El modelo evolutivo de Kimura: un enlace entre el álgebra, la estadística y la biología

por

Marta Casanellas*

Resumen. El álgebra y la estadística han estado presentes en biología evolutiva desde sus inicios. En este artículo mostramos una pequeña parte de la interrelación entre el álgebra, la estadística y la biología evolutiva centrándonos en el estudio de un modelo de mutación de nucleótidos propuesto por Kimura. Se puede explorar la estructura algebraica de este modelo mediante el uso de la transformada de Hadamard (o de Fourier sobre un grupo finito). Explicaremos cómo estas técnicas se pueden usar en métodos de selección de modelos evolutivos y de reconstrucción filogenética.

1. Introducción

Seguramente nadie se cuestiona que la estadística está detrás de los estudios en biología evolutiva, pero seguro que hay dudas sobre el papel del álgebra y la geometría en este campo. Sin embargo, el álgebra ha estado presente en biología evolutiva desde que las matemáticas se usan para modelizar la evolución de especies y de poblaciones y, en consecuencia, la biología, la estadística y el álgebra han interactuado desde hace mucho tiempo. Un ejemplo de esta interacción lo encontramos ya en el trabajo de K. Pearson [21] de 1894 y se recoge bien en la cita del biólogo J. M. Smith: «*If you can't stand algebra, keep out of evolutionary biology*» [25].

En este artículo nos proponemos mostrar el uso de técnicas algebraicas para entender el modelo evolutivo de Kimura 3-parámetros [19] (y en consecuencia también sus submodelos), y para ver cómo se pueden dar métodos de selección de modelos evolutivos y de reconstrucción filogenética basados en estas técnicas. La filogenética tiene como objetivo recomponer las relaciones ancestrales entre especies actuales a partir de sus genomas. Establecer la historia evolutiva de las especies biológicas es relevante para distintas áreas de la biología. Por ejemplo, es necesario para la detección de genes, para identificar la funcionalidad de éstos, o para el estudio de la estructura de proteínas.

Se suele modelizar la evolución de las especies en un *árbol filogenético* (véase la figura 1), donde las hojas representan las especies que viven actualmente en el planeta

Este artículo es una contribución de La Gaceta a la celebración del *Año de la Biología Matemática 2018*.

*Parcialmente financiada por la Generalitat de Catalunya, 2017 SGR-932, el proyecto MINECO/FEDER MTM2015-69135-P y la BGSMath MDM-2014-0445.

y los nodos interiores representan sus especies ancestrales. Uno de los objetivos primordiales de la filogenética es encontrar el árbol filogenético más plausible que relaciona un grupo de especies actuales dadas. Para conseguir este objetivo se usa el genoma de las especies actuales y se proponen modelos estadísticos de evolución de las secuencias de ADN.

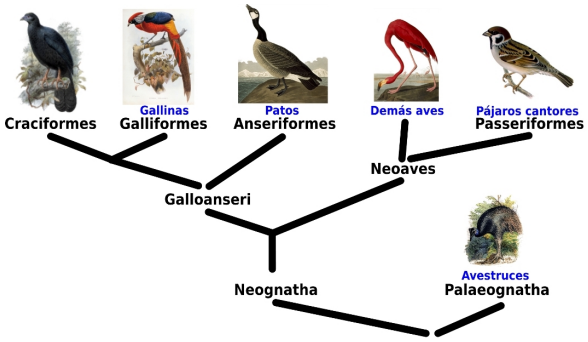

Figura 1: Un árbol filogenético con especies de aves. Autor: Ecelan, distribuida bajo la *GNU Free Documentation License*.

El modelo de Kimura 3-parámetros es uno de los modelos básicos de mutación de nucleótidos. Kimura planteó que, en ciertas partes del genoma, los nucleótidos mutan aleatoriamente; esta hipótesis es la base de la mayoría de métodos de reconstrucción filogenética que se usan actualmente. Propuso además un modelo de sustitución de nucleótidos basado en las distintas composiciones químicas de los cuatro nucleótidos. Desde el punto de vista matemático, el modelo de Kimura 3-parámetros es interesante gracias a las simetrías de sus matrices de transición. Como veremos, la transformada de Hadamard diagonaliza simultáneamente todas las matrices de transición de este modelo, independientemente de los parámetros. Este hecho permite estudiarlo a fondo y tiene consecuencias relevantes en el estudio filogenético: permite caracterizar exactamente las distribuciones de secuencias de nucleótidos que evolucionan en un árbol filogenético bajo este modelo, determinar la identificabilidad de los parámetros del modelo, y proponer métodos de reconstrucción filogenética y de selección de modelos. La transformada de Hadamard es una técnica que ha sido usada en múltiples aplicaciones, desde la computación cuántica hasta la cristalografía. En el contexto de filogenética que tratamos aquí, se corresponde con una transformada de Fourier sobre un grupo finito (véase la sección 3); sin embargo, en este artículo hemos optado por usar sólo técnicas de álgebra lineal y multilineal para presentar los resultados.

Un estudio profundo de la estructura de este modelo permite usar técnicas de álgebra lineal y multilineal para entender la geometría del espacio de distribuciones de secuencias de nucleótidos que evolucionan bajo el modelo de Kimura 3-parámetros. Este tipo de herramientas tienen cabida dentro de la geometría algebraica: el espacio de distribuciones del que hablamos se puede pensar como (un subconjunto de) una variedad algebraica y en este caso el estudio en profundidad del modelo

permite descubrir que hablamos de una variedad tórica (véase la sección 4). Ligada a la geometría algebraica encontramos el álgebra conmutativa y en este marco los resultados de la sección 4 nos permiten descubrir generadores del ideal de la variedad algebraica correspondiente. En filogenética, los polinomios del ideal de la variedad fueron llamados «invariantes filogenéticos» por los biólogos Cavender y Felsenstein en el año 1987 [7]: los definieron como polinomios que se anulan sobre cualquier distribución de secuencias de nucleótidos que evolucionan en un árbol filogenético dado (bajo un modelo evolutivo prefijado).

Todas estas técnicas se enmarcan dentro de una nueva disciplina llamada *estadística algebraica*, nombre acuñado por Pistone, Riccomagno y Wynn [22] (véase también el libro de referencia [20]). En sus aplicaciones a la filogenética, esta disciplina ha dado lugar a herramientas potentes como SVDquartets [8], Erik+2 [12] y Split Scores [1], que han sido implementadas en uno de los programas de reconstrucción filogenética más usados por los biólogos, PAUP* [29]. Aunque en este artículo nos restringimos a un modelo de evolución sencillo y consideraremos sólo árboles de cuatro especies, técnicas similares son las que han dado lugar a estos métodos más generalizados, que contemplan modelos y situaciones mucho más complejas.

2. El modelo de Kimura 3-parámetros

Las relaciones ancestrales entre un conjunto de especies se representan en un *árbol filogenético*. En él, los nodos hojas representan las especies actuales, la raíz (si existe) representa su ancestro común, los nodos interiores representan especies ancestrales y cada división en ramas representa un proceso de especiación (véase la figura 1). En términos matemáticos, un árbol filogenético es un árbol (es decir, un grafo conexo y sin ciclos) junto con una biyección entre las hojas y las especies actuales que representamos en el árbol. El árbol tiene *raíz* si hay un nodo interior distinguido desde el que se orientan las aristas de forma natural. Aunque la raíz tiene el papel destacado de representar el ancestro común a todas las especies del árbol, a menudo tratamos con árboles sin raíz puesto que la mayoría de los métodos de reconstrucción filogenética no pueden inferir la posición de la raíz. Cuando se habla de *topología* del árbol filogenético, se hace referencia a la clase de isomorfismo del grafo teniendo en cuenta que los isomorfismos tienen que preservar la biyección dada en las hojas. La topología especifica qué grupos de especies se crean en cada momento del proceso evolutivo (por ejemplo, los tres árboles filogenéticos sin raíz de la figura 2 tienen distinta topología). La longitud de una arista de un árbol filogenético normalmente representa la distancia evolutiva entre los dos nodos de la arista (en términos de cantidad de mutaciones), pero en este artículo no tendremos en cuenta esta magnitud.

El estudio evolutivo de las especies se lleva a cabo a partir del ADN de genes asociados a ellas. Debido a la simetría de la doble hélice del ADN, las moléculas de ADN las pensamos como una secuencia ordenada de caracteres en el alfabeto $\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}$ (letras que representan los nucleótidos: adenina $\mathtt{A}$, citosina $\mathtt{C}$, guanina $\mathtt{G}$, y timina $\mathtt{T}$).

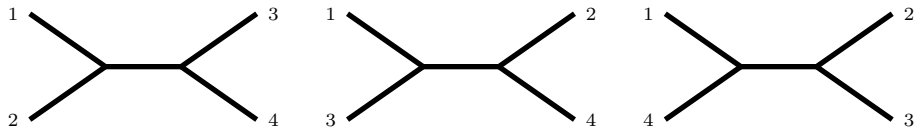

Figura 2: Las tres topologías posibles de árboles de 4 hojas sin raíz sobre el conjunto de especies $\{1, 2, 3, 4\}$ se denotan por $T_{12|34}$, $T_{13|24}$, y $T_{14|23}$ respectivamente.

Con la finalidad de reconstruir el árbol filogenético que ha dado lugar a las especies actuales, se recurre habitualmente a modelizar la evolución con procesos de Markov de sustitución de nucleótidos. El proceso de Markov en un árbol asume que los procesos evolutivos en distintas ramas sólo dependen del nodo que tengan en común. Supondremos además que las distintas posiciones de la cadena de ADN evolucionan aleatoriamente, de forma independiente y bajo las mismas probabilidades de mutación, por lo que es suficiente modelizar la mutación de un solo nucleótido como sigue.

Suponemos dado un árbol filogenético, por ejemplo el de la figura 3 donde el nodo r hace el papel de raíz (puesto que las aristas son dirigidas desde este nodo hacia las hojas). Seguidamente a cada nodo le asignamos una variable aleatoria discreta que toma valores en el conjunto $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ de los cuatro nucleótidos. Las secuencias de ADN de las especies actuales dan observaciones de las variables aleatorias en las hojas del árbol, por lo que a las variables aleatorias en las hojas del árbol se las llama variables «observadas». Por su parte, las variables aleatorias en los nodos interiores son «ocultas», puesto que no disponemos del ADN de los ancestros comunes. Las probabilidades de mutación entre un nucleótido en un nodo y el siguiente se recogen en una matriz de Markov M^i (*matriz de transición*) sobre la arista e_i que une los dos nodos.

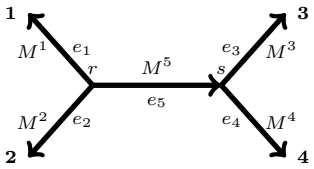

Figura 3: Proceso de Markov sobre el árbol $T_{12|34}$.

La entradas $M^i_{x,y}$ de M^i representan la probabilidad $P(y \mid x, e_i)$ de que un nucleótido x en el nodo padre a sea sustituido por un nucleótido y en el nodo hijo b a lo largo del proceso evolutivo representado por la rama $e_i : a \to b$,

$$M^i = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{pmatrix} P(\mathtt{A} \mid \mathtt{A}, e_i) & P(\mathtt{C} \mid \mathtt{A}, e_i) & P(\mathtt{G} \mid \mathtt{A}, e_i) & P(\mathtt{T} \mid \mathtt{A}, e_i) \\ P(\mathtt{A} \mid \mathtt{C}, e_i) & P(\mathtt{C} \mid \mathtt{C}, e_i) & P(\mathtt{G} \mid \mathtt{C}, e_i) & P(\mathtt{T} \mid \mathtt{C}, e_i) \\ P(\mathtt{A} \mid \mathtt{G}, e_i) & P(\mathtt{C} \mid \mathtt{G}, e_i) & P(\mathtt{G} \mid \mathtt{G}, e_i) & P(\mathtt{T} \mid \mathtt{G}, e_i) \\ P(\mathtt{A} \mid \mathtt{T}, e_i) & P(\mathtt{C} \mid \mathtt{T}, e_i) & P(\mathtt{G} \mid \mathtt{T}, e_i) & P(\mathtt{T} \mid \mathtt{T}, e_i) \end{pmatrix} \end{array}.$$

Las entradas de M^i son desconocidas para nosotros y, junto con la distribución π de nucleótidos en la raíz, son los *parámetros* del modelo. Dependiendo de los valores de los parámetros se producirán más o menos sustituciones en la arista e_i. Nótese que, puesto que las entradas de la matriz son probabilidades condicionadas, las filas de las matrices de transición M^i tienen que sumar 1.

Considerando el proceso de Markov mencionado arriba, podemos escribir la probabilidad de observar nucleótidos en las hojas del árbol en términos de los parámetros del modelo. Por ejemplo, si llamamos $p^T_{x_1x_2x_3x_4}$ a la probabilidad de observar el nucleótido x_i en la especie i del árbol $T = T_{12|34}$ de la figura 3, esta probabilidad se expresa en función de las entradas de las matrices de transición M^i y de π de la siguiente forma:

$$p^T_{x_1x_2x_3x_4} = \sum_{x_r,x_s\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}} \pi_{x_r} M^1_{x_r,x_1} M^2_{x_r,x_2} M^5_{x_r,x_s} M^3_{x_s,x_3} M^4_{x_s,x_4}, \tag{1}$$

donde π_x es la probabilidad del estado x en el nodo r, $P(X_r = x)$, y la suma se hace sobre todos los estados ocultos en los nodos r y s.

Según la estructura que se permita en las matrices de transición obtenemos modelos de evolución más o menos complejos. Por ejemplo, si no imponemos ninguna restricción, resulta el modelo más general posible, llamado el *modelo general de Markov* o *GMM* [26, 3].

Pero, teniendo en cuenta las propiedades químicas de los nucleótidos, se han propuesto otros modelos. Por ejemplo, la adenina y la guanina forman parte de las *pirimidinas*, que tienen dos anillos en su estructura molecular, mientras que la citosina y la timina son *purinas* (un solo anillo). Las mutaciones de purinas a pirimidinas (o viceversa) se llaman *transversiones*, y las mutaciones dentro del mismo grupo se llaman *transiciones* (véase la figura 4). Las transiciones son más frecuentes que las

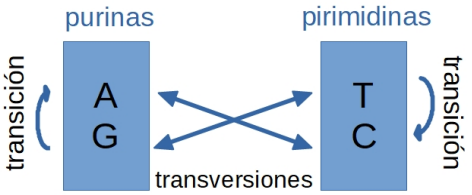

Figura 4: Las mutaciones de los nucleótidos.

transversiones porque preservan el número de anillos del nucleótido. Basándose en esta observación, Kimura propuso en 1981 (véase [19]) el siguiente modelo de sustitución de nucleótidos: un parámetro para cualquiera de las dos transiciones y un parámetro para cada tipo de transversión. Así, en este modelo, llamado *Kimura 3-parámetros* o, más brevemente, *K81*, las matrices de transición tienen 3 parámetros

libres por arista y son de la forma

$$M^i = \begin{pmatrix} a_i & b_i & c_i & d_i \\ b_i & a_i & d_i & c_i \\ c_i & d_i & a_i & b_i \\ d_i & c_i & b_i & a_i \end{pmatrix},$$

con $a_i + b_i + c_i + d_i = 1$.

Observamos que $\pi = (1/4, 1/4, 1/4, 1/4)$ es la distribución estacionaria para cualquier matriz M del modelo K81: $\pi M = \pi$. Además este modelo es *reversible* en el tiempo: $\pi_x M_{x,y} = \pi_y M_{y,x}$ para cualquier par de nucleótidos x, y, por lo que se asume que la distribución en la raíz es también la estacionaria (la uniforme en este caso). Los modelos que satisfacen esta propiedad de reversibilidad son los más usados actualmente, aunque esta propiedad es a menudo motivo de controversia [28].

Si además imponemos $b_i = d_i$, tenemos el modelo de *Kimura 2-parámetros* [18], y si nos restringimos a $b_i = c_i = d_i$ obtenemos el modelo más sencillo de *Jukes-Cantor* [15]. Todos estos modelos son ejemplos de los llamados modelos *equivariantes* [9, 5], puesto que la acción de un subgrupo $G \subseteq \mathfrak{S}_4$ deja invariantes las matrices de transición.

Si T es el árbol de la figura 3, entonces el modelo de K81 sobre T tiene 3×5 parámetros libres $b_1, \ldots, b_5, c_1, \ldots, c_5, d_1, \ldots, d_5$ (a_i se expresa en función de éstos). Las distribuciones conjuntas de nucleótidos en las hojas de T que corresponden a alguna distribución bajo el modelo K81 son los puntos de la imagen de la siguiente aplicación polinomial (si restringimos los parámetros al símplice correspondiente):

$$\begin{array}{rcl} \varphi_T : \mathbb{R}^{15} & \longrightarrow & \mathbb{R}^{4^4} \\ (M^1, \ldots, M^5) = (b_1, \ldots, d_5) & \longmapsto & p_T = (p^T_{\mathtt{AAAA}}, p^T_{\mathtt{AAAC}}, p^T_{\mathtt{AAAG}}, \ldots, p^T_{\mathtt{TTTT}}), \end{array} \tag{2}$$

donde $p^T_{x_1x_2x_3x_4}$ se calcula como en (1).

Como veremos, el modelo K81 tiene unas propiedades matemáticas relevantes que, explorándolas y sacándoles el máximo partido, contribuyen a mejorar los métodos de reconstrucción filogenética que se basan en este modelo.

3. La transformada de Hadamard

En el caso del modelo K81, se pueden usar las simetrías de las matrices de transición para descubrir una estructura extra en la parametrización (2). La clave está en un cambio de base que diagonaliza todas las matrices del modelo K81.

Sea M una matriz del modelo K81,

$$M = \begin{pmatrix} a & b & c & d \\ b & a & d & c \\ c & d & a & b \\ d & c & b & a \end{pmatrix}, \quad a + b + c + d = 1. \tag{3}$$

Observamos que el vector $(1, 1, 1, 1)^t$ (donde el superíndice t indica transposición) es un vector propio de M de valor propio $1 = a+b+c+d$. Asimismo es fácil comprobar

que los vectores $(1,1,-1,-1)^t$, $(1,-1,1,-1)^t$, y $(1,-1,-1,1)^t$ son también vectores propios de valores propios $a+b-c-d$, $a-b+c-d$ y $a-b-c+d$, respectivamente. La matriz de cambio de base es la *matriz de Hadamard*

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

y a la base de vectores propios la llamaremos $\bar{\Sigma} = \{\bar{\mathtt{A}}, \bar{\mathtt{C}}, \bar{\mathtt{G}}, \bar{\mathtt{T}}\}$, donde

$$\bar{\mathtt{A}} = (1,1,1,1)^t, \qquad \bar{\mathtt{G}} = (1,-1,1,-1)^t,$$
$$\bar{\mathtt{C}} = (1,1,-1,-1)^t, \qquad \bar{\mathtt{T}} = (1,-1,-1,1)^t.$$

Una de las propiedades relevantes de la matriz simétrica H es que su inversa es $H^{-1} = \frac{1}{4}H$.

Se llama *transformada de Hadamard* de una matriz M a

$$\bar{M} = H^{-1} \cdot M \cdot H.$$

Hemos visto pues que la transformada de Hadamard de una matriz M del modelo K81 es la matriz diagonal

$$\bar{M} = \mathrm{diag}(m_{\mathtt{A}}, m_{\mathtt{C}}, m_{\mathtt{G}}, m_{\mathtt{T}}),$$

donde usamos la notación $m_{\mathtt{A}} = a+b+c+d = 1$, $m_{\mathtt{C}} = a+b-c-d$, $m_{\mathtt{G}} = a-b+c-d$ y $m_{\mathtt{T}} = a-b-c+d$.

El uso de la transformada de Hadamard en filogenética fue introducido por Hendy y Penny en 1989 [14]. La base de vectores propios $\bar{\Sigma}$ también se conoce como *base de Fourier*. Esta nomenclatura es debida a la equivalencia que explicamos a continuación.

3.1. El modelo K81 como modelo de grupo

Consideramos la biyección entre el conjunto $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ y el grupo $G = \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$ dada por

$$\mathtt{A} \leftrightarrow (0,0), \quad \mathtt{C} \leftrightarrow (0,1), \quad \mathtt{G} \leftrightarrow (1,0), \quad \mathtt{T} \leftrightarrow (1,1).$$

Con esta biyección podemos pensar en Σ como el grupo aditivo G con elemento neutro $\mathtt{A}$ y con la operación $+$ que cumple $\mathtt{C}+\mathtt{C} = \mathtt{G}+\mathtt{G} = \mathtt{T}+\mathtt{T} = \mathtt{A}$, $\mathtt{C}+\mathtt{G} = \mathtt{T}$, $\mathtt{G}+\mathtt{T} = \mathtt{C}$, $\mathtt{C}+\mathtt{T} = \mathtt{G}$. Este grupo tiene cuatro caracteres, es decir morfismos $\chi_i : G \longrightarrow (\mathbb{C}^*, \cdot)$, que se definen como en la tabla siguiente (la entrada (i,j) de la tabla es el valor $\chi_i(j)$ de χ_i sobre el elemento j):

	A	C	G	T
$\chi_{\mathtt{A}}$	1	1	1	1
$\chi_{\mathtt{C}}$	1	1	−1	−1
$\chi_{\mathtt{G}}$	1	−1	1	−1
$\chi_{\mathtt{T}}$	1	−1	−1	1

El conjunto de caracteres de un grupo forma su *grupo dual* $\tilde{G}$, que en este caso es isomorfo a G.

Las matrices de transición del modelo K81 se pueden pensar como funciones de G en $\mathbb{C}$: la entrada (i, j) de la matriz M no depende de i, j sino sólo de su diferencia $i - j$ como elementos del grupo G. Los modelos que cumplen esta propiedad se llaman *modelos basados en grupos*. Con esta interpretación, la matriz M de (3) se corresponde con la función $f : G \longrightarrow \mathbb{C}$ que toma valores $f(\mathtt{A}) = a$, $f(\mathtt{C}) = b$, $f(\mathtt{G}) = c$, $f(\mathtt{T}) = d$ vía $M_{i,j} = f(i - j)$. Una vez tenemos la matriz de transición descrita como una función f, podemos considerar su *transformada de Fourier discreta* sobre el grupo G, $\tilde{f} : \tilde{G} \longrightarrow \mathbb{C}$ definida como

$$\tilde{f}(\chi) = \sum_{g \in G} \bar{\chi}(g) f(g),$$

donde la barra indica el conjugado (si el grupo tuviera caracteres no reales).

Si pensamos en f como el vector $(a, b, c, d) \in \mathbb{R}^4$, vemos que la transformada de Fourier no es más que una aplicación lineal cuya matriz es la tabla de caracteres del grupo. En nuestro caso, pues, la transformada de Fourier discreta coincide con la transformada de Hadamard y $\tilde{f} = (m_{\mathtt{A}}, m_{\mathtt{C}}, m_{\mathtt{G}}, m_{\mathtt{T}})$. Evans y Speed [11] introdujeron la transformada de Fourier discreta en el estudio del modelo K81, y se puede encontrar una generalización a otros modelos evolutivos en el trabajo de Sturmfels y Sullivant [27].

3.2. Cambio de coordenadas

Denotamos por Σ la base natural de $\mathbb{R}^4$. Puesto que las matrices de transición las hemos indexado usando los cuatro nucleótidos, podemos incluso identificar el conjunto $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\}$ con Σ.

Definición 3.1. Dado un vector $p \in \mathbb{R}^4$, llamamos $p_\Sigma = (p_{\mathtt{A}}, p_{\mathtt{C}}, p_{\mathtt{G}}, p_{\mathtt{T}})$ a sus coordenadas en la base Σ, y $p_{\bar{\Sigma}} = (\bar{p}_{\mathtt{A}}, \bar{p}_{\mathtt{C}}, \bar{p}_{\mathtt{G}}, \bar{p}_{\mathtt{T}})$ a sus coordenadas en la base $\bar{\Sigma}$, $p = \sum_{x \in \Sigma} p_x x = \sum_{x \in \Sigma} \bar{p}_x \bar{x}$. Las coordenadas $p_{\bar{\Sigma}}$ se llaman *coordenadas de Fourier*.

Si disponemos de las coordenadas en la base Σ, sólo necesitamos aplicar H^{-1} para obtener las coordenadas en la base $\bar{\Sigma}$,

$$p^t_{\bar{\Sigma}} = H^{-1} p^t_\Sigma = \frac{1}{4} H p^t_\Sigma.$$

Por ejemplo, para la distribución uniforme π tenemos $\pi_\Sigma = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ y $\pi_{\bar{\Sigma}} = (\frac{1}{4}, 0, 0, 0)$.

Se puede identificar el espacio de llegada de la aplicación polinómica (2), $\mathbb{R}^{4^4}$, con $\mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4$. Puesto que si consideramos una base de $\mathbb{R}^4$ (por ejemplo Σ o $\bar{\Sigma}$), se obtiene de forma natural una base de $\mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4$, podemos definir de forma análoga las coordenadas de Fourier de un tensor $p \in \mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4$:

Definición 3.2. Si p es un tensor de $\mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4$, llamamos

$$p_\Sigma = (p_{\mathtt{AAAA}}, p_{\mathtt{AAAC}}, \ldots, p_{\mathtt{TTTT}})$$

a sus coordenadas en la base $\mathtt{A}\otimes\mathtt{A}\otimes\mathtt{A}\otimes\mathtt{A},\ldots,\mathtt{T}\otimes\mathtt{T}\otimes\mathtt{T}\otimes\mathtt{T}$, y

$$p_{\bar{\Sigma}}=(\bar{p}_{\mathtt{AAAA}},\bar{p}_{\mathtt{AAAC}},\ldots,\bar{p}_{\mathtt{TTTT}})$$

a sus coordenadas en la base $\bar{\mathtt{A}}\otimes\bar{\mathtt{A}}\otimes\bar{\mathtt{A}}\otimes\bar{\mathtt{A}},\ldots,\bar{\mathtt{T}}\otimes\bar{\mathtt{T}}\otimes\bar{\mathtt{T}}\otimes\bar{\mathtt{T}}$, de forma que

$$p=\sum_{x_1,x_2,x_3,x_4\in\Sigma}p_{x_1x_2x_3x_4}x_1\otimes x_2\otimes x_3\otimes x_4=\sum_{x_1,x_2,x_3,x_4\in\Sigma}\bar{p}_{x_1x_2x_3x_4}\bar{x}_1\otimes\bar{x}_2\otimes\bar{x}_3\otimes\bar{x}_4.$$

Para cambiar de base las coordenadas de un tensor de $\mathbb{R}^4\otimes\mathbb{R}^4\otimes\mathbb{R}^4\otimes\mathbb{R}^4$ sólo hay que aplicar el producto de Kronecker de H:

$$p^t_{\bar{\Sigma}}=\left(H^{-1}\otimes H^{-1}\otimes H^{-1}\otimes H^{-1}\right)p^t_{\Sigma}=\frac{1}{4^4}\left(H\otimes H\otimes H\otimes H\right)p^t_{\Sigma}.$$

Recordamos que si M y N son matrices 4×4, el producto de Kronecker $M\otimes N$ es la matriz 16×16 que tiene en la fila (i,j) y columna (k,l) la entrada $M_{i,k}N_{j,l}$, $i,j,k,l\in\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$.

Como veremos en la sección 4, este cambio de coordenadas nos permitirá escribir de forma mucho más sencilla la expresión (1).

4. Invariantes para el modelo Kimura 3-parámetros

Los biólogos Cavender y Felsenstein [7] llamaron *invariantes filogenéticos del árbol T* a aquellos polinomios en las variables $p_{x_1x_2x_3x_4}$ que se anulan sobre cualquier punto p en la imagen de φ_T. En lenguaje de geometría algebraica, se llama invariantes filogenéticos de T a los elementos del ideal de la clausura algebraica de $\mathrm{Im}(\varphi_T)$. Describiremos primero invariantes filogenéticos para un modelo general de Markov y luego encontraremos invariantes específicos para el modelo K81.

4.1. Flattening

Definimos el *flattening* de $p\in\mathbb{R}^{4^4}$ respecto a la bipartición $12|34$ como la reorganización del vector en la matriz siguiente:

$$\mathrm{flatt}_{12|34}(p)\ =\ \begin{array}{c}\\ \text{estados en}\\ \text{las hojas}\\ 1\text{ y }2\end{array}\ \begin{array}{c}\text{estados en las hojas 3 y 4}\\ \begin{pmatrix}p_{\mathtt{AAAA}} & p_{\mathtt{AAAC}} & p_{\mathtt{AAAG}} & \cdots & p_{\mathtt{AATT}}\\ p_{\mathtt{ACAA}} & p_{\mathtt{ACAC}} & p_{\mathtt{ACAG}} & \cdots & p_{\mathtt{ACTT}}\\ p_{\mathtt{AGAA}} & p_{\mathtt{AGAC}} & p_{\mathtt{AGAG}} & \cdots & p_{\mathtt{AGTT}}\\ \cdots & \cdots & \cdots & \cdots & \cdots\\ p_{\mathtt{TTAA}} & p_{\mathtt{TTAC}} & p_{\mathtt{TTAG}} & \cdots & p_{\mathtt{TTTT}}\end{pmatrix}\end{array}. \tag{4}$$

Es conveniente indexar las filas y las columnas de la matriz con pares de nucleótidos (x,y). Análogamente podemos definir el *flattening* de p respecto a las biparticiones $13|24$ y $14|23$, $\mathrm{flatt}_{13|24}(p)$, $\mathrm{flatt}_{14|23}(p)$.

El nombre «flattening» de esta matriz se debe a que se puede considerar como un «allanamiento» de un tensor $p\in\mathbb{R}^4\otimes\mathbb{R}^4\otimes\mathbb{R}^4\otimes\mathbb{R}^4$: la matriz $\mathrm{flatt}_{12|34}(p)$ es el

resultado de «allanar» este tensor mediante un isomorfismo $\mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4 \otimes \mathbb{R}^4 \cong \mathrm{Hom}(\mathbb{R}^4 \otimes \mathbb{R}^4, \mathbb{R}^4 \otimes \mathbb{R}^4) \cong M_{16\times 16}(\mathbb{R})$.

Si $p \in \mathbb{R}^{4^4}$ proviene de un proceso de Markov en el árbol $T = T_{12|34}$ con matrices de transición $M^1, \ldots, M^5$ y distribución π en la raíz (notación de la figura 3), reescribiendo la expresión (1) obtenemos

$$p^T_{x_1x_2x_3x_4} = \sum_{x_r \in \Sigma} M^1_{x_r,x_1} M^2_{x_r,x_2} \sum_{x_s \in \Sigma} \pi_{x_r} M^5_{x_r,x_s} M^3_{x_s,x_3} M^4_{x_s,x_4}. \tag{5}$$

Podemos ahora usar el producto de Kronecker de matrices para reescribir (5) como

$$p^T_{x_1x_2x_3x_4} = \sum_{x_r \in \Sigma} (M^1 \otimes M^2)^t_{(x_1,x_2)(x_r,x_r)} \sum_{x_s \in \Sigma} \pi_{x_r} M^5_{x_r,x_s} (M^3 \otimes M^4)_{(x_s,x_s)(x_3,x_4)}. \tag{6}$$

Obtenemos, pues,

$$\mathrm{flatt}_{12|34}(p^T) = (M^1 \otimes M^2)^t \cdot \mathrm{flatt}_{12|34}(q) \cdot (M^3 \otimes M^4),$$

donde

$$q_{xyzt} = \begin{cases} \pi_x M^5_{x,z}, & \text{si } x = y, z = t, \\ 0, & \text{en caso contrario.} \end{cases}$$

Nótese que q puede entenderse como la distribución que viene del árbol T con matrices de transición en las aristas exteriores igual a la identidad (dejando la distribución π en r y la matriz M^5 en la arista interior). De forma similar obtenemos

$$\mathrm{flatt}_{13|24}(p^T) = (M^1 \otimes M^2)^t \cdot \mathrm{flatt}_{13|24}(q) \cdot (M^3 \otimes M^4),$$
$$\mathrm{flatt}_{14|23}(p^T) = (M^1 \otimes M^2)^t \cdot \mathrm{flatt}_{14|23}(q) \cdot (M^3 \otimes M^4).$$

La matriz del medio en estas expresiones es fácil de entender: en $\mathrm{flatt}_{12|34}(q)$ la entrada en la fila (x_1, x_2) y columna (x_3, x_4) es 0 si $x_1 \neq x_2$ o bien $x_3 \neq x_4$, y es $\pi_{x_1} M^5_{x_1,x_3}$ si $x_1 = x_2$ y $x_3 = x_4$; por otra parte, la entrada en la fila (x_1, x_3) y columna (x_2, x_4) de $\mathrm{flatt}_{13|24}(q)$ es 0 si $(x_1, x_3) \neq (x_2, x_4)$, y es $\pi_{x_1} M^5_{x_1,x_3}$ si $(x_1, x_3) = (x_2, x_4)$; y la entrada en la fila (x_1, x_4) y columna (x_2, x_3) de $\mathrm{flatt}_{14|23}(q)$ es 0 si $(x_1, x_4) \neq (x_2, x_3)$, y es $\pi_{x_1} M^5_{x_1,x_3}$ si $(x_1, x_4) = (x_2, x_3)$. Así, $\mathrm{flatt}_{13|24}(q)$ y $\mathrm{flatt}_{14|23}(q)$ son matrices diagonales mientras que $\mathrm{flatt}_{12|34}(q)$ tiene sólo 4 de sus 16 filas distintas de 0.

Para parámetros suficientemente genéricos, $\mathrm{flatt}_{12|34}(q)$ tiene rango 4, mientras que $\mathrm{flatt}_{13|24}(q)$ y $\mathrm{flatt}_{14|23}(q)$ tienen rango 16. Por otra parte, es bien conocido que el rango de un producto de Kronecker de matrices es el producto de rangos. Así pues, hemos demostrado el siguiente resultado:

Teorema 4.1 ([2]). *Si $p \in \mathbb{R}^{4^4}$ proviene de un proceso de Markov en $T_{12|34}$ en el que la distribución en la raíz y las matrices de transición son suficientemente genéricas, entonces* $\mathrm{flatt}_{12|34}(p)$ *tiene rango* 4, *mientras que* $\mathrm{flatt}_{13|24}(p)$ *y* $\mathrm{flatt}_{14|23}(p)$ *tienen rango* 16.

Este resultado, debido a Allman y Rhodes, ha sido usado recientemente para proponer métodos de reconstrucción filogenética. La idea básica es usar el resultado de Eckart-Young [10] que da la distancia (en norma de Frobenius o espectral) de una matriz al conjunto de matrices de cierto rango dado. Para calcular esta distancia sólo hace falta encontrar la descomposición en valores singulares (SVD) de la matriz. Por ejemplo, los métodos Split Scores [1] y Erik+2 [12] presentan adaptaciones de esta idea y SVDQuartets de [8] amplía el modelo base al modelo de coalescencia.

También podemos considerar el *flattening* del tensor en coordenadas de Fourier. Llamamos $\overline{\text{flatt}}_{ij|kl}(p)$ al *flattening* del vector p expresado en dichas coordenadas, es decir $\text{flatt}_{ij|kl}\left((H^{-1}\otimes H^{-1}\otimes H^{-1}\otimes H^{-1})p\right)$. Así, por ejemplo, $\overline{\text{flatt}}_{12|34}(p)$ es la matriz (4), pero en coordenadas $\bar{p}_\Sigma$ en vez de p_Σ. Como veremos en el siguiente apartado, cuando analicemos el modelo K81 es conveniente considerar $\overline{\text{flatt}}_{12|34}(p)$.

4.2. Parametrización monomial

La observación de que el cambio de coordenadas mencionado arriba, tanto en el espacio de parámetros como en el espacio de distribuciones, transforma la parametrización polinomial (2) en una aplicación monomial fue demostrada por primera vez por Evans y Speed, pero de hecho está subyacente en el trabajo de Hendy y Penny citado anteriormente. La demostración original de Evans y Speed se basaba en usar las propiedades de la transformada de Fourier de una convolución, pero aquí vamos a dar una demostración usando técnicas de álgebra lineal y multilineal.

Teorema 4.2. *Sea p en la imagen de la aplicación polinomial* (2)*, $p=\varphi_T(M^1,M^2,M^3,M^4,M^5)$, donde M^i es una matriz de Kimura 3-parámetros asociada a la arista e_i del árbol $T=T_{12|34}$ de la figura 3. Si los valores propios de M^i son $(m^i_{\mathtt{A}},m^i_{\mathtt{C}},m^i_{\mathtt{G}},m^i_{\mathtt{T}})$, entonces se tiene*

$$\bar{p}_{x_1x_2x_3x_4}=\begin{cases}\dfrac{1}{4^4}\cdot m^1_{x_1}m^2_{x_2}m^5_{x_1+x_2}m^3_{x_3}m^4_{x_4}, & \text{si } x_1+x_2=x_3+x_4,\\ 0, & \text{en caso contrario}\end{cases}$$

(donde la suma de nucleótidos se calcula como en la sección 3.1).

Como consecuencia de este teorema tenemos que la aplicación (2) escrita en coordenadas de Fourier es una aplicación monomial.

Demostración. Consideramos primero el caso en que M^1,M^2,M^3,M^4 sean la matriz identidad y llamamos q al vector correspondiente. En estas circunstancias,

$$q_{x_1x_2x_3x_4}=\begin{cases}\dfrac{1}{4}M^5_{x,y}, & \text{si } x=x_1=x_2 \text{ e } y=x_3=x_4,\\ 0, & \text{en caso contrario.}\end{cases}$$

Tenemos

$$q^t_\Sigma=\frac{1}{4^4}\left(H\otimes H\otimes H\otimes H\right)q^t_\Sigma,$$

luego

$$\begin{aligned}\bar{q}_{x_1x_2x_3x_4} &= \frac{1}{4^4}\sum_{y_1,y_2,y_3,y_4} H_{x_1,y_1}H_{x_2,y_2}H_{x_3,y_3}H_{x_4,y_4}q_{y_1y_2y_3y_4}\\ &= \frac{1}{4^5}\sum_{y,z} H_{x_1,y}H_{x_2,y}H_{x_3,z}H_{x_4,z}M^5_{y,z}\\ &= \frac{1}{4^5}\sum_{y} H_{x_1,y}H_{x_2,y}\sum_{z} H_{x_3,z}H_{x_4,z}M^5_{y,z}.\end{aligned}$$

Es fácil ver que la matriz H cumple $H_{x_3,z}H_{x_4,z} = H_{x_3+x_4,z}$ (de hecho, esto es inmediato si se piensa H como la tabla de caracteres del grupo $\mathbb{Z}/2\mathbb{Z}\times\mathbb{Z}/2\mathbb{Z}$). Así,

$$\begin{aligned}\bar{q}_{x_1x_2x_3x_4} &= \frac{1}{4^5}\sum_{y} H_{x_1+x_2,y}\sum_{z} H_{x_3+x_4,z}M^5_{y,z} = \frac{1}{4^5}\sum_{y} H_{x_1+x_2,y}(H(M^5)^t)_{x_3+x_4,y}\\ &= \frac{1}{4^5}\left(H(H(M^5)^t)^t\right)_{x_1+x_2,x_3+x_4} = \frac{1}{4^5}\left(HM^5H^t\right)_{x_1+x_2,x_3+x_4}\\ &= \frac{1}{4^4}\left(HM^5H^{-1}\right)_{x_1+x_2,x_3+x_4} = \frac{1}{4^4}\bar{M}^5_{x_1+x_2,x_3+x_4},\end{aligned}$$

y obtenemos

$$\bar{q}_{x_1x_2,x_3x_4} = \begin{cases}\dfrac{1}{4^4}\cdot m^5_{x_1+x_2}, & \text{si } x_1+x_2 = x_3+x_4,\\ 0, & \text{en caso contrario.}\end{cases}$$

Ahora, puesto que

$$\mathrm{flatt}_{12|34}(p) = (M^1\otimes M^2)^t\cdot\mathrm{flatt}_{12|34}(q)\cdot(M^3\otimes M^4),$$

tenemos que $\overline{\mathrm{flatt}}_{12|34}(p)$ es igual a

$$\begin{aligned}&\mathrm{flatt}_{12|34}\left((H^{-1}\otimes H^{-1}\otimes H^{-1}\otimes H^{-1})p\right)\\ &\quad= (H^{-1}\otimes H^{-1})\cdot\mathrm{flatt}_{12|34}(p)\cdot(H^{-1}\otimes H^{-1})^t\\ &\quad= (H^{-1}\otimes H^{-1})\cdot(M^1\otimes M^2)^t\cdot\mathrm{flatt}_{12|34}(q)\cdot(M^3\otimes M^4)\cdot(H^{-1}\otimes H^{-1})^t\\ &\quad= (H^{-1}\otimes H^{-1})\cdot(M^1\otimes M^2)^t\cdot(H\otimes H)\cdot(H^{-1}\otimes H^{-1})\cdot\mathrm{flatt}_{12|34}(q)\\ &\qquad\cdot(H^{-1}\otimes H^{-1})^t\cdot(H\otimes H)^t\cdot(M^3\otimes M^4)\cdot(H^{-1}\otimes H^{-1})^t\\ &\quad= (\bar{M}_1\otimes\bar{M}_2)^t\cdot\overline{\mathrm{flatt}}_{12|34}(q)\cdot(\bar{M}_3\otimes\bar{M}_4).\end{aligned}$$

La primera igualdad es fácil de demostrar usando álgebra lineal, y en la última hemos usado que el producto de Kronecker cumple las propiedades $(A\otimes B)(M\otimes N) = AM\otimes BN$ y $(A\otimes B)^t = A^t\otimes B^t$.

Junto con lo que hemos probado anteriormente, y usando que $\bar{M}_1$, $\bar{M}_2$, $\bar{M}_3$ y $\bar{M}_4$ son matrices diagonales, obtenemos

$$\bar{p}_{x_1x_2x_3x_4} = \begin{cases}\dfrac{1}{4^4}\cdot m^1_{x_1}m^2_{x_2}m^5_{x_1+x_2}m^3_{x_3}m^4_{x_4}, & \text{si } x_1+x_2 = x_3+x_4,\\ 0, & \text{en caso contrario.}\end{cases}$$

□

4.3. Invariantes del modelo y de la topología

El teorema 4.2 tiene distintas consecuencias, pero la más directa es que permite el cálculo de los invariantes filogenéticos. En efecto, si una variedad tiene una parametrización monomial, su ideal se puede generar por binomios. Usando este teorema, Sturmfels y Sullivant dieron una colección de invariantes para este modelo (y sus submodelos) en [27]. En esta sección explicaremos cómo calcular los invariantes más relevantes para el modelo K81 en árboles de cuatro hojas. Los lectores interesados en geometría algebraica habrán notado que otra de las consecuencias de este teorema es que permite demostrar que la variedad $\overline{\mathrm{Im}(\varphi_T)}$ es una variedad tórica.

Nótese que, como estamos en el grupo $\mathbb{Z}/\mathbb{Z}_2 \times \mathbb{Z}/\mathbb{Z}_2$,

$$x_1 + x_2 = x_3 + x_4 \Leftrightarrow x_1 + x_3 = x_2 + x_4 \Leftrightarrow x_1 + x_4 = x_3 + x_4.$$

Por otra parte, el teorema 4.2 se puede aplicar tanto al árbol $T_{12|34}$ como, permutando las hojas $\{1, 2, 3, 4\}$, a los otros dos árboles $T_{13|24}$ y $T_{14|23}$. Entonces obtenemos que

$$\bar{p}_{x_1x_2x_3x_4} = 0 \quad \text{si} \quad x_1 + x_2 \neq x_3 + x_4 \tag{7}$$

se cumple si p es un tensor que está en la imagen de φ_T, siendo T cualquiera de los tres árboles $T_{12|34}, T_{13|24}, T_{14|23}$.

Hemos visto pues que las ecuaciones (7) dan invariantes filogenéticos para los tres árboles (por ejemplo $\bar{p}_{\mathtt{AAAC}}$). A los polinomios que son invariantes filogenéticos para todos los árboles los llamamos *invariantes del modelo*. Los invariantes del modelo no aportan ninguna información sobre la forma del árbol, sino que son pura consecuencia del modelo evolutivo que se ha escogido, en este caso el Kimura 3-parámetros.

Sin embargo, vamos a ver que hay polinomios que son invariantes filogenéticos para algunos árboles pero no para todos; este tipo de invariantes se llaman *invariantes de la topología*. Consideramos un tensor p como en el teorema 4.2 y nucleótidos $x_1, x_2, x_3, x_4, x_1', x_2', x_3', x_4'$ que cumplan $x_1 + x_2 = x_1' + x_2' = x_3 + x_4 = x_3' + x_4'$. Entonces, por el teorema 4.2,

$$\bar{p}_{x_1x_2x_3x_4}\bar{p}_{x_1'x_2'x_3'x_4'} = \bar{p}_{x_1'x_2'x_3x_4}\bar{p}_{x_1x_2x_3'x_4'}. \tag{8}$$

En efecto, los dos lados de la igualdad coinciden con

$$\frac{1}{4^8} \cdot m^1_{x_1} m^2_{x_2} m^3_{x_3} m^4_{x_4} m^1_{x_1'} m^2_{x_2'} m^3_{x_3'} m^4_{x_4'} (m^5_{x_1+x_2})^2$$

según el teorema 4.2. La ecuación (8) se puede escribir también como

$$\det \begin{pmatrix} \bar{p}_{x_1x_2x_3x_4} & \bar{p}_{x_1x_2x_3'x_4'} \\ \bar{p}_{x_1'x_2'x_3x_4} & \bar{p}_{x_1'x_2'x_3'x_4'} \end{pmatrix} = 0. \tag{9}$$

Dicho de otra forma, si reordenamos las filas y columnas de la matriz $\overline{\mathrm{flatt}}_{12|34}(p)$ como $\mathtt{AA}, \mathtt{CC}, \mathtt{GG}, \mathtt{TT}, \mathtt{AC}, \mathtt{CA}, \mathtt{GT}, \mathtt{TG}, \mathtt{AG}, \mathtt{CT}, \mathtt{GA}, \mathtt{TG}, \mathtt{AT}, \mathtt{CG}, \mathtt{GC}, \mathtt{TA}$ (de forma que los nucleótidos que indexan las cuatro primeras filas/columnas suman $\mathtt{A}$, y las siguientes

C, G y T, respectivamente) obtenemos una matriz diagonal por bloques:

$$\begin{pmatrix} B_{\mathtt{A}} & & & \\ & B_{\mathtt{C}} & & \\ & & B_{\mathtt{G}} & \\ & & & B_{\mathtt{T}} \end{pmatrix}$$

donde

$$B_{\mathtt{A}} = \begin{pmatrix} \bar{p}_{\mathtt{AAAA}} & \bar{p}_{\mathtt{AACC}} & \bar{p}_{\mathtt{AAGG}} & \bar{p}_{\mathtt{AATT}} \\ \bar{p}_{\mathtt{CCAA}} & \bar{p}_{\mathtt{CCCC}} & \bar{p}_{\mathtt{CCGG}} & \bar{p}_{\mathtt{CCTT}} \\ \bar{p}_{\mathtt{GGAA}} & \bar{p}_{\mathtt{GGCC}} & \bar{p}_{\mathtt{GGGG}} & \bar{p}_{\mathtt{GGTT}} \\ \bar{p}_{\mathtt{TTAA}} & \bar{p}_{\mathtt{TTCC}} & \bar{p}_{\mathtt{TTGG}} & \bar{p}_{\mathtt{TTTT}} \end{pmatrix}, \quad B_{\mathtt{C}} = \begin{pmatrix} \bar{p}_{\mathtt{ACAC}} & \bar{p}_{\mathtt{ACCA}} & \bar{p}_{\mathtt{ACGT}} & \bar{p}_{\mathtt{ACTG}} \\ \bar{p}_{\mathtt{CAAC}} & \bar{p}_{\mathtt{CACA}} & \bar{p}_{\mathtt{CAGT}} & \bar{p}_{\mathtt{CATG}} \\ \bar{p}_{\mathtt{GTAC}} & \bar{p}_{\mathtt{GTCA}} & \bar{p}_{\mathtt{GTGT}} & \bar{p}_{\mathtt{GTTG}} \\ \bar{p}_{\mathtt{TGAC}} & \bar{p}_{\mathtt{TGCA}} & \bar{p}_{\mathtt{TGGT}} & \bar{p}_{\mathtt{TGTG}} \end{pmatrix},$$

$$B_{\mathtt{G}} = \begin{pmatrix} \bar{p}_{\mathtt{AGAG}} & \bar{p}_{\mathtt{AGCT}} & \bar{p}_{\mathtt{AGGA}} & \bar{p}_{\mathtt{AGTC}} \\ \bar{p}_{\mathtt{CTAG}} & \bar{p}_{\mathtt{CTCT}} & \bar{p}_{\mathtt{CTGA}} & \bar{p}_{\mathtt{CTTC}} \\ \bar{p}_{\mathtt{GAAG}} & \bar{p}_{\mathtt{GACT}} & \bar{p}_{\mathtt{GAGA}} & \bar{p}_{\mathtt{GATC}} \\ \bar{p}_{\mathtt{TCAG}} & \bar{p}_{\mathtt{TCCT}} & \bar{p}_{\mathtt{TCGA}} & \bar{p}_{\mathtt{TCTC}} \end{pmatrix}, \quad B_{\mathtt{T}} = \begin{pmatrix} \bar{p}_{\mathtt{ATAT}} & \bar{p}_{\mathtt{ATCG}} & \bar{p}_{\mathtt{ATGC}} & \bar{p}_{\mathtt{ATTA}} \\ \bar{p}_{\mathtt{CGAT}} & \bar{p}_{\mathtt{CGCG}} & \bar{p}_{\mathtt{CGGC}} & \bar{p}_{\mathtt{CGTA}} \\ \bar{p}_{\mathtt{GCAT}} & \bar{p}_{\mathtt{GCCG}} & \bar{p}_{\mathtt{GCGC}} & \bar{p}_{\mathtt{GCTA}} \\ \bar{p}_{\mathtt{TAAT}} & \bar{p}_{\mathtt{TACG}} & \bar{p}_{\mathtt{TAGC}} & \bar{p}_{\mathtt{TATA}} \end{pmatrix}.$$

Los pares de nucleótidos que indexan las filas y las columnas de cada bloque $B_{\mathtt{x}}$ suman x, y cada bloque $B_{\mathtt{x}}$ tiene rango 1 debido a (9). En el caso del modelo K81, el teorema 4.1 se traduce en esta condición de rango 1 para cada bloque (en particular, la matriz del *flattening* sigue teniendo rango 4).

Para poder estimar el árbol filogenético en base a modelos evolutivos, se asume que las matrices de transición son «próximas» a la identidad (puesto que, si están muy lejos de la identidad, las secuencias pueden mutar casi independientemente unas de otras y difícilmente se podría representar su evolución en un árbol filogenético). Como consecuencia, podemos suponer que los valores propios de las matrices de transición son distintos de cero (o incluso que no están muy lejos de 1). En particular, podemos suponer que la entrada $\bar{p}_{\mathtt{xAxA}}$ del bloque $B_{\mathtt{x}}$ es distinta de 0 y podemos traducir la condición de que los bloques tengan rango uno a 9 ecuaciones por bloque (los 9 menores 2×2 que contienen esta entrada, igualados a cero). Es decir, las ecuaciones del tipo (8) se pueden reducir a 36 ecuaciones.

Se puede demostrar (véase [4]) que estas 36 ecuaciones de grado 2, junto con otras 12 ecuaciones de grado 3, forman un conjunto de 48 ecuaciones que caracterizan los puntos de la imagen de la aplicación polinomial (2) que caen dentro del símplice de probabilidades. Las doce ecuaciones necesarias para completar las 36 de arriba son invariantes del modelo y, por lo tanto, no aportan información sobre el árbol escogido. En consecuencia, los 36 menores 2×2 que hemos seleccionado dan lugar a los únicos invariantes filogenéticos relevantes para determinar la topología de un árbol de cuatro hojas bajo el modelo K81. Este estudio se puede ampliar a árboles de cualquier número de hojas e incluso a otros modelos equivariantes (véase [5, 6]). Nótese que, aunque la parametrización monomial aligera el cálculo de invariantes, no es posible obtenerlos mediante *software* de álgebra computacional más que para árboles de cinco hojas (incluso usando la aplicación monomial); véase la Small Trees Webpage [13] donde hay un listado de los árboles y modelos que se han podido calcular usando algún paquete de álgebra computacional.

5. Conclusiones

Como hemos podido ver, un estudio álgebro-geométrico de las propiedades del modelo evolutivo permite encontrar las ecuaciones que caracterizan las distribuciones de secuencias de ADN que han evolucionado siguiendo un árbol y un modelo determinados. Estudios de este tipo permiten determinar también la dimensión de la variedad filogenética correspondiente y decidir si los parámetros del modelo son identificables (es decir, si hay, salvo permutaciones, un único conjunto de parámetros con significado estadístico para cada distribución en las hojas). Esto pasa por estudiar las antiimágenes de un punto genérico de la imagen de (2).

Las ecuaciones que describen la imagen de (2) las hemos clasificado entre invariantes del modelo e invariantes de la topología. Los invariantes del modelo se han usado para diseñar métodos de selección del modelo más apropiado para los datos (véase [16]). Sin embargo, es difícil que se puedan usar los invariantes de la topología directamente para seleccionar el árbol más apropiado para unas secuencias dadas. Se tendría que decidir primero qué ecuaciones se usan, evaluarlas sobre el punto de datos, decidir estadísticamente si esta evaluación es suficientemente cercana a cero... y, lo que es peor, el número mínimo de ecuaciones necesarias es exponencial en el número de hojas. Por eso es también una buena idea desarrollar primero buenos métodos para reconstruir árboles de cuatro hojas y luego usar alguno de los métodos basados en cuartetos para reconstruir el árbol total (véase, por ejemplo, [23]).

No obstante, las ecuaciones que describen la variedad son también necesarias para otro aspecto de la reconstrucción filogenética, la inferencia de parámetros: es habitual usar la función de verosimilitud para encontrar los parámetros que maximizan la probabilidad de observar los datos dados. Sin embargo, los métodos numéricos habituales no garantizan encontrar un máximo global de esta función. Recientemente (véase [17], por ejemplo), se han usado los invariantes filogenéticos y técnicas de geometría algebraica numérica para encontrar el máximo global (si existe) de la función de verosimilitud.

En vez de usar directamente las ecuaciones para la estimación de la topología, es mejor usar las condiciones de rango que se han mencionado en el teorema 4.1. Como hemos dicho, usando la descomposición en valores singulares es muy fácil calcular la distancia de una matriz al conjunto de matrices de un rango dado. Esta aproximación, que requiere de un estudio algebraico previo de las variedades, está tomando relevancia entre la comunidad de biólogos que se dedican a métodos de reconstrucción de la topología del árbol. Se está convenciendo a los biólogos de que no tener que estimar los parámetros es una gran mejora, sobre todo para poder considerar modelos más complejos. Y los modelos que hemos descrito aquí son, de hecho, más generales que los que consideran habitualmente los biólogos (donde asumen homogeneidad de las velocidades de mutación a lo largo del tiempo). Remitimos el lector interesado en la diferencia entre ambos modelos al artículo [24].

La generalización de estas técnicas a nuevos modelos, la verificación de los resultados teóricos sobre datos simulados, y el trabajo codo a codo con biólogos es lo que permitirá que estas técnicas se asimilen como propias entre los filogeneticistas.

Referencias

[1] E. S. Allman, L. S. Kubatko y J. A. Rhodes, Split scores: A tool to quantify phylogenetic signal in genome-scale data, *Syst. Biol.* **66** (2016), 620–636.

[2] E. S. Allman y J. A. Rhodes, Phylogenetic ideals and varieties for the general Markov model, *Adv. in Appl. Math.* **40** (2008), 127–148.

[3] D. Barry y J. Hartigan, Asynchronous distance between homologous DNA sequences, *Biometrics* **43** (1987), 261–276.

[4] M. Casanellas y J. Fernández-Sánchez, Geometry of the Kimura 3-parameter model, *Adv. in Appl. Math.* **41** (2008), 265–292.

[5] M. Casanellas y J. Fernández-Sánchez, Relevant phylogenetic invariants of evolutionary models, *J. Math. Pures Appl. (9)* **96** (2011), 207–229.

[6] M. Casanellas, J. Fernández-Sánchez y M. Michalek, Complete intersection for equivariant models, *Adv. in Math.* **315** (2017), 285–323.

[7] J. Cavender y J. Felsenstein, Invariants of phylogenies in a simple case with discrete states, *J. Classification* **4** (1987), 57–71.

[8] J. Chifman y L. Kubatko, Quartet inference from SNP data under the coalescent model, *Bioinformatics* **30** (2014), 3317–3324.

[9] J. Draisma y J. Kuttler, On the ideals of equivariants tree models, *Math. Ann.* **344** (2009), 619–644.

[10] C. Eckart y G. Young, The approximation of one matrix by another of lower rank, *Psychometrika* **1** (1936), 211–218.

[11] S. Evans y T. Speed, Invariants of some probability models used in phylogenetic inference, *Ann. Statist.* **21** (1993), 355–377.

[12] J. Fernández-Sánchez y M. Casanellas, Invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages, *Syst. Biol.* **65** (2016), 280–291.

[13] L. García y J. Porter, Small phylogenetic trees webpage, `http://www.shsu.edu/~ldg005/small-trees/`.

[14] M. D. Hendy y D. Penny, A framework for the quantitative study of evolutionary trees, *Syst. Zool.* **38** (1989), 297–309.

[15] T. Jukes y C. Cantor, Evolution of protein molecules, *Mammalian Protein Metabolism* (N. H. Munro, ed.), 21–132, Academic Press, 1969.

[16] A. Kedzierska, M. Drton, R. Guigó y M. Casanellas, SPIn: model selection for phylogenetic mixtures via linear invariants, *Mol. Biol. Evol.* **29** (2012), 929–937.

[17] D. Kosta y K. Kubjas, Geometry of symectric group-based models, `https://arxiv.org/abs/1705.09228`.

[18] M. Kimura, A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences, *J. Mol. Evol.* **16** (1980), 111–120.

[19] M. Kimura, Estimation of evolutionary sequences between homologous nucleotide sequences, *Proc. Natl. Acad. Sci. U.S.A.* **78** (1981), 454–458.

[20] L. Pachter y B. Sturmfels (Eds.), *Algebraic Statistics for Computational Biology*, Cambridge University Press, 2005.

[21] K. Pearson, Contributions to the mathematical theory of evolution, *Phil. Trans. Roy. Soc. London A* **185** (1894), 71–110.

[22] G. Pistone, E. Riccomagno y H. Wynn, *Algebraic Statistics: Computational Commutative Algebra in Statistics*, Chapman & Hall/CRC, 2000.

[23] V. Ranwez y O. Gascuel, Quartet-based phylogenetic inference: improvements and limits, *J. Mol. Evol.* **18** (2001), 1103–1116.

[24] J. Roca-Lacostena y J. Fernández-Sánchez, Embeddability of Kimura 3ST Markov matrices, *J. Theor. Biol.* **445** (2018), 128–135.

[25] J. M. Smith, *Evolutionary Genetics* (2nd ed.), Oxford University Press, 1998.

[26] M. Steel, Recovering a tree from the leaf colourations it generates under a Markov model, *Appl. Math. Lett.* **7** (1994), 19–24.

[27] B. Sturmfels y S. Sullivant, Toric ideals of phylogenetic invariants, *J. Comput. Biol.* **12** (2005), 204–228.

[28] J. Sumner, P. Jarvis, J. Fernández-Sánchez, B. Kaine, M. Woodhams y B. Holland, Is the general time-reversible model bad for molecular phylogenetics?, *Syst. Biol.* **61** (2012), 1069–1978.

[29] D. L. Swofford, *PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods)*, 2002, software package. Disponible en `http://paup.phylosolutions.com/`.

Marta Casanellas, Departament de Matemàtiques, ETSEIB, Universitat Politècnica de Catalunya, Avinguda Diagonal 647, 08028 Barcelona
Correo electrónico: `marta.casanellas@upc.edu`