ARBOR Ciencia, Pensamiento y Cultura
CLXXXVI 746 noviembre-diciembre (2010) 1023-1033 ISSN: 0210-1963
doi: 10.3989/arbor.2010.746n1251

RECONSTRUCCIÓN FILOGENÉTICA USANDO GEOMETRÍA ALGEBRAICA

PHYLOGENETIC RECONSTRUCTION USING ALGEBRAIC GEOMETRY

Marta Casanellas
Jesús Fernández-Sánchez
Ministerio de Educación y Ciencia, MTM2009-14163-C02-2 y Generalitat de Catalunya, 2009 SGR 1284

ABSTRACT: *A new approach to phylogenetic reconstruction has been emerging in the last years. Given an evolutionary model, the joint probability distribution of the nucleotides for these species satisfy some algebraic constraints called invariants. These invariants have theoretical and practical interest, since they can be used to infer phylogenies. In this paper, we explain how to use these invariants to design algorithms for phylogenetic reconstruction and we show how the application of tools and theoretical results coming from commutative algebra and algebraic geometry can improve the performance and the efficiency of these algorithms.*

KEY WORDS: *molecular evolution; phylogenetic reconstruction; evolutionary model; algebraic variety.*

RESUMEN: Una nueva aproximación a la reconstrucción filogenética basada en la geometría algebraica está ganando fuerza en los últimos años. Fijado un modelo evolutivo para un conjunto de especies, las distribuciones teóricas de los nucleótidos de estas especies satisfacen ciertas relaciones algebraicas que llamamos invariantes. Estos invariantes son de interés teórico y práctico dado que se pueden utilizar para inferir filogenias. En este artículo, explicamos cómo usar los invariantes para implementar algoritmos de reconstrucción filogenética y mostramos cómo el uso de técnicas y resultados teóricos procedentes del álgebra conmutativa y la geometría algebraica puede contribuir en la mejora en la eficacia y la eficiencia de estos algoritmos.

PALABRAS CLAVE: evolución molecular; reconstrucción filogenética; modelo evolutivo; variedad algebraica.

1. Introducción

Hace ya 150 años de la publicación de *On the Origin of Species by Means of Natural Selection,* el trabajo donde Charles Darwin proponía que todas las especies están relacionadas entre ellas a través del "árbol de la vida". Desde entonces, ha habido un creciente interés en biología por conocer este árbol de la vida, pero todavía quedan muchas lagunas por llenar y muchos aspectos por entender. En las últimas décadas, y muy especialmente desde la obtención de amplias bases de datos de genomas, la biología evolutiva ha empezado a interactuar con ciertas áreas de las matemáticas. Aunque J. M. Smith ya apuntaba que el álgebra también sería fundamental para la biología comparativa: "if you can't stand algebra, keep out of evolutionary biology", no ha sido hasta el nuevo milenio que la geometría algebraica ha empezado a tener un papel destacado. En este artículo damos una idea sobre cómo interactúan la biología (particularmente la filogenética) y la geometría algebraica y mostramos nuevos resultados sobre datos simulados que prueban que esta interacción es potencialmente fructífera.

La filogenética actual pretende inferir las relaciones ancestrales (llamadas *filogenias*) entre especies actuales a partir de sus genomas[1]. Las *filogenias* se representan generalmente mediante un árbol (con o sin raíz) donde las hojas se etiquetan con las especies actuales, los nodos internos representan ancestros comunes y las ramas entre los nodos los procesos evolutivos entre las especies representadas por éstos. La longitud de las ramas representa la *distancia evolutiva* entre las especies que aparecen en los extremos de la rama o, dicho de otra forma, la cantidad de mutaciones que se han producido entre las dos especies (véase Figura 1).

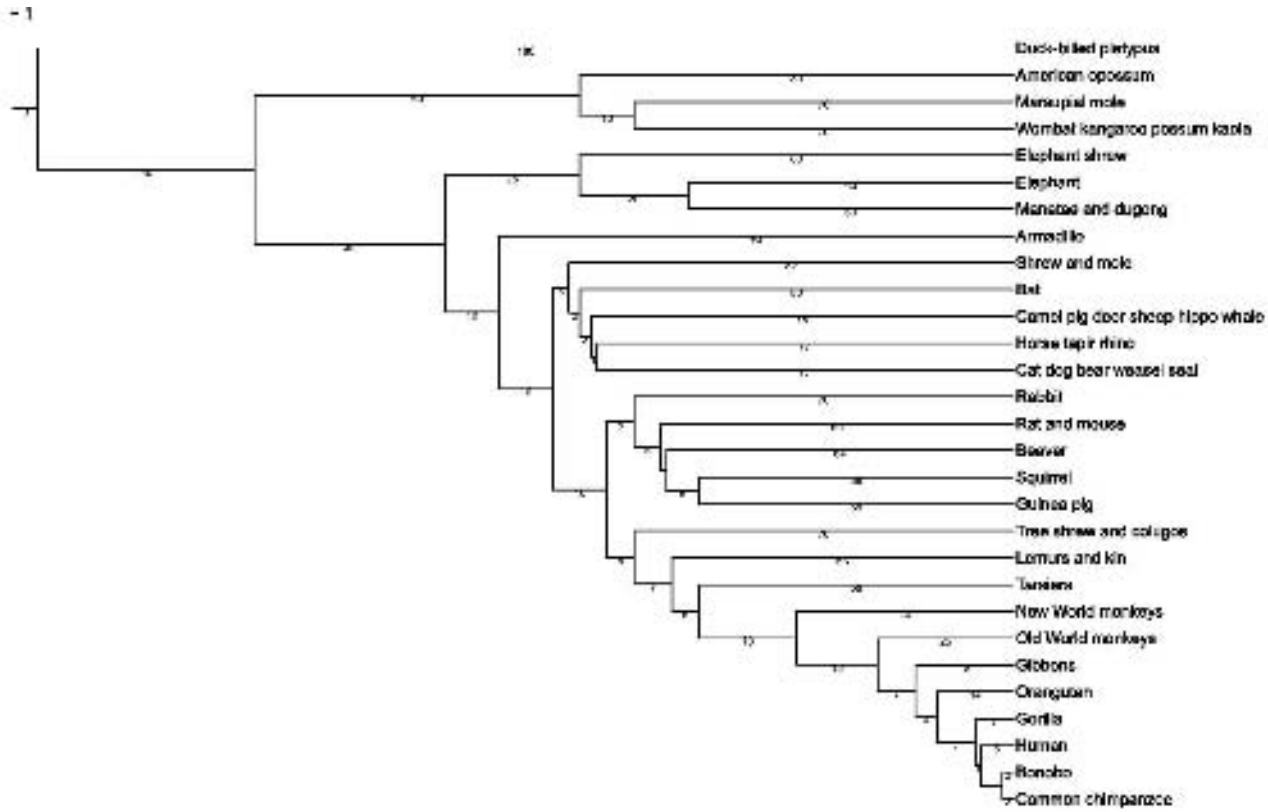


Figura 1. Árbol filogenético de diversas especies de mamíferos

Existen diversas aproximaciones para reconstruir árboles filogenéticos. Algunos métodos se basan en modelos evolutivos preestablecidos, mientras que otros definen distancias entre especies que luego se utilizan para establecer el árbol filogenético. Dentro de este grupo, uno de los métodos más usados es el *Neighbor-Joining* (véase [SN87]). Dentro del primer grupo, el algoritmo de Felsenstein para obtener el árbol que maximiza la verosimilitud de un modelo evolutivo probabilístico es uno de los métodos más conocidos (véase [Fel81]). Algunos modelos evolutivos permiten además la utilización de resultados y técnicas procedentes de la geometría algebraica, así que nosotros nos centraremos en estos modelos. Aunque existen métodos de reconstrucción filogenética bastante fiables (hay estudios hechos con datos simulados) bajo ciertas hipótesis evolutivas, muchos de ellos no son fiables en modelos con un gran número de parámetros. Por ejemplo, se ha demostrado (véase [GG98, YY99]) que las velocidades de mutación de las especies pueden diferir en las distintas ramas del árbol (en este caso, hablamos de un árbol *no homogéneo*). Cualquier modelo que contemple esta posibilidad involucra un gran número de parámetros con la dificultad computacional que esto lleva consigo, así que los modelos usados normalmente en filogenética involucran sólo árboles homogéneos. Es en este punto dónde puede resultar crucial la contribución de la geometría algebraica. Los métodos basados en geometría algebraica no resultan penalizados por el número de parámetros y, como ya mostramos en [CFS07], incluso pueden dar lugar a mejores resultados sobre árboles no homogéneos. En este artículo explicamos nuevos resultados teóricos que mejoran los métodos de reconstrucción filogenética basados en geometría algebraica (véase [CFS08], [CF10]), y probamos mediante simulaciones que estas mejoras teóricas conllevan mejoras en la práctica.

ARBOR CLXXXVI 746 noviembre-diciembre [2010] 1023-1033 ISSN: 0210-1963 doi: 10.3989/arbor.2010.746n1251

Detallamos ahora la estructura del artículo. En la siguiente sección presentamos algunos de los modelos evolutivos algebraicos más utilizados y discutimos por qué razón la utilización de la geometría algebraica para su estudio puede ser de gran utilidad en problemas complejos de filogenética. En la sección 3 damos una idea de los resultados teóricos de geometría algebraica que permiten mejorar los métodos de reconstrucción filogenética basados en geometría algebraica. En la última sección mostramos nuevos resultados sobre el uso de métodos algebraicos en datos simulados.

2. Modelos de evolución y variedades algebraicas

Para representar el proceso de evolución entre especies vamos a dar un modelo estadístico bajo las siguientes hipótesis:

(i) Los árboles son *binarios,* es decir, de la raíz del árbol salen dos ramas y cada rama se divide en dos más hasta llegar a las hojas. Normalmente no es posible inferir a partir de los datos la posición de la raíz del árbol, por lo que los métodos filogenéticos reconstruyen árboles sin raíz. Sin embargo, en esta sección consideraremos árboles con raíz, pues son más intuitivos y pueden facilitar la comprensión de lector.

(ii) La evolución de la especie asociada a un nodo del árbol sólo depende de la especie representada en el nodo inmediatamente superior.

(iii) Las mutaciones ocurren aleatoriamente y la probabilidad de que se produzca una mutación es siempre positiva.

(iv) Las distintas posiciones de la cadena de ADN evolucionan de forma independiente y bajo las mismas probabilidades de mutación.

Como es habitual en filogenética, supondremos que partimos de un *alineamiento* de las secuencias de ADN de las especies. Debido a diversos procesos de mutación, supresión o inserción de nucleótidos, las secuencias de ADN de un mismo gen procedente de varias especies no son idénticas en general, sino que presentan zonas parecidas y zonas que directamente no se pueden comparar. Incluso, las zonas parecidas pueden no encontrarse en el mismo lugar del genoma (por ejemplo, los genomas de distintas especies tienen distinto número de nucleótidos, de cromosomas y de genes). Por este motivo, antes de estudiar las relaciones ancestrales entre dos especies es importante conocer qué partes del genoma se corresponden. Esta información se recoge en un "buen alineamiento" de las secuencias de ADN a considerar: cada columna representa los nucleótidos en las secuencias que han evolucionado a partir de un mismo nucleótido en el ancestro común.

Ejemplo: Supongamos que partimos del siguiente alineamiento de secuencias

Homo sapiens	AACTTCGAGGCTTACCGCTG
Gorilla gorilla	AACGTCTATGCTCACCGATG
Pan troglodytes	AAGGTCGATGCTCACCGATG

La probabilidad de obtener este alineamiento varía en función del árbol filogenético que relaciona las especies (ver figura 2) y entendemos que el árbol correcto es aquél que maximiza esta probabilidad. Nuestro objetivo sería pues conocer el árbol correcto para el alineamiento dado.

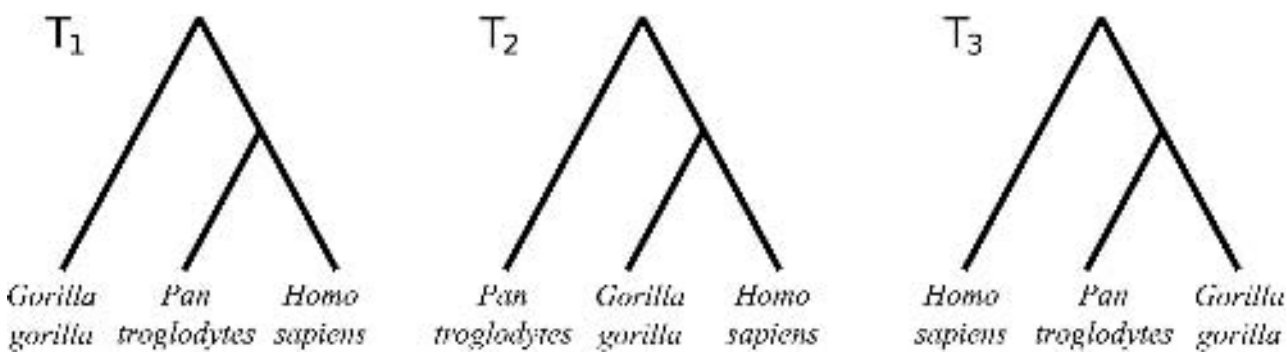


Figura 2. Los tres posibles árboles de tres hojas con raíz

Puesto que suponemos que todas las posiciones del genoma evolucionan bajo las mismas probabilidades, es suficiente modelar una posición. Representaremos el modelo en el árbol de la forma que aparece en la figura 3.

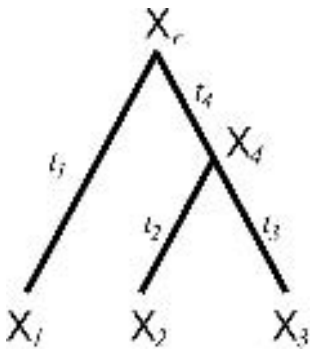


Figura 3. Modelo estadístico en el árbol T_1 de tres hojas con raíz

A cada vértice *i* le asignamos una variable aleatoria X_i discreta que toma valores en el conjunto de los 4 nucleótidos, que representamos mediante {A, C, G, T}. Podemos entender que cada columna del alineamiento es una cierta observación del vector aleatorio $X = (X_1, X_2, X_3)$, por lo que a las variables aleatorias X_i en las hojas del árbol las llamaremos variables "observadas". Por su parte, las variables aleatorias en los nodos interiores son "ocultas" puesto que no disponemos de observaciones de ellas.

Siguiendo un proceso de Markov, asociamos una matriz S_e a cada rama *e*. Las entradas de S_e son las probabilidades $P(x|y,t_e)$ de que un nucleótido *y* en el nodo padre sea sustituido por un nucleótido *x* en el nodo hijo a lo largo del proceso evolutivo representado por la rama *e*. La longitud t_e de la rama representa la distancia evolutiva entre especies, luego a mayor t_e mayor serán las probabilidades de que se produzca alguna mutación.

$$S_e = \begin{array}{c} \\ \mathrm{A} \\ \mathrm{C} \\ \mathrm{G} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \end{array} \\ \begin{pmatrix} P(\mathrm{A}\mid\mathrm{A},t_e) & P(\mathrm{C}\mid\mathrm{A},t_e) & P(\mathrm{G}\mid\mathrm{A},t_e) & P(\mathrm{T}\mid\mathrm{A},t_e) \\ P(\mathrm{A}\mid\mathrm{C},t_e) & P(\mathrm{C}\mid\mathrm{C},t_e) & P(\mathrm{G}\mid\mathrm{C},t_e) & P(\mathrm{T}\mid\mathrm{C},t_e) \\ P(\mathrm{A}\mid\mathrm{G},t_e) & P(\mathrm{C}\mid\mathrm{G},t_e) & P(\mathrm{G}\mid\mathrm{G},t_e) & P(\mathrm{T}\mid\mathrm{G},t_e) \\ P(\mathrm{A}\mid\mathrm{T},t_e) & P(\mathrm{C}\mid\mathrm{T},t_e) & P(\mathrm{G}\mid\mathrm{T},t_e) & P(\mathrm{T}\mid\mathrm{T},t_e) \end{pmatrix} \end{array}$$

Estas probabilidades son desconocidas para nosotros y, junto con la distribución π_A, π_C, π_G, π_T de nucleótidos en la raíz, son los *parámetros* del modelo. Las matrices S_e se llaman *matrices de sustitución* o *de transición.* Según la estructura que presenten estas matrices obtenemos diferentes modelos (algebraicos) de evolución. Por ejemplo, si no imponemos ninguna restricción en las entradas de S_e, resulta el modelo más general posible, llamado el *modelo general de Markov* (cf. [Ste94, BH87]):

$$S_e = \begin{pmatrix} a_e & b_e & c_e & d_e \\ e_e & f_e & g_e & h_e \\ j_e & k_e & l_e & m_e \\ n_e & o_e & p_e & q_e \end{pmatrix}$$

Si imponemos que $\pi_A = \pi_T$, $\pi_C = \pi_G$ y que las matrices de sustitución tengan la estructura

$$S_e = \begin{pmatrix} a_e & b_e & c_e & d_e \\ e_e & f_e & g_e & h_e \\ h_e & g_e & f_e & e_e \\ d_e & c_e & b_e & a_e \end{pmatrix}$$

obtenemos el modelo conocido como *Strand symmetric* (cf. [CS05]). Si imponemos $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$ y la matrices de sustitución tienen la forma

$$S_e = \begin{pmatrix} a_e & b_e & c_e & d_e \\ b_e & a_e & d_e & c_e \\ c_e & d_e & a_e & b_e \\ d_e & c_e & b_e & a_e \end{pmatrix}$$

entonces obtenemos la versión algebraica del modelo de *Kimura 3-parámetros* (cf. [Kim81]). Si además imponemos $b_e = d_e$, tenemos el modelo algebraico del *Kimura 2-parámetros* (cf. [Kim80]) y si nos restringimos a $b_e = c_e = d_e$, obtenemos la versión algebraica del modelo de *Jukes-Cantor* (cf. [JC69]). Todos estos modelos son ejemplos de los llamados modelos *equivariantes* (ver [DK09, CF10]).

La hipótesis (iv) implica que la probabilidad de que la evolución en uno de los árboles de la figura 2 haya dado lugar al alineamiento dado arriba para las secuencias de *Homo sapiens, Gorilla gorilla* y *Pan troglodytes* es igual a

$$(p^T_{AAA})^4 * p^T_{CCG} * p^T_{TGG} * (p^T_{TTT})^3 * (p^T_{CCC})^4 * \\ * p^T_{GTG} * p^T_{GTT} * (p^T_{GGG})^3 * p^T_{TCC} * p^T_{CAA}$$

ARBOR CLXXXVI 746 noviembre-diciembre [2010] 1023-1033 ISSN: 0210-1963 doi: 10.3989/arbor.2010.746n1251

donde p^T_{xyz} denota la probabilidad de observar *x, y, z* en las hojas *Gorilla gorilla* (X_1), *Pan troglodytes* (X_2) y *Homo sapiens* (X_3) *del árbol T* respectivamente:

$$p^T_{xyz} = \mathrm{Prob}(X_1 = x, X_2 = y, X_3 = z \mid T).$$

Si denotamos $\pi = (\pi_A, \pi_C, \pi_G, \pi_T)$ la distribución de nucleótidos en la raíz, considerando el proceso de Markov (hipótesis (ii)) en el árbol de la figura 3, la probabilidad de observar los nucleótidos *x, y, z* en las hojas se expresa en función de las entradas de las matrices de sustitución de la siguiente forma:

$$p^T_{xyz} = \sum_{x_r, x_4 \in \{A,C,G,T\}} \pi_{x_r} S_1(x_1, x_r) S_4(x_4, x_r) S_2(x_2, x_4) S_3(x_3, x_4). \quad (1)$$

Bajo los modelos evolutivos que hemos descrito, la distribución conjunta en las hojas se expresa como función polinómica en los parámetros [véase ecuación (1)]. Por esta razón, estos modelos se llaman modelos *algebraicos.* Todo modelo evolutivo equivariante $\mathcal{M}$ con *d* parámetros libres sobre un árbol *T* de *n* hojas, tiene asociado la siguiente aplicación polinomial:

$$\begin{array}{rcl} \varphi^{\mathcal{M}}_T : \mathbb{R}^d & \to & \mathbb{R}^{4^n} \\ \theta = (\theta_1, \ldots, \theta_d) & \mapsto & (p^T_{AA\ldots A}, p^T_{AA\ldots C}, p^T_{AA\ldots G}, \ldots, p^T_{TT\ldots T}) \end{array} \quad (2)$$

De esta forma el modelo de Jukes-Cantor en un árbol de tres hojas (con raíz) tiene asociada la siguiente aplicación polinomial:

$$\begin{array}{rcl} \varphi^{JC}_T : \mathbb{R}^4 & \to & \mathbb{R}^{64} \\ (a_1, a_2, a_3, a_4) & \mapsto & (p^T_{AAA}, p^T_{AAC}, p^T_{AAG}, \ldots, p^T_{TTT}) \end{array}$$

Aunque los parámetros del modelo son probabilidades y por lo tanto están en el intervalo [0,1], nos interesa olvidar esta restricción y considerar estas aplicaciones polinomiales definidas sobre $\mathbb{R}^d$ o incluso, sobre $\mathbb{C}^d$, con el objetivo de poder aplicar resultados y técnicas de la geometría algebraica en este contexto.

Al considerar distintos valores para los parámetros del modelo obtenemos distintos puntos en el espacio de llegada. Todos estos puntos están sobre una variedad algebraica que denotamos por $V_{\mathcal{M}}(T)$. Una *variedad algebraica* es un conjunto de puntos que son solución de un sistema de ecuaciones polinómicas: $V_{\mathcal{M}}(T) = \{p \in \mathbb{R}^{4^n} \mid f_1(p) = 0, \ldots, f_r(p) = 0\}$ para ciertos polinomios $f_1, \ldots, f_r$; decimos entonces que $f_1, \ldots, f_r$ *definen* la variedad $V_{\mathcal{M}}(T)$. La imagen de una aplicación polinomial no es en general una variedad algebraica, pero siempre podemos considerar la menor variedad algebraica que contiene la imagen. De hecho, la imagen de φ_T comprende "casi todos" los puntos de la variedad $V_{\mathcal{M}}(T)$; el resto de puntos pertenecen a una variedad algebraica de dimensión menor; en tal caso decimos que la imagen de $\varphi^{\mathcal{M}}_T$ es un *abierto denso* de $V_{\mathcal{M}}(T)$. Si un polinomio *f* se anula sobre todo punto de la imagen de $\varphi^{\mathcal{M}}_T$ (o equivalentemente sobre la variedad $V_{\mathcal{M}}(T)$) entonces *f* es una relación que cumplen las probabilidades teóricas $p^T_{x_1,\ldots,x_n}$, independientemente de los parámetros del modelo de las que proceden. En cierta forma *f* es "invariante". Los biólogos Cavender, Felsenstein y Lake en los años ochenta propusieron la siguiente definición (cf. [CF87], [Lak87]):

Definición 2.1. Dado un árbol filogenético *T* de *n* hojas y un modelo evolutivo $\mathcal{M}$, los polinomios que se anulan sobre cualquier punto *p* de $V_{\mathcal{M}}(T)$ se llaman *invariantes* de *T.* Los polinomios que se anulan sobre todos los puntos de $V_{\mathcal{M}}(T)$ pero que no se anulan sobre todos los puntos de $V_{\mathcal{M}}(T')$ para algún otro árbol *T'* de *n* hojas se llaman *invariantes filogenéticos* de *T.*

Los invariantes filogenéticos permiten distinguir entre distintos árboles y, por lo tanto, pueden ser usados para inferir la *topología* del árbol filogenético. Si bien la longitud de las ramas del árbol también es de gran interés para los biólogos, en este artículo nos centramos en recuperar la topología del árbol, es decir, en la forma del árbol con los nombres de las especies a considerar en las hojas (en términos matemáticos, hablamos de la topología del grafo del árbol con las hojas etiquetadas). Los invariantes han sido usados para estudiar la adecuación del modelo escogido y también para deducir divisiones ancestrales entre grupos de especies ([San90]). No ha sido hasta muy recientemente que los matemáticos han comenzado a interesarse por esta aplicación de la geometría algebraica en la filogenética y a estudiar la geometría de estas variedades.

Ejemplo 2.2. Para cualquier árbol *T* de 3 hojas (véase figura 2) bajo el modelo de Jukes-Cantor las siguientes igualdades (que se pueden deducir fácilmente de la simetría de las matrices de sustitución) dan lugar a invariantes de *T:*

$$p_{AAA} = p_{CCC} = p_{GGG} = p_{TTT} \quad \text{4 términos}$$
$$p_{AAC} = p_{AAG} = p_{AAT} = \ldots = p_{TTG} \quad \text{12 términos}$$
$$p_{ACA} = p_{AGA} = p_{ATA} = \ldots = p_{TGT} \quad \text{12 términos}$$
$$p_{CAA} = p_{GAA} = p_{TAA} = \ldots = p_{GTT} \quad \text{12 términos}$$
$$p_{ACG} = p_{ACT} = p_{AGT} = \ldots = p_{CGT} \quad \text{24 términos}$$

Evidentemente también se cumple $\Sigma_{x_1, x_2, x_3}\ p_{x_1\ x_2\ x_3} - 1 = 0$. Estos 60 invariantes algebraicos se anulan sobre la variedad asociada a cualquier árbol de 3 hojas (véase figura 2) bajo el modelo de Jukes-Cantor, luego no son invariantes filogenéticos. Sin embargo, para cualquiera de los tres árboles T existen polinomios de grado 3 que se anulan en $V_{\mathcal{M}}(T)$ pero no se anulan sobre las variedades de los otros dos árboles. Se trata, por tanto, de invariantes filogenéticos.

Dado un alineamiento de n especies, denotamos por $\rho_{x_1 \ldots x_n}$ la frecuencia relativa de aparición de la n-upla x_1, ..., x_n como columna del alineamiento. Por ejemplo, en el caso del alineamiento dado en la página 1025 tenemos:

$$\rho_{AAA} = 4/20 \quad \rho_{CAA} = 1/20 \quad \rho_{CCC} = 4/20 \quad \rho_{CCG} = 1/20 \quad \rho_{GGG} = 3/20$$
$$\rho_{GTG} = 1/20 \quad \rho_{GTT} = 1/20 \quad \rho_{TCC} = 1/20 \quad \rho_{TGG} = 1/20 \quad \rho_{TTT} = 3/20$$

y 0 para cualquier otra combinación de nucleótidos. Si el alineamiento se hubiera generado siguiendo uno de los árboles de la figura 2 bajo uno de los modelos descritos, los invariantes filogenéticos se anularían sobre el vector de frecuencias relativas $\rho = (\rho_{AAA}, \ldots, \rho_{TTT})$, o equivalentemente, ρ sería un punto de la variedad algebraica asociada.

En la práctica, es claro que el genoma de la especies no evoluciona siguiendo ningún modelo evolutivo sobre un árbol filogenético. De todas formas, si el modelo evolutivo es apropiado al alineamiento dado, es de esperar que al evaluar los invariantes filogenéticos del árbol "correcto" sobre las frecuencias relativas obtendremos valores cercanos a 0. Se hace necesario por tanto proponer métodos para la inferencia del árbol correcto en esta situación. Algoritmos basados en geometría algebraica para reconstrucción filogenética han sido descritos en [Eri05] y [CGS05].

3. Resultados teóricos

Para árboles de 4 hojas o más no es factible usar programas de álgebra computacional para encontrar todos los invariantes asociados a un árbol (incluso para modelos sencillos como Jukes-Cantor). Necesitamos resultados teóricos que proporcionen algoritmos para obtener los invariantes filogenéticos necesarios. En esta sección explicamos los resultados más relevantes que hemos obtenido en este sentido y aquéllos que serán usados en la siguiente sección.

A partir de ahora, todos los árboles que aparecen son árboles *trivalentes* y sin raíz, esto es, en cada vértice interior del árbol concurren exactamente tres ramas. Para $n \geq 3$, denotaremos por $\mathcal{T}_n$ el conjunto de las topologías posibles de árboles de n hojas, trivalentes, etiquetados y sin raíz. Por ejemplo, si $n = 4$, la figura 4 muestra las tres topologías posibles en $\mathcal{T}_4$ con hojas etiquetadas por $\{1, 2, 3, 4\}$.

Teorema 3.1 ([DK09]). *Sea T un árbol filogenético de n especies evolucionando bajo un modelo $\mathcal{M}$ de los mencionados más arriba. Entonces existe un algoritmo que permite obtener una lista completa de los invariantes asociados a T a partir de los invariantes de un árbol de 3 hojas bajo el modelo $\mathcal{M}$ y de ciertas restricciones para el rango de matrices asociadas a las ramas de T. Estas restricciones se traducen en ecuaciones algebraicas que llamaremos "invariantes de ramas".*

Este resultado ha sido probado en casos particulares por E. Allman y J. Rhodes [AR07], B. Sturmfels y S. Sullivant [SS05], M. Casanellas y S. Sullivant [CS05] pero han sido J. Draisma y J. Kuttler quienes han demostrado una versión general del resultado para todos los modelos equivariantes.

Es necesario observar que, mientras los invariantes de ramas son relativamente fáciles de obtener, no pasa lo mismo con los invariantes asociados a árboles de 3 hojas. De hecho, no se conocen los invariantes asociados a árboles de 3 hojas bajo el modelo general de Markov o el *strand symmetric model.* Por lo tanto, el resultado anterior aún no se puede llevar a la práctica en estos casos.

De todas maneras, es lógico esperar que, en general, no sea necesario calcular una lista completa de invariantes sino que sea suficiente aplicar ciertos invariantes bien seleccionados, al menos para ciertas aplicaciones en filogenética. Esta idea se ve plasmada, por ejemplo, en el resultado principal del artículo [CFS08], que presentamos a continuación.

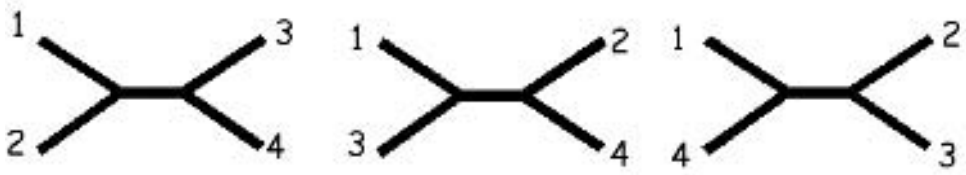


Figura 4. Las tres topologías posibles de árboles trivalentes de 4 hojas

Antes de enunciar el resultado, recordemos que la codimensión de una variedad algebraica es la dimensión del espacio ambiente menos la dimensión de la variedad. Es importante notar que para definir cualquier variedad algebraica, se necesitan como mínimo tantos polinomios como la codimensión de la variedad. En el caso de una variedad algebraica asociada a un árbol de *n* hojas que evoluciona bajo el modelo de Kimura con 3 parámetros *(K3)*, esta codimensión viene dada por

$$\mathrm{codim} V_{K3}(T) = 4^n - 3(2n-3).$$

Aunque el siguiente resultado hace referencia exclusivamente al modelo de Kimura con 3 parámetros, es de esperar que resultados análogos sean ciertos también para otros modelos algebraicos. En la siguiente sección probaremos que en datos simulados sobre árboles de 4 hojas es efectivamente suficiente considerar estos invariantes.

Teorema 3.2 ([CFS08]). *Sea T un árbol filogenético de n especies evolucionando bajo el modelo Kimura 3-parámetros. Los puntos de V(T) que tienen significado biológico*[2] *están contenidos en un abierto denso donde la variedad puede ser definida por* $4^n - 3(2n - 3)$ *invariantes.*

Además presentamos un algoritmo que permite obtener una lista completa de invariantes con esta propiedad.

Ejemplo 3.3. Sea *T* un árbol de 4 hojas sin raíz evolucionando bajo el modelo Kimura 3-parámetros. Mediante programas de álgebra computacional se puede comprobar que una lista completa de invariantes de la variedad $V(T)$ está formada por 8002 polinomios (véase Small Trees webpage). Sin embargo, usando este resultado podemos afirmar que para aplicaciones a la biología es suficiente considerar 48 invariantes (véase [CFS08]).

Los dos resultados anteriores se centran en buscar listas de invariantes, sin preocuparse si estos invariantes son *filogenéticos* o no. Sin embargo, para resolver problemas de reconstrucción filogenética sólo los invariantes filogenéticos tienen interés. En un estudio reciente nos hemos centrado en encontrar los invariantes filogenéticos de $V_{\mathcal{M}}(T)$ para cualquier modelo $\mathcal{M}$ y cualquier topología de árbol sobre *n* hojas. Para ello, nos hemos inspirado en el siguiente resultado combinatórico.

Teorema 3.4 (Buneman, [Bun71]). *Sea T el grafo de un árbol binario. Entonces T se puede reconstruir conociendo sólo las biparticiones de las hojas de T que inducen sus ramas.*

En vista de este resultado, es natural esperar que, para aplicaciones en reconstrucción filogenética, sólo sea necesario considerar los invariantes de rama mencionados más arriba. Esto es lo que probamos en [CF10].

Teorema 3.5. *Sea T un árbol filogenético de n especies evolucionando bajo un modelo equivariante* $\mathcal{M}$ *de los considerados en este artículo. Para aplicaciones en reconstrucción filogenética es suficiente considerar los invariantes de rama del árbol T. Más concretamente:*

Sea $\mathcal{T}_n$ *el conjunto de topologías de árboles binarios de n hojas. Existen abiertos densos* $U_T \subset V_{\mathcal{M}}(T)$ *para cada* $T \in \mathcal{T}_n$ *tales que, si p es un punto en* $U_{T\in\mathcal{T}}\, U_T$*, entonces p pertenece a* $V_{\mathcal{M}}(T_0)$ *si y sólo si p es anulado por todos los invariantes de ramas de* T_0*.*

Este resultado nos dice que es suficiente utilizar los invariantes de rama para construir el árbol correcto. En particular, no es necesario conocer todos los invariantes de un árbol de 3 hojas mencionado anteriormente.

Ejemplo 3.6. Continuando con el Ejemplo 3.3, sea T un árbol de 4 hojas sin raíz evolucionando bajo el modelo Kimura 3-parámetros. Entonces de los 48 invariantes del Ejemplo 3.3 sólo 36 son invariantes filogenéticos. En la siguiente sección usaremos estos invariantes en simulaciones y mostraremos cómo en efecto es suficiente considerar los 36 invariantes filogenéticos.

4. Resultados en simulaciones

En esta sección presentamos la aplicación de los teoremas anteriores en un estudio con simulaciones en los árboles de 4 hojas sin raíz sobre el modelo de Kimura 3-parámetros. Para ello usaremos el algoritmo presentado en [CGS05] que se resume como sigue. Sean T_1, T_2, T_3 las tres topologías posibles de árboles de 4 hojas sin raíz con hojas etiquetadas por $\{1, 2, 3, 4\}$ (véase figura 4).

Algoritmo de reconstrucción filogenética

Input: Un alineamiento de las especies 1, 2, 3, 4 y una colección de invariantes I_i para cada árbol T_i, $i \in \{1, 2, 3\}$.

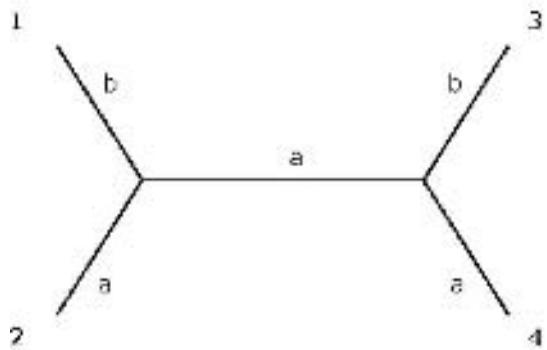
Procedimiento: sea p el punto cuyas coordenadas son las frecuencias relativas de las columnas del alineamiento dado. Para cada topología de árbol T_i calculamos el valor $s(T_i) := \Sigma_{f \in I_i} |f(p)|$.

Output: El árbol T_i con menor $s(T_i)$.

Para aplicar los resultados teóricos de la sección anterior usaremos este algoritmo en tres versiones distintas (siempre considerando el modelo Kimura 3-parámetros):

A1. I_i es una colección completa de invariantes de $V(T_i)$ (véase Teorema 3.1). En nuestro caso ésta está formada por los 8002 polinomios mencionados en el Ejemplo 3.3.
A2. I_i son los 48 polinomios mencionados en el Ejemplo 3.3 para cada árbol T_i (véase Teorema 3.2).
A3. I_i son los 36 invariantes de ramas mencionados en el Ejemplo 3.6 para cada árbol T_i (véase Teorema 3.5).

Para poner a prueba estar 3 versiones del algoritmo, usamos la idea de J. Huelsenbeck [Hue95] que establece un espacio de árboles donde examinar la eficacia de distintos algoritmos. Lo describimos a continuación. Fijamos la topología T_1 y denotamos por a la longitud de las dos ramas inferiores y la rama interior y por b la longitud de las dos ramas superiores (véase figura 5). Al variar a y b entre 0.01 y 0.75 obtenemos el espacio de árboles representado a la derecha de la figura 5. Luego, para cada par (a, b) y para cada longitud prefijada l y



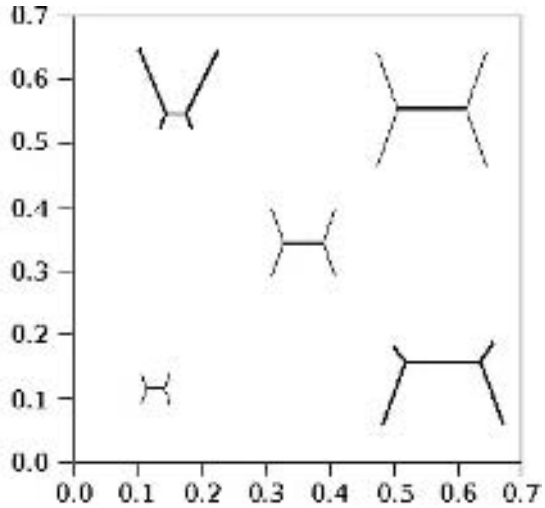


Figura 5. El parámetro a asigna la longitud de la rama interior y de dos ramas exteriores opuestas, mientras que la rama b asigna la longitud de las ramas restantes. Los dos parámetros a y b varían desde 0,01 hasta 0,75 en incrementos de 0.02

ARBOR CLXXXVI 746 noviembre-diciembre [2010] 1023-1033 ISSN: 0210-1963 doi: 10.3989/arbor.2010.746n1251

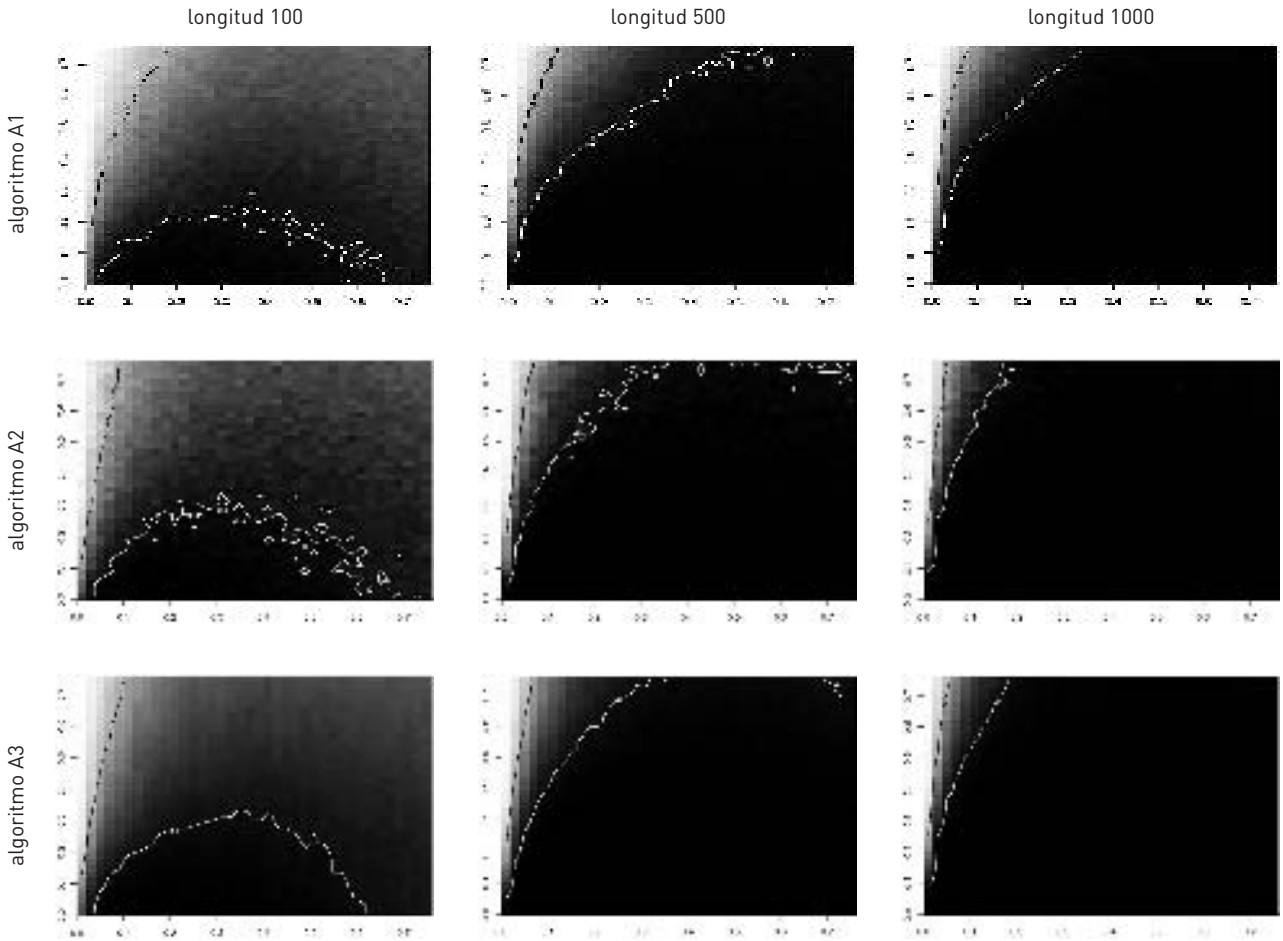


Figura 6. Resultados obtenidos en la simulación de las 3 versiones del algoritmo en datos simulados con Seq-Gen usando los parámetros $\gamma = 0{,}12$, $\alpha = 0{,}73$, $\beta = 0{,}13$. En blanco representamos las curvas de nivel del 95% y en negro las curvas de nivel del 33%

con la ayuda del programa Seq-Gen [RG97], generamos 1000 alineamientos de secuencias de longitud *l* procedentes del árbol T_1 con estas longitudes de rama. Mediante la escala de grises, representamos la proporción de aciertos obtenida por el algoritmo a la hora de escoger el árbol T_1 como el árbol correcto, de forma que tonalidades oscuras representan mayor éxito que tonalidades más claras. De esta forma, el color negro en un punto (*a, b*) indica que el algoritmo ha elegido el árbol T_1 el 100% de las veces, mientras que el color blanco indica que no lo ha elegido nunca.

Los resultados para secuencias de longitudes 100, 500 y 1000 para las 3 versiones distintas del algoritmo se muestran en la figura 6.

Podemos ver claramente que el Teorema 3.2 no sólo sirve para dar un algoritmo mucho más rápido puesto que reduce el número de invariantes sino que también nos proporciona un algoritmo mucho más eficiente. Por otra parte vemos que considerar sólo invariantes de rama como proponemos en el Teorema 3.5 proporciona un algoritmo (A3) igual de eficiente que el algoritmo A2 pero más rápido (dado que se usan menos invariantes).

Aunque estas simulaciones muestran claramente cómo los resultados teóricos de la sección 3 mejoran la eficiencia y eficacia del algoritmo, es necesario comparar este algoritmo con los métodos usados por biólogos. Uno de ellos es el método de Neighbor-joining [SN87] basado en

un algoritmo voraz de construcción iterativa de cerezas en el árbol. Éste es un algoritmo muy eficaz y en árboles homogéneos nuestro algoritmo A3 no supera de ningún modo al Neighbor-joining. Sin embargo, presentamos aquí un estudio con datos simulados que muestra que el algoritmo A3 es potencialmente mejor que Neighbor-joining en árboles no homogéneos. Usamos el árbol no homogéneo de 4 hojas descrito en la figura 3 de [CFS07] y producimos para cada longitud $l \in \{100, 200, ..., 1000\}$ un alineamiento de secuencias evolucionando bajo este árbol. En la gráfica de la figura 7 vemos la eficacia del algoritmo A3 (línea continua) y de Neighbor-joining (línea discontinua).

Aunque todavía hay que mejorar el algoritmo presentado aquí (o incluso sería mejor dar un algoritmo totalmente nuevo basado en invariantes filogenéticos) y hacer simulaciones sobre árboles con mayor número de hojas, los resultados mostrados en este artículo sugieren cuáles son los invariantes que realmente hay que usar y en qué casos pueden ser más eficaces los algoritmos basados en invariantes que los métodos de reconstrucción filogenética comúnmente usados.

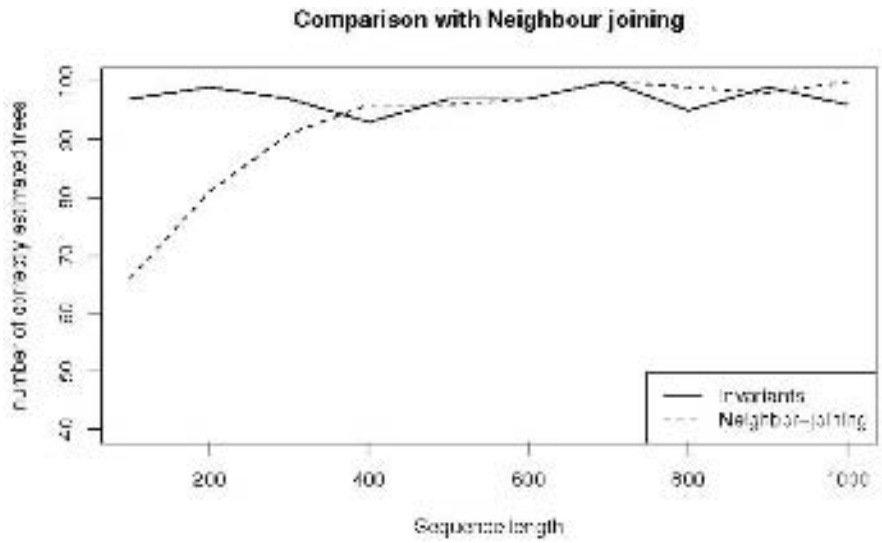


Figura 7

NOTAS

1 El genoma de una especie eucariota es el contenido de ADN que hay en el núcleo de sus células. Las unidades básicas que componen el ADN son los *nucleótidos.* Hay cuatro distintos en el ADN: A denota *adenina*, C *citosina*, G *guanina* y T *timina.* En nuestro contexto, dar el genoma de una especie equivale a dar una secuencia ordenada de carácteres en las letras A, C, G, T.

2 Para una definición precisa de punto con significado biológico, véase [CFS08].

REFERENCIAS

[AR07] E.S. Allman y J. A. Rhodes. "Phylogenetic ideals and va-

Recibido: 27 de diciembre de 2009
Aceptado: 25 de enero de 2010

rieties for the general Markov model", *Adv. in Appl. Math.,* 40: 127-148, 2007.

[BH87] D. Barry y J. A. Hartigan. "Asynchronous distance between homologous DNA sequences", *Biometrics,* 43(2): 261-276, 1987.

[Bun71] P. Buneman. "The recovery of trees from measures of dissimilarity", in Edinburgh University Press, editor, *Mathematics in the Archaeological and Historical Sciences,* pp. 387-395, 1971.

[CF87] J. Cavender y J. Felsenstein. "Invariants of phylogenies in a simple case with discrete states", *J. Classification,* 4: 57-71, 1987.

[CFS07] M. Casanellas y J. Fernández-Sánchez. "Performance of a new invariants method on homogeneous and nonhomogeneous quartet trees", *Mol. Biol. Evol.,* 24(1): 288-293, 2007.

[CFS08] M. Casanellas y J. Fernández-Sánchez. "Geometry of the Kimura 3-parameter model", *Adv. in Appl. Math,* 41: 265-292, 2008.

[CF10] M. Casanellas y J. Fernández-Sánchez. *Relevant phylogenetic invariants of evolutionary models,* 2009. Aparecerá en *J. Mathèmatiques Pures et Appliquées.*

[CGS05] M. Casanellas, L. D. García y S. Sullivant. "Catalog of small trees", in L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for computational biology,* chapter 15, Cambridge University Press, 2005.

[CS05] Casanellas y S. Sullivant. "The strand symmetric model", in L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for computational biology,* chapter 16, Cambridge University Press, 2005.

[DK09] J. Draisma y J. Kuttler. "On the ideals of equivariants tree models", *Mathematische Annalen,* 344: 619-644, 2009.

[Eri05] N. Eriksson. "Tree construction using singular value decomposition", in L. Pachter and B. Sturmfels, editors, *Algebraic Statistics for computational biology,* chapter 19, Cambridge University Press, pp. 347-358, 2005.

[Fel81] J. Felsenstein. "Evolutionary trees from DNA sequences: a maximum likelihood approach", *J. Mol. Evol.,* 17: 368-376, 1981.

[GG98] N. Galtier y M. Gouy. "Inferring pattern and process: maximum likelihood implementation of a non-homogeneous model of DNA sequence evolution for phylogenetic analysis", *Mol. Biol. Evol.,* 154(4): 871-879, 1998.

[Hue95] J. P. Huelsenbeck. "Performance of phylogentic methods in simulation", *Syst. Biol.,* 44: 17-48, 1995.

[JC69] T. H. Jukes y C. R. Cantor. "Evolution of protein molecules", in *Mammalian Protein Metabolism,* pp. 21-132, 1969.

[Kim80] M. Kimura. "A simple method for estimating evolutionary rates of base substitution through comparative studies of nucleotide sequences", *J. Mol. Evol.,* 16: 111-120, 1980.

[Kim81] M. Kimura. "Estimation of evolutionary sequences between homologous nucleotide sequences", *Proc. Nat. Acad. Sci.,* USA, 78: 454-458, 1981.

[Lak87] J. A. Lake. "A rate-independent technique for analysis of nucleaic acid sequences: evolutionary parsimony", *Mol. Biol. Evol.,* 4: 167-191, 1987.

[RG97] A. Rambaut y N. C. Grassly. "Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees", *Comput. Appl. Biosci.,* 13: 235-238, 1997.

[San90] D. Sankoff. "Designer invariants for large phylogenies", *Mol. Biol. Evol.,* 7: 255-269, 1990.

[SN87] N. Saitou y M. Nei. "The neighbor joining method: a new method for reconstructing phylogenetic trees", *Mol. Biol. Evol.,* 4(4): 406-425, 1987.

[SS05] B. Sturmfels y S. Sullivant. "Toric ideals of phylogenetic invariants", *J. Comput. Biol.,* 12: 204-228, 2005.

[Ste94] M. A. Steel. "Recovering a tree from the leaf colourations it generates under a markov model", *Applied Mathematics Letters,* 7: 19-24, 1994.

[YY99] Z. Yang y A. D. Yoder. "Estimation of the transition/transversion rate bias and species sampling", *J. Mor. Evol.,* 48: 274-283, 1999.