Tequio *4*(11), 2021: 41-52
ISSN: 2594-0546

Modelo de estimación de pesos de árbol filogenético para un cuartet, aplicando conjugación de Hadamard

Estimation model of phylogenetic tree weights on a quartet through Hadamard conjugation

Ernesto Álvarez-González[1]

Fecha de recepción: 30 de octubre de 2020
Fecha de aceptación: 23 de diciembre de 2020

Resumen - El tema de este artículo es la parte de la filogenética algebraica que propone una metodología para inferir los valores esperados de tres tipos de sustituciones de nucleótidos sobre las ramas de un árbol filogenético que explica las relaciones ancestrales de un conjunto de linajes asociado. Se ilustra una herramienta conocida como conjugación de Hadamard, que por relacionar tanto la distribución de probabilidad de los diferentes patrones de sustitución sobre las hojas del árbol filogenético, como el conjunto completo de valores esperados de sustituciones sobre sus ramas, promete ser un recurso de reconstrucción filogenética. Con base en esta técnica se construye una función de verosimilitud para un cuartet asociado a un alineamiento de cuatro secuencias de nucleótidos.

Palabras clave: Reconstrucción filogenética, conjugación de Hadamard, estimación de máxima verosimilitud.

Abstract - The subject on this paper is that part of algebraic phylogenetics which proposes a method to infere the expected values of three kinds of nucleotide substitutions on the branches of a phylogenetic tree that explains the ancestral relations among an associated set of lineages. A tool known as Hadamard conjugation is presented, which --because it connects both the probability distribution of the different substitution patterns on the leaves of the phylogenetic tree, and the complete set of expected values of substitutions on its branches-- may indeed be a resource to phylogenetic reconstruction. Based on this, a likelihood function is built for a quartet associated with an alignment of four nucleotide sequences.

Keywords: Phylogenetic reconstruction, Hadamard Conjugation, Maximum Likelihood Estimation.

1. Introducción

La reconstrucción filogenética es un área de investigación actual que ofrece explicar todas las relaciones ancestrales de un conjunto de especies (Cifuentes, 2015, p. 1). Esta reconstrucción demanda el conocimiento de un árbol filogenético como modelo de especiación, así como de un modelo de evolución molecular (Hendy &

[1] Estudiante de Doctorado, Facultad de Matemáticas, Universidad Complutense de Madrid, España. Profesor de Matemáticas en la Escuela de Ciencias de la Universidad Autónoma "Benito Juárez" de Oaxaca, México. ORCID: 0000-0001-5795-8752

Charleston, 1993, p. 232). En el caso de modelos binarios de evolución de caracteres, donde sólo hay dos estados observables, existen metodologías que permiten determinar en la última etapa el mejor árbol filogenético que explica la información existente (Hendy, 1989, pp. 317-318). Dichas metodologías son viables, ya que hay modelos de evolución para estos caracteres que proporcionan relaciones invertibles entre los valores esperados de cambios de estado sobre los diferentes caminos del árbol filogenético y la distribución de probabilidad de ambos caracteres sobre sus hojas (Hendy, 1989, p. 315). En el caso de modelos de evolución molecular, como el de Kimura 3-Parámetros, la conjugación de Hadamard proporciona una relación entre los valores esperados de tres tipos de sustitución molecular (transiciones, transversiones tipo I y transversiones tipo II) sobre los diferentes caminos del árbol filogenético y la distribución de probabilidad de los patrones de sustitución observables en sus hojas (Hendy & Snir, 2005, p. 15). Dicha relación no es invertible, por lo que la metodología de reconstrucción filogenética, en lugar de terminar con un árbol filogenético que explique mejor los datos observados, fija a uno de éstos desde el principio y toma como parámetros suyos dichas ternas de valores esperados (Chor, Hendy & Snir, 2006, p. 628). En consecuencia, el objetivo es determinar los valores óptimos para estas ternas que maximizan la probabilidad de que el árbol filogenético propuesto al inicio explique mejor los datos observados. La construcción de una función de verosimilitud es una opción adecuada para lograr dicha maximización (Chor, Kethan & Snir, 2003, p.78). En el contexto de la reconstrucción filogenética estas funciones son polinomios, cuyos valores extremos demandan metodologías de programación no lineal y de teorías algebraicas de resolución de sistemas de ecuaciones (Casanellas & Fernández-Sánchez, 2010, p. 1023).

El objetivo de este artículo es dar a conocer la herramienta de conjugación de Hadamard y ejemplificar su aplicación en el contexto de la reconstrucción filognética para el caso del cuartet de la Figura 2 como modelo de especiación del alineamiento de la Tabla 1, enfatizando cómo proponer la función de verosimilitud L(T) que maximice la probabilidad de que dicho cuartet describa mejor los datos observados en la Tabla 1. Esta metodología que concluye con la construcción de la función de verosimilitud L(T) es la que los autores del presente manuscrito definen como "Modelo de estimación de pesos de árbol filogenético".

Chor, Hendy & Snir (2006) propusieron este Modelo de estimación de pesos de árbol filogenético por primera vez dentro del contexto de un peine con tres hojas, bajo la restricción del modelo de evolución molecular tipo Jukes-Cantor, sujeto a la condición de reloj molecular, abriendo la posibilidad de aplicarlo al caso que se plantea en el presente manuscrito.

Las técnicas de optimización no lineal y de resolución de los sistemas de ecuaciones polinomiales que surgen para maximizar la función de verosimilitud, resultado de la aplicación del modelo, no se contemplan en este documento.

1.1 Definiciones

Definición **1.1.1**. Un alineamiento de m linajes es una identificación de nucleótidos entre las secuencias asociadas a éstos. El alineamiento puede representarse mediante una tabla, donde cada renglón está relacionado con una especie distinta y donde los nucleótidos pertenecientes a la misma columna provienen del mismo nucleótido del ancestro que comparten. Estas columnas pueden enumerarse en sitio 1, sitio 2, etcétera. Un alineamiento puede originar carácteres vacíos, llamados gaps, lo que sugiere que los linajes que las contienen pudieron haber perdido nucleótidos durante el proceso evolutivo (o bien los linajes que no los contienen pudieron haber ganado nucleótidos durante el proceso evolutivo). Cada columna de nucleótidos se conoce como un patrón de carácteres.

A lo largo de un proceso de evolución de linajes es posible que los nucleótidos muten aleatoriamente (Casanellas, 2018, p. 242). Kimura (1981) propone un modelo de mutación que establece tres diferentes tipos de sustitución (o mutación), junto con probabilidades fijas para éstas.

Definición **1.1.2**. En congruencia con el modelo de Kimura tres parámetros, se consideran tres tipos de sustitución junto con sus tasas infinitesimales de cambio: transiciones ($A \overset{\alpha}{\leftrightarrow} G$, $T \overset{\alpha}{\leftrightarrow} C$), transversiones tipo I ($A \overset{\beta}{\leftrightarrow} T$, $G \overset{\beta}{\leftrightarrow} C$) y transversiones tipo II ($A \overset{\gamma}{\leftrightarrow} C$, $T \overset{\gamma}{\leftrightarrow} G$). Para simplificar la notación, más adelante se identifican las transiciones con el entero 1; las transversiones tipo I, con el entero 2, y las transversiones tipo II, con el entero 3. El entero 0 identifica "no sustitución". Las tasas infinitesimales de cambio satisfacen la relación $\alpha + \beta + \gamma \leq 1$ para admitir la posibilidad de no cambio.

Definición **1.1.3**. Supongamos que hay un alineamiento de tamaño m. Al fijar su i-ésimo linaje, $1 \leq i \leq m$, sobre cada sitio se pueden considerar las sustituciones con respecto a dicho lugar. Esto origina un patrón de sustitución.

Ejemplo **1.1.1**. La siguiente tabla es un alineamiento de cuatro linajes con 16 sitios (sin gaps).

Tabla 1: Alineamiento de 4 linajes con 16 sitios, sin gaps.

site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\sigma_1 =$	C	C	A	T	C	A	A	A	C	G	T	G	T	G	A	C
$\sigma_2 =$	A	C	A	G	C	A	A	T	G	T	T	A	T	C	T	C
$\sigma_3 =$	C	C	A	T	T	G	A	A	G	A	T	G	C	G	T	T
$\sigma_4 =$	A	C	A	G	T	A	G	T	G	T	T	A	C	C	A	G

Con respecto al linaje 2 de la tabla anterior, la siguiente tabla recopila 16 patrones de sustitución:

Tabla 2: Patrones de Substitución del alineamiento de la tabla 1 con relación al linaje 2.

site	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
$\sigma_2 \to \sigma_1$	3	0	0	3	0	0	0	$T \to A = 2$	2	3	0	1	0	2	2	$C \to C = 0$
$\sigma_2 \to \sigma_3$	3	0	0	3	1	1	0	$T \to A = 2$	0	2	0	1	1	2	0	$C \to T = 1$
$\sigma_2 \to \sigma_4$	0	0	0	0	1	0	1	$T \to T = 0$	0	0	0	0	1	0	2	$C \to G = 2$

Michael & Sagi (2005) usan la siguiente notación para representar patrones de sustitución:

Elijamos un linaje de referencia, por ejemplo $i \in [n] = \{1,2,3,\cdots n\}$. Sean $A, B \subset [n] \setminus \{i\} = [n]_i$. La pareja ordenada (A, B) es el patrón de sustitución que cumple lo siguiente:

- $A \setminus B$: conjunto de linajes que se obtienen mediante una transición a partir del linaje de referencia.
- $B \setminus A$: conjunto de linajes que se obtienen mediante una transversión tipo I a partir del linaje de referencia.
- $A \cap B$: conjunto de linajes que se obtienen mediante una transversión tipo II a partir del linaje de referencia.

- $[n]\backslash(A \cup B)$: conjunto de linajes que comparten el mismo carácter que el linaje de referencia.

Ejemplo **1.1.2**. De la Tabla 2 del ejemplo 1.1.2, se fijó el linaje 2. Con respecto a éste, los patrones de sustitución en los sitios 8 y 16 tienen la siguiente representación alternativa, respectivamente:

Sitio 8 Los caracteres en los linajes 1 y 3 se obtienen del carácter en el linaje 2 a partir de una transversión tipo I: $(\emptyset, \{1,3\})$;

Sitio 16 El carácter en el linaje 3 se obtiene del carácter en el linaje 2 mediante una transición; el carácter en el linaje 4 se obtiene del carácter del linaje 2 a través de una transversión tipo I: $(\{3\}, \{4\})$.

Esta última notación para los patrones de sustitución servirán en las siguientes secciones para identificar tanto los renglones como las columnas de las matrices espectrales asociadas al teorema 3.3.1.

2. Distribución de probabilidad sobre un árbol filogenético

Casanellas (2018) explica que la reconstrucción de un árbol filogenético que da lugar a las especies actuales recurre a la modelación de su evolución con procesos de Markov de sustitución de nucleótidos. De acuerdo con el modelo de Kimura 3-Parámetros, hay tres sustituciones de nucleótidos, lo que implica la existencia de una matriz de transición 4×4 que identifica las probabilidades de las diferentes sustituciones que se pueden observar, dependiendo de los nucleótidos presentes en los vértices asociados. Ya sea que se parta desde una raíz del árbol filogenético o desde algún otro vértice suyo, se puede construir una distribución de probabilidad para los caracteres que se observan en sus hojas, tomando en cuenta a las matrices de transición.

Aclaremos con un ejemplo muy sencillo cómo se puede construir una distribución de probabilidad sobre el árbol filogenético de la Figura 1.

Ejemplo **2.0.1**. Consideremos la siguiente distribución de caracteres sobre los vértices del árbol filogenético de la Figura 2: $\chi_o(1) = \chi_e(5) = A$, $\chi_o(2) = T$, $\chi_o(3) = \chi_o(4) = C$ y $\chi_e(6) = G$. La letra griega χ se reserva en este ejemplo para denotar una función que va del conjunto de los vértices del cuartet de la Figura 1 al conjunto de nucleótidos. El número entre paréntesis hace referencia al vértice. El subíndice identifica si el vértice es "observado" o "Escondido" (en el contexto de la teoría filogenética, sólo las hojas son observadas). En resumen: los vértices 1 y 5 muestran adenina, el vértice 2 muestra timina, los vertices 3 y 4 muestran citocina y el vértice 6 muestra guanina.

Proponemos las siguientes matrices de transición tipo Kimura 3-Parámetros (observe su simetría) sobre las ramas del árbol filogenético de la Figura 2:

$$M_1 = M_2 = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} 0.7 & 0.15 & 0.05 & 0.1 \\ 0.15 & 0.7 & 0.1 & 0.05 \\ 0.05 & 0.1 & 0.7 & 0.15 \\ 0.1 & 0.05 & 0.15 & 0.7 \end{pmatrix} \end{array}$$

$$M_{124} = M_4 = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} 0.6 & 0.2 & 0.05 & 0.15 \\ 0.2 & 0.6 & 0.15 & 0.05 \\ 0.05 & 0.15 & 0.6 & 0.2 \\ 0.15 & 0.05 & 0.2 & 0.6 \end{pmatrix} \end{array}$$

$$M_{12} = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} 0.75 & 0.15 & 0.025 & 0.075 \\ 0.15 & 0.75 & 0.075 & 0.025 \\ 0.025 & 0.075 & 0.75 & 0.15 \\ 0.075 & 0.025 & 0.15 & 0.75 \end{pmatrix} \end{array}$$

Proponemos la distribución de probabilidades uniforme sobre la raíz del árbol: $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

La probabilidad de observar dicha distribución de nucleótidos sobre los vértices del cuartet es:

$$p_{(ATCC|AG)} = \frac{1}{4} M_1(A|A) M_2(T|A) M_{12}(G|A) M_{124}(C|G) M_4(C|G) =$$
$$\frac{1}{4}(0.7)(0.1)(0.15)(0.15)(0.15) = 5.90625 \times 10^{-25}$$

Observe que todas las probabilidades involucradas en el cálculo anterior son condicionadas.

2.1 Función de verosimilitud

El problema de filogenética que se formula en el presente documento es el siguiente: se propone un árbol filogenético para un conjunto finito de linajes, cada uno identificado por su secuencia de nucleótidos. Se propone un modelo de evolución molecular especificado por las matrices de transición sobre sus ramas. Las entradas de estas matrices son parámetros del modelo. Se propone una distribución de probabilidad uniforme sobre su raíz (si es que tiene) o bien sobre el vértice que se determine, a partir del cual se consideran caminos convergentes en las hojas. La distribución de probabilidad de los patrones de carácter para el árbol se plantea como en el ejemplo 2.0.1. Si dicho árbol tiene *n* hojas, los patrones de carácter (que en esta sección se identifican por secuencias de nucleótidos $\gamma_1, \gamma_2, \gamma_3, \cdots \gamma_n$) también son *n*-dimensionales. Se define una función de verosimilitud $L(T)$ como sigue:

$$L = \prod_{\Gamma^m} p_{\gamma_1,\gamma_2,\gamma_3,\cdots\gamma_n}{}^{f_{\gamma_1,\gamma_2,\gamma_3,\cdots\gamma_n}}, \qquad (1)$$

donde n indica el número de linajes, $\Gamma = \{A, G, T, C\}$, $p_{\gamma_1,\gamma_2,\gamma_3,\cdots\gamma_n}$ se calcula como en el ejemplo 2.0.1 y $f_{\gamma_1,\gamma_2,\gamma_3,\cdots\gamma_n}$ es la frecuencia relativa observada del mismo patrón de carácter.

Más adelante, en la sección 3.4, se formulará una función de verosimilitud cuyos parámetros sean los valores esperados de sustitución (también llamados pesos) de cada rama del árbol filogenético, en lugar de las entradas

de las matrices de transición. De hecho esta segunda versión de la función de verosimilitud es la que se usará sobre el cuartet de la Figura 1.

3. Conjugación de Hadamard

Conjugación de Hadamard es una relación que involucra los pesos de un árbol filogenético con la distribución de probabilidad de los patrones de sustitución asociados a un alineamiento de un conjunto finito de linajes.

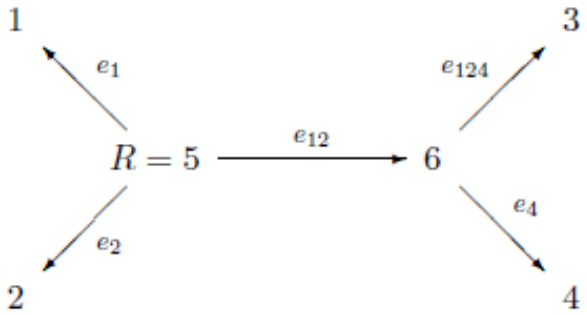
Para establecer dicha relación es necesario construir dos matrices espectrales: la matriz "Espectro de longitud de borde" y la matriz "Espectro de secuencia espectral". La primera incluye los pesos del árbol filogenético; la segunda contiene las probabilidades de todos los patrones de sustitución asociados al alineamiento.

3.1 Espectro de longitud de borde

Antes de dar una definición de esta matriz, es importante aclarar la notación. Supongamos que tenemos *n* linajes. Fijemos el i-ésimo (puede ser cualquier otro). Toda pareja ordenada de subconjuntos (ajenos entre sí) A, B de $[n] = \{1,2,\cdots n\}$ es una bipartición del conjunto de linajes, si $A \cup B = [n]$. Se acostumbra a identificar a dichas parejas con el subconjunto que no incluye al i-ésimo linaje (linaje de referencia).

Sobre la base de un árbol filogenético, hacer un corte sobre cualquiera de sus ramas también produce una bipartición: dicho corte descompone el árbol en dos subárboles complementarios (cada subárbol tiene asociado un conjunto de hojas). Cada rama se identifica con la bipartición asociada al corte de ésta (o mejor dicho, con el subconjunto componente de la bipartición que excluye a la hoja de referencia). En el siguiente ejemplo, los lados se denotan con la letra e más un subíndice asociado a la bipartición del corte.

Ejemplo **3.1.1**. Fijemos la tercera hoja del árbol filogenético de la Figura 1. Sus biparticiones se identifican por el subconjunto que excluye la tercera hoja:



Nota: Reservamos el símbolo $e(T)$ para denotar el conjunto de lados (o ramas) del árbol filogenético T.

Para un árbol filogenético T que explique las relaciones ancestrales de n linajes, se define su matriz Espectro de longitud de borde, Q, de la siguiente manera:

$$q_{A,B} = \begin{cases} q_{e_A}(1) & si \quad e_A \in e(T) \quad y \quad si \quad B = \emptyset \\ q_{e_B}(2) & si \quad e_B \in e(T) \quad y \quad si \quad A = \emptyset \\ q_{e_A}(3) & si \quad e_A \in e(T) \quad y \quad si \quad A = B \\ -K_T & si \quad\quad\quad\quad\quad\quad A = B = \emptyset \\ 0 & otro \quad\quad\quad\quad\quad caso. \end{cases}$$

Reservamos el símbolo $q_{e_\Delta}(j)$, $\Delta \in 2^{[n]_i}$ y $j \in [3]$ para denotar el valor esperado de sustituciones tipo j sobre la rama e_Δ. Las columnas y filas de la matriz Q están ordenadas lexicográficamente, de acuerdo con los subconjuntos de $[n]_i$, siendo i la hoja de referencia. Observe que el tamaño de la matriz Q es $2^{n-1} \times 2^{n-1}$. $K_T = \sum_{\Delta \in 2^{[n]_i}} (q_{e_\Delta}(1) + q_{e_\Delta}(2) + q_{e_\Delta}(3))$.

Ejemplo **3.1.2**. Con respecto al tercer linaje del árbol filogenético de la Figura 1, la sucesión de subconjuntos de $\{1,2,4\}$, cuyos elementos están ordenados lexicográficamente, es la siguiente:

$$\begin{aligned} 0 &\rightarrow \emptyset \\ 1 &\rightarrow \{1\} \\ 10 &\rightarrow \{2\} \\ 11 &\rightarrow \{1,2\} \\ 100 &\rightarrow \{4\} \\ 101 &\rightarrow \{1,4\} \\ 110 &\rightarrow \{2,4\} \\ 111 &\rightarrow \{1,2,4\} \end{aligned}$$

En este último caso, observe que, como el tercer linaje es el de referencia, su lugar lo ocupa el cuarto linaje: 1, 2, 4. El tercer linaje se omite sólo para fines de representación de las biparticiones asociadas.

En congruencia con este ordenamiento para los conjuntos A y B, la matriz Espectro de longitud de borde para el árbol filogenético de la figura 2 es:

$$Q = \begin{array}{c} \\ \emptyset \\ \{1\} \\ \{2\} \\ \{1,2\} \\ \{4\} \\ \{1,4\} \\ \{2,4\} \\ \{1,2,4\} \end{array} \begin{array}{c} \begin{array}{cccccccc} \emptyset & \{1\} & \{2\} & \{1,2\} & \{4\} & \{1,4\} & \{2,4\} & \{1,2,4\} \end{array} \\ \begin{pmatrix} -K & q_1(2) & q_2(2) & q_{12}(2) & q_4(2) & 0 & 0 & q_{124}(2) \\ q_1(1) & q_1(3) & 0 & 0 & 0 & 0 & 0 & 0 \\ q_2(1) & 0 & q_2(3) & 0 & 0 & 0 & 0 & 0 \\ q_{12}(1) & 0 & 0 & q_{12}(3) & 0 & 0 & 0 & 0 \\ q_4(1) & 0 & 0 & 0 & q_4(3) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ q_{124}(1) & 0 & 0 & 0 & 0 & 0 & 0 & q_{124}(3) \end{pmatrix} \end{array}$$

Note que como el árbol filogenético de la Figura 1 no tiene ramas asociadas a las biparticiones $\{1,4\}$ y $\{2,4\}$, los renglones y columnas correspondientes a éstas en la matriz Q tienen sólo ceros.

3.2 Espectro de secuencia espectral

Para un conjunto de n linajes, esta matriz es de tamaño $2^{n-1} \times 2^{n-1}$, pues sus renglones y columnas están ordenadas del mismo modo como lo están los renglones y columnas de la matriz Espectro de longitud de borde. A la matriz Espectro de secuencia espectral se le denota por la letra P.

Observe que, para n linajes distintos, después de fijar a uno de éstos, hay a lo más $4^{n-1} = 2^{n-1} \times 2^{n-1}$ diferentes patrones de sustitución.

El siguiente ejemplo ilustra cómo se pueden aproximar las entradas de la matriz P, partiendo del alineamiento de los cuatro linajes de la Tabla 2 del ejemplo 1.1.2:

Ejemplo **3.2.1**.La Tabla 2 de la sección 1.1 resume los patrones de sustitución de un alineamiento de cuatro linajes. Éstos pertenecen a una matriz Espectro de secuencia espectral de tamaño $2^{4-1} \times 2^{4-1} = 2^3 \times 2^3 = 8 \times 8$:

$$P = \begin{array}{c} \\ \emptyset \\ \{1\} \\ \{3\} \\ \{1,3\} \\ \{4\} \\ \{1,4\} \\ \{3,4\} \\ \{1,3,4\} \end{array} \begin{array}{c} \begin{array}{cccccccc} \emptyset & \{1\} & \{3\} & \{1,3\} & \{4\} & \{1,4\} & \{3,4\} & \{1,3,4\} \end{array} \\ \begin{pmatrix} \frac{3}{16} & \frac{1}{16} & 0 & \frac{2}{16} & 0 & \frac{1}{16} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{16} & 0 & 0 & 0 & 0 \\ \frac{1}{16} & 0 & 0 & 0 & \frac{1}{16} & 0 & 0 & 0 \\ \frac{1}{16} & 0 & 0 & \frac{2}{16} & 0 & 0 & 0 & 0 \\ \frac{1}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{2}{16} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

3.3 Conjugación de Hadamard

Teorema **3.3.1** Sea Q la matriz Espectro de longitud de borde de un árbol filogenético con n hojas. Sea P su matriz Espectro de secuencia espectral.

Se cumple lo siguiente:

$$H_n P H_n = \exp(H_n Q H_n), \qquad (2)$$

donde H_n es la matriz de Hadamard de tamaño $2^{n-1} \times 2^{n-1}$, obtenida inductivamente de esta manera:

1. $H_1 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$

2. $H_n = H_1 \otimes H_{n-1}$.

En este caso, el término exp de la ecuación 2 se refiere a la función exponencial $\exp: \mathbb{R} \to \mathbb{R}$, que se evalúa independientemente en cada entrada de la matriz producto $H_n Q H_n$.

El símbolo $\otimes$ del teorema 3.3.1 significa el producto de Kronecker. El primer factor en un tal producto es la matriz indicadora, como lo ilustra el siguiente ejemplo:

Ejemplo **3.3.1** (Producto de Kronecker). Sean A y B las siguientes matrices:

$$A = \begin{pmatrix} \frac{13}{5} & -7 \\ 0 & \pi \end{pmatrix}, B = \begin{pmatrix} 5 & 21 \\ -6 & 2 \end{pmatrix}.$$

$$A \otimes B = \begin{pmatrix} (\frac{13}{5})(5) & (\frac{13}{5})(21) & (-7)(5) & (-7)(21) \\ (\frac{13}{5})(-6) & (\frac{13}{5})(2) & (-7)(-6) & (-7)(2) \\ (0)(5) & (0)(21) & (\pi)(5) & (\pi)(21) \\ (0)(-6) & (0)(2) & (\pi)(-6) & (\pi)(2) \end{pmatrix} = \begin{pmatrix} 13 & \frac{273}{5} & -35 & -147 \\ -\frac{78}{5} & \frac{26}{5} & 42 & -14 \\ 0 & 0 & 5\pi & 21\pi \\ 0 & 0 & -6\pi & 2\pi \end{pmatrix}$$

De las dos matrices espectrales P y Q del teorema 3.3.1, la matriz P se puede aproximar a través de un alineamiento de secuencias de nucleótidos. Pasa lo contrario con la matriz Q, porque está asociada a un proceso evolutivo desconocido. Tampoco cobra mucho sentido despejar esta última del teorema 3.3.1, pues no se puede garantizar que todas las entradas de la matriz $H_n P H_n$ sean positivas y por lo tanto no puede aplicarse sobre sus términos la función logaritmo natural.

3.4 Conjugación de Hadamard y su papel en la reconstrucción filogenética

Al final de la sección 3.3 se aclaró la dificultad de despejar la matriz Q de la ecuación 2. Benny *et al.* (2006) sugieren una metodología para proponer otra función de verosimilitud alterna a la que se plantea en la sección 2.1, tomando como parámetros del modelo de evolución los pesos de las ramas asociadas al árbol filogenético propuesto. En resumen, dicha metodología es la siguiente:

- Calcular la matriz P de la ecuación 2;
- construir una función de verosimilitud, haciendo los siguientes cambios sobre los elementos de la función 1:

    1. $p_{\gamma_1,\gamma_2,\gamma_3} \to P(X,Y)$ y
    2. $f_{\gamma_1,\gamma_2,\gamma_3} \to f(X,Y)$,

    donde $P(X,Y)$ representa la probabilidad del patrón de sustitución (X,Y). Asimismo, $f(X,Y)$ es la frecuencia relativa (observada) del patrón de substitución (X,Y) con relación a un alineamiento disponible.

    La función de verosimilitud resultante es:

$$L(T) = \prod_{X,Y \subseteq \{1,2\}} P(X,Y)^{f(X,Y)}. \qquad (3)$$

Para ilustrar esta metodología sobre el cuartet de la Figura 1 como modelo de evolución para los linajes presentes en la Tabla 1, se suponen las siguientes restricciones:

- Hay una única tasa de sustitución fija sobre el árbol filogenético ($\alpha = \beta = \gamma$);

- imponemos la condición de reloj molecular sobre el árbol filogenético: es la misma distancia evolutiva desde la raíz del cuartet de la Figura 1 hacia cualquier hoja suya.

Estas dos restricciones implican lo siguiente:

- $M_1 = M_2$, $M_{124} = M_4$ y $M_1 = M_{12} \times M_4$;
- $q_1 = q_2$, $q_{12} = q_4$ y $q_1 = q_{12} + q_4$.

Bajo estas consideraciones, la matriz Espectro de longitud de borde, Q, se reduce como sigue:

$$Q = \begin{array}{c} \\ \emptyset \\ \{1\} \\ \{2\} \\ \{1,2\} \\ \{4\} \\ \{1,4\} \\ \{2,4\} \\ \{1,2,4\} \end{array} \begin{array}{c} \begin{array}{cccccccc} \emptyset & \{1\} & \{2\} & \{1,2\} & \{4\} & \{1,4\} & \{2,4\} & \{1,2,4\} \end{array} \\ \begin{pmatrix} -K & q_1 & q_1 & q_1 - q_4 & q_4 & 0 & 0 & q_4 \\ q_1 & q_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ q_1 & 0 & q_1 & 0 & 0 & 0 & 0 & 0 \\ q_1 - q_4 & 0 & 0 & q_1 - q_4 & 0 & 0 & 0 & 0 \\ q_4 & 0 & 0 & 0 & q_4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ q_4 & 0 & 0 & 0 & 0 & 0 & 0 & q_4 \end{pmatrix} \end{array}$$

donde $K = 9q_1 + 3q_4$.

La matriz Espectro de secuencia espectral, P, se puede obtener del teorema 3.3.1, ejecutando el siguiente código escrito sobre el sistema de cómputo Maple:

```
> with(LinearAlgebra):
Construcción de la matriz Espectro de Longitud de
Borde Q asociada al Árbol filogenético de la figura 2:
> K := 9*q[1]+3*q[4] :
> Q := Matrix( [ [-K,q[1],q[1],q[1]-q[4],q[4],0,0,q[4]],
[q[1],q[1],0,0,0,0,0,0],[q[1],0,q[1],0,0,0,0,0],
[q[1]-q[4],0,0,q[1]-q[4],0,0,0,0],[q[4],0,0,0,q[4],0,0,0],
[0,0,0,0],[0,0,0,0],[q[4],0,0,0,0,0,0,q[4] ] ] ):
Construcción inductiva de la matriz de Hadamard, H_3 :
> H_1 := Matrix( [ [1,1],[1,-1] ] ) :
> H_2 := KroneckerProduct(H_1,H_1) :
> H_3 := KroneckerProduct(H_1,H_2) :
```

Los siguientes pasos se requieren para despejar la matriz P del teorema 3.3.1, correspondiente a la matriz Espectral de Longitud de Borde Q previamente construida:

```
> HQH := H_3.Q.H_3 :
La función exponencial del teorema 3.3.1 se aplica término
a término sobre los elementos de la matriz HQH:
> for i from 1 to 8 do
```

```
for j from 1 to 8 do
E[i, j] := exp(-HQH[i, j] ):
end do:
end do:
```

Calculemos la inversa de la matriz H_3, que redenotaremos por KK para no confundirla con la entrada de la esquina superior izquierda de la matriz Q:

```
> KK := MatrixInverse(H_3):
```

La siguiente línea de código construye la matriz exp(HQH) del teorema 3.3.1:

```
> EHQH := Matrix( [
[E[1, 1], E[1, 2], E[1, 3], E[1, 4], E[1, 5], E[1, 6], E[1, 7], E[1, 8] ],
[E[2, 1], E[2, 2], E[2, 3], E[2, 4], E[2, 5], E[2, 6], E[2, 7], E[2, 8] ],
[E[3, 1], E[3, 2], E[3, 3], E[3, 4], E[3, 5], E[3, 6], E[3, 7], E[3, 8] ],
[E[4, 1], E[4, 2], E[4, 3], E[4, 4], E[4, 5], E[4, 6], E[4, 7], E[4, 8] ],
[E[5, 1], E[5, 2], E[5, 3], E[5, 4], E[5, 5], E[5, 6], E[5, 7], E[5, 8] ],
[E[6, 1], E[6, 2], E[6, 3], E[6, 4], E[6, 5], E[6, 6], E[6, 7],E [6, 8] ],
[E[7 ,1], E[7, 2], E[7, 3], E[7, 4], E[7, 5], E[7, 6], E[7, 7], E[7, 8] ],
[E[8, 1], E[8, 2], E[8, 3], E[8, 4], E[8, 5], E[8, 6], E[8, 7], E[8, 8] ]
] ) :
```

A esta última matriz la multiplicamos por KK por ambos lados. Esta es la matriz P del teorema 3.3.1:

```
> P := KK.EHQH.KK:
```

Antes de construir la función de verosimilitud en congruencia con la ecuación 3, se hace el siguiente cambio de variables: $x = \exp(q_1)$ y $y = \exp(q_4)$.

La función de verosimilitud para el cuartet de la Figura 1, en congruencia con el alineamiento de la Tabla 1, es la siguiente:

$$L(T) = \frac{2^{7/8}}{256}(12x^{12}y^4 + 9x^8y^8 + 12x^{12} + 12x^8y^4 + 15x^8 + 3y^8 + 1)^{3/8}(-4x^{12}y^4 - 3x^8y^8 - 4x^{12} + 4x^8y^4 + 3x^8 + 3y^8 + 1)^{1/8}(-4x^{12}y^4 + 9x^8y^8 - 4x^{12} - 4x^8y^4 - x^8 + 3y^8 + 1)^{5/16}(4x^{12}y^4 - 3x^8y^8 + 4x^{12} - 4x^8y^4 - 5x^8 + 3y^8 + 1)^{1/16}((y^4 - 1)(4x^{12} + x^8y^4 - 3x^8 - y^4 - 1))^{3/16}(-(y^4 - 1)(4x^{12} + 3x^8y^4 + 7x^8 + y^4 + 1))^{1/16}$$

4. Conclusiones

El modelo de Estimación de pesos de árbol filogenético que se ilustra en este artículo puede aplicarse a una diversidad de árboles filogenéticos, sujetos a modelos de evolución molecular, como Kimura 3-Parámetros, Kimura 2-Parámetros o Jukes-Cantor. El caso que se ilustra en el presente manuscrito supone válido el último de estos modelos. La razón de haber elegido Jukes-Cantor es que simplifica el modelo de estimación de pesos de árbol filogenético, pues las matrices de transición asociadas tienen más simetrías que aquellas en

correspondencia con los otros modelos de evolución molecular y porque las tres tasas infinitesimales de mutación se reducen a una sola. La condición de reloj molecular también ayuda mucho, ya que disminuye el número de parámetros y simplifica la matriz Espectral de longitud de borde (de hecho, permite establecer relaciones lineales entre diferentes pesos).

El siguiente problema que queda abierto con relación al ejemplo desarrollado en la sección 3.4, no tanto por su relevancia biológica (pues el alineamiento de las secuencias de nucleótidos es ficticio) sino para resaltar la naturaleza de las técnicas que se decidieran usar, es cómo maximizar la función de verosimilitud $L(T)$ de la misma sección.

Referencias

Carnevali, G., Cetzal-Ix, W., Balam R.N., Leopardi, C., & Romero-González, G. A. (2013). A combined evidence phylogenetic re-circumscription and a taxonomic revision of *Lophiarella* (*Orchidaceae: Oncidiinae*). *Systematic Botany, 38*(1), 46-63.

Casanellas, M. (2018). El modelo evolutivo de Kimura: un enlace entre el álgebra, la estadística y la biología. *La Gaceta de la RSME, 21*(2), 241-257. Recuperado de https://gaceta.rsme.es/abrir.php?id=1444

Casanellas, M., Sánchez, F. (2010). Reconstrucción Filogenética usando Geometría Algebraica. ARBOR *Ciencia, Pensamiento y Cultura*, 186 1023-1033, doi: 10.3989/arbor.2010.746n1251

Chor, B., Hendy, M. D., & Snir, S. (2006). Maximum Likelihood Jukes-Cantor Triplets: Analytic Solutions. *Molecular Biology and Evolution, 23*(3), 626-632.

Chor, B., Khetan, A., & Snir, S. (2003). Maximum Likelihood on Four TaxaPhylogenetic Trees: Analytic Solutions. *RECOMB 03: Proceedings of the Seventh Annual Conference on Research in Computational Molecular Biology,* 76-83. https://doi.org/10.1145/640075.640084

Cifuentes-Fontanals, L. (2015). *Application of algebraic techniques to phylogenetic reconstruction* (Bachelor's degree thesis). Depto. Matemática Aplicada I, Facultad de Matemáticas y Estadística, Universidad Politécnica de Cataluña.

Dariusz, L., Szlachetko, Mytnik-Ejsmont, J., & Romowicz, A., (2006). Genera et species Orchidalium. 14. Oncidieae. *Polish Botanical Journal,* 51, 53-55. Recuperado de http://maxbot.botany.pl/cgi-bin/pubs/data/article pdf?id=1732

Hendy, M. D. (1989). The relationship between simple evolutionary tree models and observable sequence data. *Sistematic Zoology, 38*(4), 310-321.

Hendy, M. D., & Charleston, M. A. (1993) Hadamard conjugation: a versatile tool for modelling nucleotide sequence evolution. *New Zealand Journal of Botany, 31*(3), 231-237.

Hendy, M. D., & Snir, S., (2005). Hadamard Conjugation for the Kimura 3st Model: Combinatorial Proof using Pathsets. *arXiv: q-bio/0505055v2 [q-bio.PE]*. Recuperado de https://arxiv.org/pdf/q-bio/0505055.pdf

Kimura, M. (1981). Estimation of evolutionary sequences between homologous nucleotide sequences. *Proceedings of the National Academy of Sciences of the Unites States of America, 78*(1), 454-458.

Simmons, M. P., & Ochoterena, H., (2000). Gaps as characters in sequence-based phylogenetic analyses. *Systematic Biology, 49*(2), 369-381. https://doi.org/10.1093/sysbio/49.2.369