70227`.

[8] M. Steel, M. Hendy, L. Székely, and P. Erdös. *Spectral analysis and a closest tree method for genetic sequences*. Applied mathematics letters **5**(1992), no. 6, 63–67. DOI: `10.1016/0893-9659(92)90016-3`.

[9] M. Steel, M. Hendy, and D. Penny. *Reconstructing phylogenies from nucleotide pattern probabilities: a survey and some new results*. Discrete Applied Mathematics **88**(1998), no. 1-3, 367–396. DOI: `10.1016/S0166-218X(98)00080-8`.