Publicación cuatrimestral
enero - abril 2021
vol. 4, no. 11

ISSN: 2594-0546

Tequio
Revista de Divulgación, Investigación e Innovación

Filogenia, matemáticas y biología comparada

Obra gráfica de
Almaluz Guzmán

11

UABJO
Universidad Autónoma Benito Juárez de Oaxaca
Oaxaca, México

TEQUIO no. 11, enero-abril 2021, es una publicación cuatrimestral editada y distribuida por la Universidad Autónoma "Benito Juárez" de Oaxaca, con domicilio en Edificio de Rectoría Planta Baja, Secretaría Académica, Dirección de Investigación, Ciudad Universitaria, Avenida Universidad s/n, colonia Cinco Señores, Oaxaca de Juárez, Oaxaca, C.P. 68120, México. Teléfono 951-50-20-700 ext. 20148, correo electrónico: publicacionesuabjo2016@gmail.com, dirección electrónica: http://www.uabjo.mx/publicaciones. Editora responsable: Dra. Gisela Fuentes Mascorro. Reserva al uso exclusivo **No.04-2017-070614533300-203** para la difusión vía red de cómputo, **ISSN: 2594-0546, expedidos por el Instituto Nacional del Derecho de Autor, Certificado de Licitud de Título y Contenido: 16973**, en la **Comisión Calificadora de Publicaciones y Revistas Ilustradas de la Secretaría de Gobernación.** Fecha de actualización: 1 de septiembre de 2020 en la Dirección de Servicios Editoriales, ubicada en el circuito interior de la Ciudad Universitaria.
TEQUIO es un espacio para difundir la investigación, las reflexiones teóricas y el conocimiento científico entre la comunidad universitaria y el público en general. Las opiniones expresadas por los autores no necesariamente reflejan la postura de la Universidad Autónoma "Benito Juárez" de Oaxaca (UABJO).

Contenido

"Iris/radiales"
Cerámica /esmalte
12 x 60 x 60 cm
2014

"Tricomas"
Cerámica/engobes y esmalte
9.5 x 14 x 14 cm
2016

"Semilla voladora"
Cerámica/Raku
20 x 16.5 x 5 cm
2017

Editorial

La filogenética es la parte de la sistemática que estudia las relaciones ancestro-descendientes entre los seres vivos. Al inicio, esta rama se desarrolló sobre una base descriptiva, clasificando las especies según sus rasgos. Sin embargo, en el siglo XX se produjo un cambio de paradigma impulsado por los avances tecnológicos que revelaron el papel fundamental de las proteínas y del ADN en la evolución, lo que condujo a los actuales enfoques cuantitativos y computacionales en filogenética. Hoy en día, la biología y las matemáticas, haciendo uso de herramientas de la bioinformática, pueden utilizar el volumen de datos disponibles para plantear hipótesis robustas sobre la historia de la vida.

Dado que el objetivo principal de la filogenética es inferir las relaciones ancestro-descendientes entre diferentes linajes, es necesario hacer uso de árboles filogenéticos que permitan visualizar ese patrón. También es pertinente emplear métodos estadísticos y algebraicos para obtener una primera hipótesis de las relaciones evolutivas entre las distintas progenies, de acuerdo con los datos disponibles. Este proceso recibe el nombre de inferencia filogenética.

En este número de *Tequio* se describen métodos clásicos de inferencia filogenética, como la máxima parsimonia, la máxima verosimilitud y la inferencia bayesiana, así como procedimientos alternativos, entre ellos la conjugación de Hadamard o la teoría de politopos que están relacionados con el enfoque basado en la distancia. La información que aquí se presenta no pretende ser una revisión exhaustiva, ya que se trata de un campo bastante desarrollado. En su lugar, se quiere introducir el tema, brindando diferentes puntos de vista, explicando conexiones que son de interés tanto para biólogos/as como para matemáticos/as.

En las siguientes páginas veremos cómo la máxima parsimonia (MP) estima una filogenia utilizando el principio de la Navaja de Ockham, según el cual la respuesta más simple o que implica el menor número de pasos evolutivos tiende a ser la mejor. Esta idea es radicalmente diferente de lo que proponen formas más novedosas de reconstrucción filogenética, como la máxima verosimilitud (MV) y la inferencia bayesiana (IB), que tienen como punto de partida los mismos conceptos de verosimilitud o probabilidad de una filogenia. El primero de estos dos conceptos parte de secuencias de ADN o caracteres morfológicos provenientes de varias especies o linajes, bajo un modelo concreto de evolución. La diferencia entre la MV y la IB radica en el principio estadístico sobre el que se sustenta cada una. La MV reporta al árbol más verosímil, mientras que la IB calcula la probabilidad posterior de cada filogenia, aplicando el Teorema de Bayes. En este número se describen los pasos necesarios para inferir una filogenia usando la MP, la MV y la IB.

Otra herramienta que también presentamos es la conjugación de Hadamard. Ésta es una ecuación matricial que relaciona las longitudes de las ramas de un árbol filogenético (propuesto desde el principio) con su distribución de probabilidad de los patrones de sustitución. El enfoque se centra en

cómo construir dicha relación ejemplificándolo sobre un cuartet cuyas hojas están asociadas a cuatro secuencias de caracteres ficticias, partiendo tanto del modelo de evolución molecular de Jukes-Cantor, como de la condición de reloj molecular, y en cómo aplicarla para determinar los valores óptimos de las longitudes de sus ramas a fin de maximizar la probabilidad de que dicho árbol explique las secuencias de nucleótidos observadas.

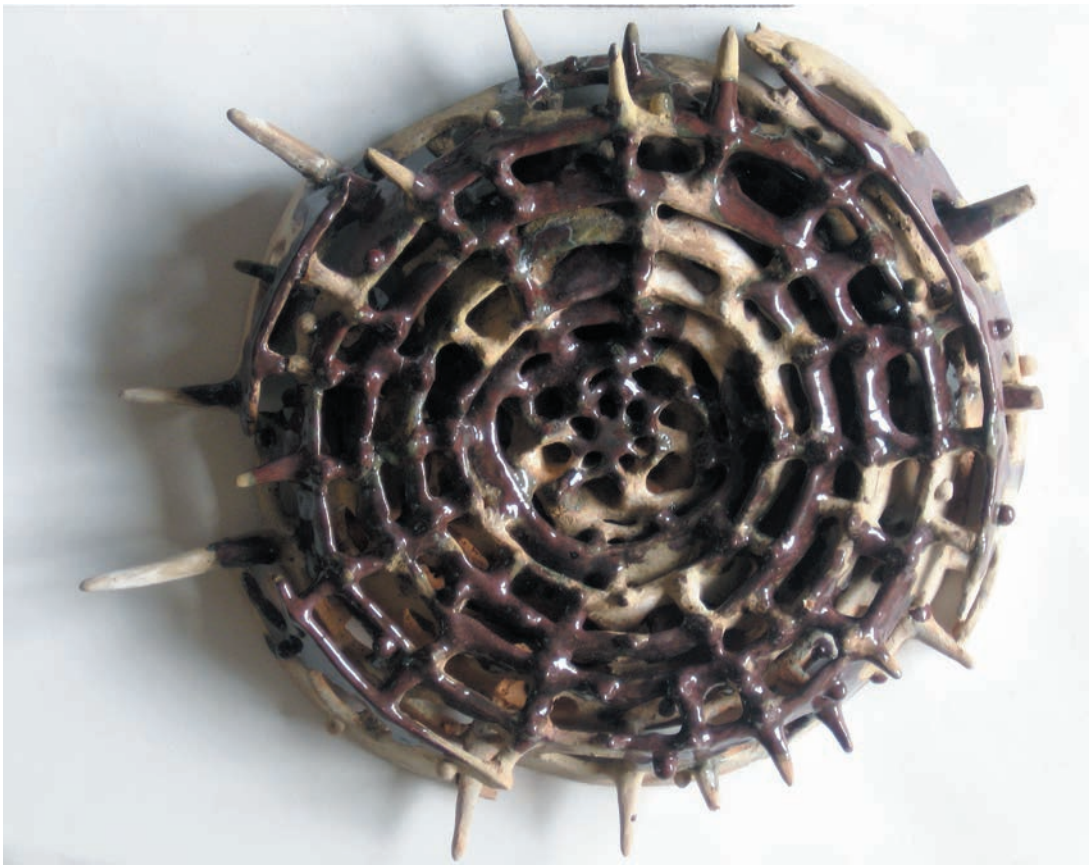
Los métodos sucintamente descritos aquí no son los únicos que pueden utilizarse para hacer reconstrucciones de las relaciones ancestro-descendientes. Existe, además, una familia de algoritmos muy útiles cuando se tienen cantidades masivas de datos; por ejemplo, aquellos que se enfocan en la distancia las calculan entre pares de especies para construir un árbol filogenético. Al respecto, se introducirán tres politopos, de los cuales el Tight Span y el de Lipschitz están relacionados con representaciones de la métrica que preservan distancia, mientras que el politopo de evolución mínima equilibrado se vincula con el principio de evolución mínima, que es un principio de optimización utilizado en filogenética y que en varios aspectos es similar al MP, ya que persigue el árbol con la menor longitud total, aunque la ruta para lograrlo puede ser distinta.

Teniendo una hipótesis filogenética robusta se pueden utilizar metodologías auxiliares para tratar de responder preguntas más profundas o establecer metas más ambiciosas. Algunas de estas directrices son el reloj molecular y la reconstrucción de estados ancestrales de caracteres. La primera de ellas permite establecer un marco temporal que ayuda a entender cómo y bajo qué condiciones evolucionó un grupo. La reconstrucción de caracteres estima los estados de una especie variable (como parte de su proceso evolutivo) para relacionarlos con eventos relevantes. Adicionalmente, se incluyen teorías que responden qué tan fuerte es la asociación entre la evolución de un conjunto de caracteres (p. e. componentes del nicho, morfología) con la filogenia.

Ernesto Álvarez-González
Pablo Duchén-Bocángel
Carlos Luis Leopardi-Verde
Linard Hoessly
Guadalupe Jeanett Escobedo-Sarti

"Aeonis"
Cerámica/ esmaltes y engobes
Dimensiones varias
2014

"Biocosmos"
Grafito sobre papel
40 x 50 cm
2013



"Diatomea I"
Cerámica/esmaltes y engobes
16.5 x 25 x 25 cm
2011

Tequio *4*(11), 2021: 7-25
ISSN: 2594-0546

Filogenias: conceptos y generalidades
Phylogenies: concepts and generalities

Carlos Luis Leopardi-Verde[1*] y Guadalupe Jeanett Escobedo-Sarti[2]

Fecha de recepción: 16 de noviembre de 2020
Fecha de aceptación: 5 de enero de 2021

Resumen - Cualquier investigación en biología es un ejercicio de comparación y eso incluye el estudio de la evolución. La indagación de patrones evolutivos en cualquiera de sus dos enfoques (micro o macroevolutivo) establece retos metodológicos para cualquier persona interesada en estos temas. Dichos enfoques tienen el interés común de comprender el origen de las relaciones de parentesco entre los organismos estudiados, aunque las escalas temporales y el nivel de organización en el que se concentran son diferentes. Actualmente, las filogenias son la herramienta más robusta para elucidar las hipótesis de relaciones ancestro-descendiente entre un conjunto de organismos. Estas representaciones consisten en proyecciones diagramáticas bidimensionales (cladogramas) o multidimensionales (redes) que pueden ser estimadas con diferentes aproximaciones (máxima parsimonia, máxima verosimilitud e inferencia bayesiana), según los datos disponibles y el propósito de la investigación. En esta revisión se presenta una introducción a los métodos disponibles para la construcción de filogenias, incluyendo la perspectiva tradicional que utiliza diagramas basados en dicotomías y las nuevas tendencias que tratan de visualizar patrones más complejos a través de redes evolutivas.

Palabras clave: Evolución, máxima parsimonia, máxima verosimilitud, inferencia bayesiana, redes filogenéticas.

Abstract - Any research in biology is an exercise of comparison that includes the study of evolution. The investigation of evolutionary patterns in either of its two approaches (micro or macroevolutionary) raises methodological challenges for any researcher interested in these topics. These approaches have a common interest in understanding the origin the relationships between the studied organisms, although the temporal scales and the level of organization in which they focus are different. Currently, phylogenies are the most robust tool to explain ancestor-descendant relationships between a set of organisms. These diagrams, which are two-dimensional (cladograms) or multidimensional (networks), can be estimated with different approximations (maximum parsimony, maximum likelihood, and bayesian inference) according to the data available and the purpose of the investigation. This review presents an introduction to the methods available for the construction of phylogenies, including the traditional perspective that uses diagrams based on dichotomies and the new trends that try to visualize more complex patterns through evolutionary networks.

Keywords: Evolution, maximum parsimony, maximum likelihood, bayesian inference, phylogenetic networks.

[1] Facultad de Ciencias Biológicas y Agropecuarias de la Universidad de Colima, Km. 40 Autopista Colima-Manzanillo, Crucero de Tecomán, Tecomán, Colima, México, C.P. 28930. *Autor de correspondencia: cleopardi@ucol.mx. ORCID: https://orcid.org/0000-0001-5172-5114
[2]ORCID: https://orcid.org/0000-0002-4901-971X

Introducción

Una parte importante del pensamiento humano se basa en comparaciones, por ello podemos saber si hace más o menos calor en un día soleado, si un alimento es dulce o salado, opinar sobre un sistema político, etcétera. De la misma manera, el pensamiento comparado es una parte esencial de cualquier persona que se dedica a la investigación de la biología evolutiva en sus distintos ámbitos, debido a que a través de las comparaciones se pueden descubrir y describir los patrones que caracterizan a los grandes fenómenos evolutivos (Harvey & Pagel, 1991).

La biología evolutiva comparte con otras ciencias –como la astronomía y la geología– la tarea de interpretar fenómenos que son imposibles de comprender si no se conoce su pasado (Harvey & Pagel, 1991). De hecho, la biología evolutiva puede considerarse parte de la sistemática, que para autores como Simpson (2019) incluye tanto a la taxonomía como al estudio de la evolución. La taxonomía se encarga de identificar a los seres vivos, describirlos, nombrarlos y clasificarlos, todo esto con reglas establecidas; en el caso de la botánica, en el Código Internacional de Nomenclatura de Algas, Hongos y Plantas (Turland *et al.,* 2018).

El estudio de la evolución se centra en comprender los patrones y procesos que dieron origen a los organismos tal como los vemos hoy, o en otras palabras, busca elucidar la historia de la vida. Esa historia puede analizarse a grandes rasgos en dos escalas: la microevolutiva, que está enfocada al estudio del cambio genético que acumulan las distintas poblaciones que conforman una metapoblación; normalmente, la escala temporal de esta aproximación está prácticamente circunscrita al presente y quizás a algunas decenas o cientos de años. La segunda escala es la macroevolutiva, en donde se estudian los cambios de linajes completos (conjuntos de especies) a lo largo de periodos de tiempo que corresponden más a la escala geológica, desde miles hasta millones de años (Herron & Freeman, 2014). La herramienta básica con la que se cuenta en biología evolutiva para expresar las hipótesis de relaciones ancestro-descendiente que conforman esa historia y para hacer inferencias sobre los patrones y procesos que pudieran explicarlas son las filogenias.

El enfoque microevolutivo utiliza marcadores moleculares altamente variables o secuencias y numerosas muestras de miembros de la misma especie. Según Hayward, Tollenaere, Dalton-Morgan & Batley (2015), un marcador molecular es un loci genético (una parte del genoma) que puede ser fácilmente rastreado y cuantificado en una población (o en un linaje) y que podría estar asociado con un gen o un carácter de interés. Esta información es procesada con herramientas estadísticas que permiten conocer el grado de variación o diferenciación entre las poblaciones e interpreta las relaciones entre éstas, básicamente utilizando un modelo de red en el que potencialmente todas las poblaciones pueden intercambiar información genética entre sí, lo que se expresa a través de alelos (las diferentes versiones que puede tener un gen) o haplotipos (conjunto de polimorfismos que se heredan juntos) compartidos o únicos. Este enfoque presupone que las especies están conformadas por poblaciones genéticamente vinculadas a través del intercambio ocasional de material genético, sea como gametos o en forma de individuos (Templeton, 2006; Herron & Freeman, 2014).

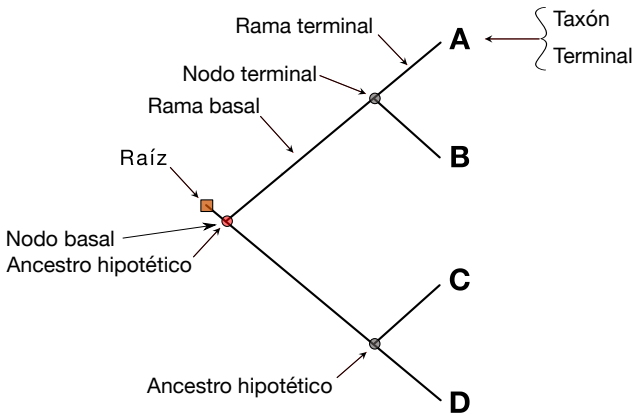
El enfoque macroevolutivo utiliza caracteres morfológicos y/o moleculares, estos últimos, aunque pueden ser de diversa índole, usualmente son secuencias de ADN de regiones específicas con grados de variación proporcionales a la profundidad con la que se quiera hacer el estudio (Wiley & Lieberman, 2011). En plantas, por ejemplo, para investigar grupos con divergencias profundas se han utilizado secuencias de regiones con un alto

grado de conservación, como *rbcL*, pero para estudiar grupos de reciente divergencia se ocupan secuencias de regiones altamente variables como ITS o ETS (Soltis & Soltis, 1999).

Los estudios macroevolutivos emplean herramientas estadísticas de diversa naturaleza para tratar de elucidar el grado de relación que guardan dos entidades (usualmente especies) entre sí. Comúnmente expresan la información generada en forma de filogenias, que son diagramas en forma de árboles en los que se presentan las relaciones hipotéticas entre un conjunto de ancestros y sus descendientes (Figura 1). Las filogenias idealmente son árboles dicotómicos en los que la raíz representa el origen o ancestro común a todas las entidades incluidas, cada punto de bifurcación (nodo) se interpreta como un ancestro e implica la aparición de nuevos linajes. La línea que conecta dos nodos se conoce como rama y la parte final en donde se encuentran las entidades incluidas se llama ramas terminales o simplemente terminales. El enfoque dicotómico tradicional presupone que, una vez formados los linajes, ya nunca más intercambian información genética entre sí (Knowles & Carstens, 2007). Aunque esta es la interpretación tradicional, en la literatura hay propuestas discordantes con tal visión y por ello se han planteado formas diferentes de análisis, de las cuales algunas son análogas a las redes con las que se interpretan las relaciones entre las poblaciones (Huson, Rupp & Scornavacca, 2010).

Figura 1.

Una filogenia hipotética y sus partes. En la figura se representa un cladograma perfectamente dicotómico con todas las ramas con la misma longitud. Note que tiene una raíz que marca el origen de la filogenia, el cual representa un ancestro hipotético al igual que cada nodo. Están las ramas que van conectando los nodos de la filogenia con los ancestros hipotéticos hasta que se llega a las ramas terminales que representan a los taxones incluidos en el muestreo.

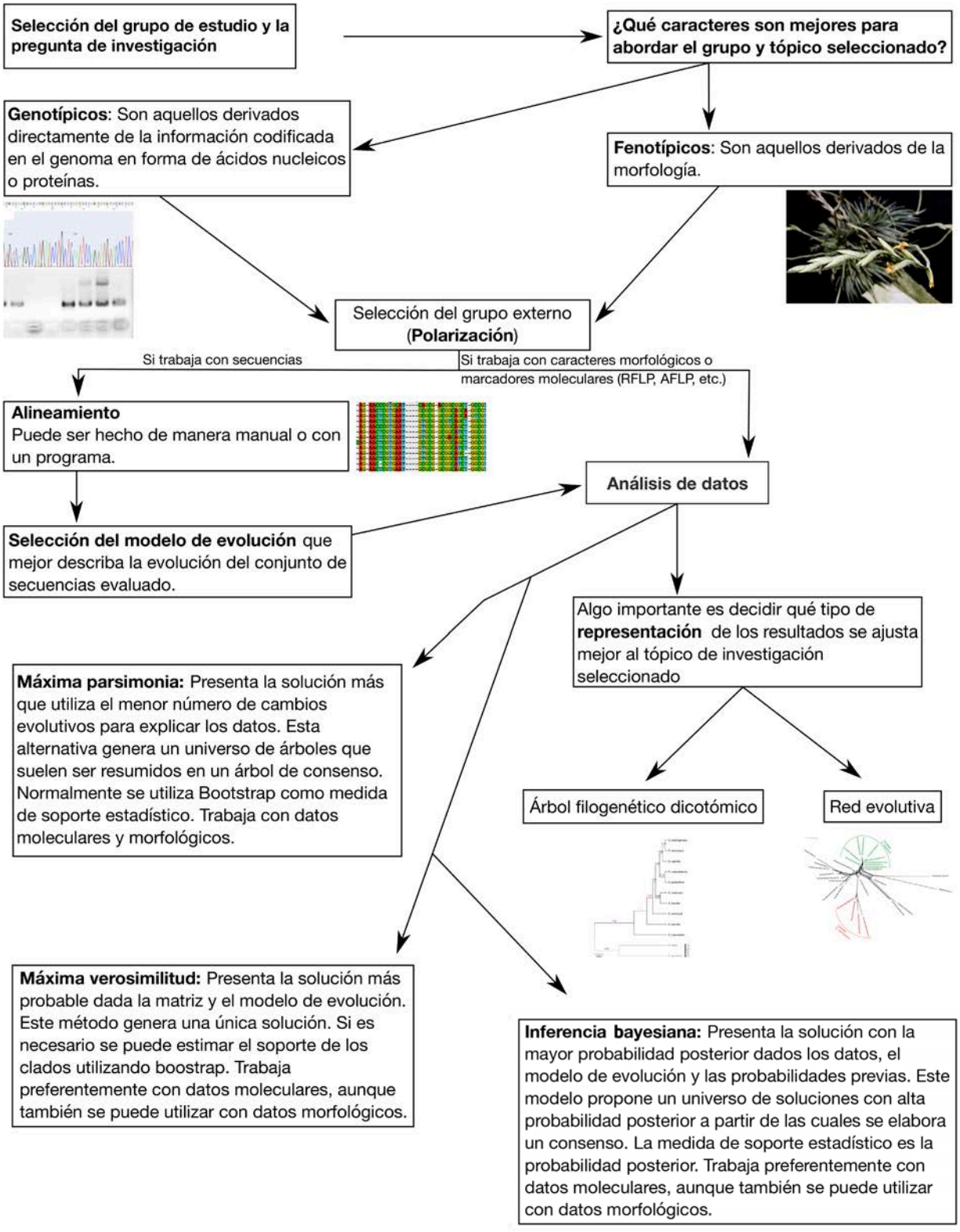


Es importante recordar que las filogenias expresan una hipótesis de relaciones evolutivas entre linajes y por lo mismo no deben ser confundidas con los dendrogramas o fenogramas. Estos últimos son diagramas dicotómicos que muestran similitud entre entidades en función de algo, por ejemplo, morfología (fenogramas) o en ecología se utilizan para comparar qué tanto se parecen dos comunidades (análisis de conglomerados).

Dado que esta contribución versa sobre el uso de métodos filogenéticos en un contexto tradicional, a continuación, se presentarán los elementos teóricos relacionados con una reconstrucción filogenética utilizando caracteres moleculares y/o morfológicos desde una perspectiva amplia, como se resume en la Figura 2. En una segunda contribución se aborda la utilidad de las herramientas filogenéticas para estudiar fenómenos evolutivos (evolución de caracteres, por mencionar alguno).

Figura 2.

Diagrama de flujo de una reconstrucción filogenética. Para detalles, consulte el documento.



La reconstrucción de una filogenia

La reconstrucción de una filogenia es un proceso cuyo primer paso requiere definir el grupo de estudio y su circunscripción. Tradicionalmente esto se hacía con criterios empíricos, sin un enfoque evolutivo y usualmente con el propósito de poner a prueba determinado esquema de clasificación. Actualmente, con base en la información que existe en la literatura, se hacen cada vez con más frecuencia propuestas menos conservadoras, con la intención de responder preguntas concretas que pueden ser prácticamente de cualquier índole imaginable (p. e. Givnish *et al.,* 2011; Smith & Hendricks, 2013).

Para elaborar una reconstrucción filogenética luego de la selección del grupo de estudio es necesario precisar el grupo externo. Este punto es esencial, pues se ocupa para definir la polaridad de la filogenia y por lo mismo es la base para establecer qué es lo "ancestral" y lo "derivado". Lo que sigue es la elección del tipo de caracteres, que pueden ser de las dos fuentes más comunes, a las que por simplicidad se denomina aquí caracteres de origen fenotípico y los de origen genotípico. Lo anterior es relevante porque no es lo mismo utilizar caracteres fenotípicos o genotípicos codificados como una matriz binaria o multiestado, que utilizar secuencias de ADN, cuyo abanico de opciones de análisis es mucho mayor.

En el caso de las secuencias de ADN, el primer paso es establecer un alineamiento en el cual se maximice la comparabilidad (y compatibilidad) entre las secuencias que conforman la muestra, luego se debe especificar el modelo evolutivo que se utilizará para analizar la matriz de secuencias alineadas (ver sección de modelos más adelante). Posteriormente se hace el análisis, que debe estar guiado por la pregunta que se desea responder, ya que cada uno de los paradigmas para el análisis de datos (máxima parsimonia, máxima verosimilitud[3] e inferencia bayesiana) presenta fortalezas y debilidades. Además, los productos de cada una de estas formas de análisis tienen propiedades diferentes, por lo que pueden ser usados para actividades tan distintas como establecer el grado de discrepancia de los datos utilizados, la tasa de evolución de un linaje, elaborar una escala temporal o entender cómo ha sido el proceso de cambio de un atributo.

Caracteres fenotípicos

Los caracteres fenotípicos son atributos que podemos observar o medir de la información contenida en el genoma de los organismos a analizar; pueden ser desde la presencia de determinado tipo de indumento, cambios en la forma, los tamaños, las proporciones, los ángulos, los colores, etcétera (Poe & Wiens, 2000). En los caracteres fenotípicos se pueden incluir también elementos como las fragancias, las características fisiológicas, los compuestos químicos, entre otros. Por supuesto, el uso de características con una homología discutible como las fragancias o los colores requiere de una justificación basada en investigación previa. Un carácter homólogo es aquel que tiene el mismo origen entre los taxa incluidos en el estudio (Herron & Freeman, 2014).

En este sentido, un carácter puede definirse como cualquier atributo que presenta variación entre dos o más grupos de organismos, rasgos que para ser empleados en los análisis filogenéticos deben ser constantes (Wagner, 2001). Este concepto es apropiado para describir cualquier actividad de la vida cotidiana y es una parte esencial de la biología comparada, que tiene por función examinar y capturar los patrones biológicos, y elaborar teorías sobre los procesos que podrían explicarlos (Eldredge & Cracraft, 1980).

Utilizar caracteres fenotípicos normalmente ofrece dos restricciones que deben salvarse; la primera es determinar, más allá de la duda razonable, si el atributo seleccionado cumple con el criterio de homología. La

[3] También se conoce como máxima probabilidad.

segunda tiene que ver con la definición precisa (sin ambigüedades) de los estados de carácter; esto es, de las diferentes formas en las que puede presentarse ese atributo (Wagner, 2001; Poe & Wiens, 2000). En muchos casos, los rasgos elegidos marcan discontinuidades muy claras, por ejemplo "verticilos trímeros" y "verticilos tetra- o pentámeros", el primero es un carácter que define a las monocotiledóneas, mientras que el segundo distingue a las eudicotiledóneas (Simpson, 2019).

Los caracteres fenotípicos pueden ser utilizados de varias maneras: (i) codificando de forma discreta a aquellos rasgos que podrían resultar informativos (Wagner, 2001; Poe & Wiens, 2000); (ii) empleando caracteres continuos discretizados a través de alguno de los métodos propuestos a lo largo de los últimos 30 años (para una revisión, consultar García-Cruz & Sosa, 2006); (iii) haciendo uso de caracteres continuos de forma directa; a diferencia de los enfoques anteriores, éstos pueden ser analizados mediante la parsimonia cuadrática propuesta por Goloboff, Mattoni & Quinteros (2006) o con inferencia bayesiana modelados bajo difusión browniana (*brownian Diffusion*), o una de las variantes del modelo de procesos estocásticos conocido en inglés como *Random Walks* (Lemey, Rambaut, Welch & Suchard, 2010).

Caracteres genotípicos

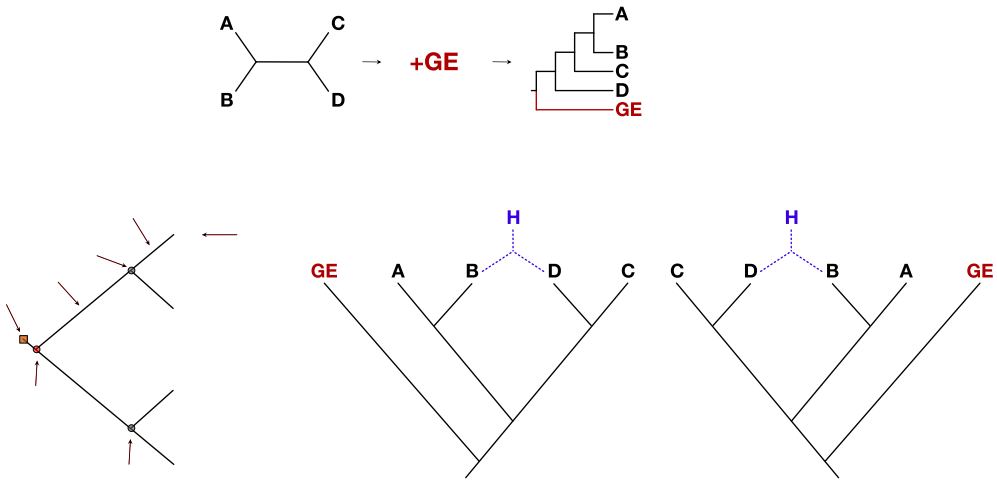
Los caracteres genotípicos se derivan del estudio directo de la información contenida en el genoma; esto puede ser a través de marcadores moleculares como los RAPD, AFLP, SSR, entre otros, así como del estudio directo de las secuencias de ADN. En la actualidad la fuente predilecta de información para construir filogenias es el uso de secuencias de ADN, que puede ser de regiones específicas del núcleo, cloroplasto o mitocondrias o una combinación de éstas, según la profundidad de las relaciones o grupo de organismos a estudiar (p. e. Jiang, Zhu, Song, Li, Yang & Yu, 2014; Leopardi-Verde, Carnevali & Romero-González, 2017). La elección de las regiones a usar normalmente está relacionada con la tasa de evolución; por ejemplo, en plantas la tasa de evolución del genoma del cloroplasto es mucho mayor que la de las mitocondrias y por lo mismo es más informativo (Soltis & Soltis 1999). En contraste, en los animales hay genes mitocondriales que son lo suficientemente variables como para que puedan ser utilizados como códigos de barras (p. e. Prado, Pozo, Valdez-Moreno & Hebert, 2011). Cabe mencionar que en los estudios más recientes ya se empiezan a ocupar cantidades masivas de datos, al punto que en algunos casos ya se emplean conjuntos genómicos completos (Kim *et al.,* 2020; Koenen *et al.,* 2020).

Polarización

Se dice que dos o más estados de un carácter están polarizados cuando se establece, bajo algún criterio, cuál fue el primero en evolucionar (Wiley & Lieberman, 2011); la misma idea aplica al escoger el grupo externo en una reconstrucción filogenética. La polarización es fundamental para el análisis filogenético, pues a través de ella es posible dar dirección a un árbol filogenético (Figura 3). Una consecuencia de esto es que se identifican las sinapomorfías que diagnostican a los grupos monofiléticos que emergen de un estudio. Antes de continuar, es importante aclarar que una apomorfía es un carácter único de un grupo que se considera una sinapomorfía cuando es compartido por todos los miembros de un grupo evolutivamente derivado. Mientras que si es compartido por todos los miembros del grupo analizado se le denomina simplesiomorfía, por ejemplo, el pelo es un carácter sinapomórfico de los mamíferos si se ve en el contexto de la filogenia de los vertebrados; sin embargo, si nos enfocamos exclusivamente en la filogenia de los mamíferos este carácter es considerado simplesiomórfico, porque todos lo comparten.

Figura 3.

El efecto de polarizar. De izquierda a derecha se muestra un árbol no enraizado, al que se añade un grupo externo (GE) para dar dirección a la filogenia.



Se han descrito muchos métodos para la polarización, no obstante, el más ampliamente aceptado es la comparación con un grupo externo, por dos razones: la primera, porque sus supuestos son confiables y sencillos; la segunda es porque la información que se requiere para su aplicación en la mayor parte de los casos está disponible y se recolecta al mismo tiempo que para el grupo de organismos de interés (Bryan, 2001). La comparación con un grupo externo establece la ancestría identificando el estado compartido con el taxón o taxa designados para este fin.

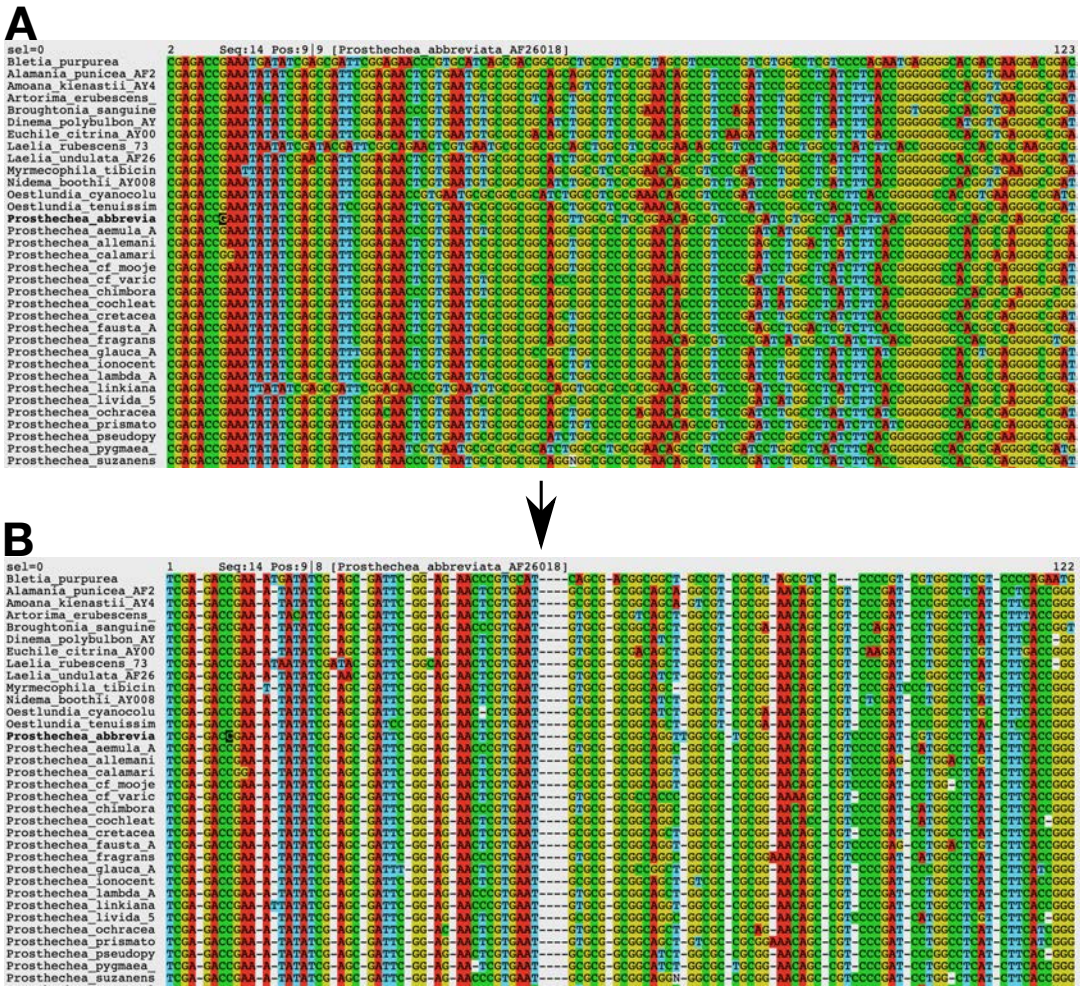
El método ontogénico se basa en la comparación de los patrones de desarrollo de las especies y establece la ancestría de un estado de carácter utilizando como criterio el orden en que los distintos estados de ese carácter aparecen en la ontogenia de los organismos. A pesar de que este método se basa en evidencia primaria, debido a lo difícil que es hacer estudios de patrones de desarrollo para muchos organismos y a que existen diversas corrientes de interpretación para esos resultados, se considera que es un criterio poco práctico (Harvey & Pagel, 1991; Bryan, 2001). Un tercer método es el del uso del registro fósil, que establece que los estados ancestrales son aquellos que aparecen primero en los fósiles. Debido a que la evidencia fósil es fragmentaria, este criterio ha caído en desuso (Bryan, 2001). Cabe notar que el único criterio válido para la polarización de una filogenia, cuando se trabaja usando como información base secuencias de ADN, es la selección del grupo externo; los otros dos son útiles cuando se intenta polarizar sólo caracteres morfológicos.

Métodos para la reconstrucción de una filogenia

Alineamiento. Consiste en organizar las secuencias (ADN, ARN o proteínas) considerando cada posición comparable de la matriz (p.e. la posición 355 del alineamiento) como una hipótesis de homología posicional (Figura 4). Usualmente estas hipótesis deben ser lo más parsimoniosas posible (se prefieren las opciones más simples sobre las más complejas) y persiguen inferir las relaciones evolutivas sin necesidad de un conocimiento previo de los eventos evolutivos que dieron origen a esos cambios (Morrison, 2006)

Figura 4.

Ejemplo de un alineamiento. A) Segmento de una matriz de ITS sin alinear. B) Segmento de la matriz de ITS luego del alineamiento.



Los alineamientos pueden hacerse manualmente o con ayuda de programas de cómputo. Los manuales presentan el problema de ser difíciles de repetir; mientras que los elaborados con programas, aunque pueden ser reproducidos con facilidad, usualmente requieren ser refinados a mano. Los programas pueden seguir tres métodos básicos para alinear las secuencias: global, por bloques o combinados; también pueden ser progresivos, iterativos o ser una mezcla de ambos (Mount, 2001; Pivorano & Heringa, 2008).

Los métodos globales utilizan las secuencias completas durante el alineamiento (p. e. ClustalW y MUSCLE). Por su parte, los que funcionan por bloques identifican patrones en las secuencias y los usan para elaborar el alineamiento (p. e. MAFFT). Sólo los métodos más recientes, como T-Coffee, son capaces de incorporar estas dos metodologías (Mount, 2001; Pivorano & Heringa, 2008).

Los métodos progresivos intentan construir el alineamiento generando un "boceto" de árbol filogenético, para lo que usan una distancia genética y algoritmos como el método de grupo de pares no ponderados con

media aritmética (UPGMA); este boceto es empleado para ir añadiendo una a una las secuencias a ser alineadas. Por conveniencia, los algoritmos alinean las secuencias por pares. Por ello, cuando se utiliza un alineamiento progresivo, es de extrema importancia que las dos primeras secuencias se alineen correctamente, ya que su resultado afecta al resto del proceso (Mount, 2001). En los alineamientos progresivos, una vez que se coloca un gap, en esa posición siempre habrá un gap (Mount, 2001); ClustalW es un método que funciona de esta manera. Los gaps también se conocen como indels y son llamados así porque representan una inserción o una deleción en el genoma, reciben este nombre porque no es posible a *priori* saber de qué se trata.

Los métodos iterativos se diferencian de los progresivos en que repiten tantas veces como sea necesario el alineamiento de las secuencias, presentando al final una secuencia de consenso. En la actualidad, es común el uso de algoritmos que combinan ambas aproximaciones (son globales e iterativos), como MUSCLE y T-Cofee (Pivorano & Heringa, 2008).

Recurrir a uno u otro tipo de algoritmos es una decisión multifactorial. Edgar & Batzoglou (2006) hacen una serie de recomendaciones al respecto, por ejemplo, si se es muy tradicional, la opción lógica es ClustalW que da resultados bastante buenos, aunque es un algoritmo lento y menos preciso que los métodos más modernos. Si se manejan 500-1000 secuencias en la matriz que será alineada, MUSCLE es más eficiente que ClustalW en el uso del tiempo y su carácter iterativo disminuye sustancialmente los errores de identidad que pueden ocurrir en los alineamientos de ClustalW. Otra opción muy buena cuando se tienen más de 500 secuencias o secuencias con patrones muy complejos es MAFFT. T-Coffee produce alineamientos muy buenos, pero es lento, demanda amplios recursos del sistema y no puede ser utilizado con más de 50 secuencias a la vez.

Para hacer los alineamientos hay muchas alternativas, desde utilizar los ejecutables de los algoritmos directamente, hasta emplear programas que los integran, como MEGA, Seaview, Geneious, entre otros. Otra opción es recurrir a servidores dedicados a alineamientos. En el sitio web del libro de Lemey, Salemi & Vandamme (2009)[4] existe una lista de algoritmos disponibles e indica dónde pueden ser utilizados o descargados. Adicional a esto, el servidor del Laboratorio Europeo de Biología Molecular (EMBL-EBI) pone a disposición del público, libre y gratuitamente, algunos de los algoritmos más populares.[5]

Selección del modelo de evolución. Para reconstruir una hipótesis de relaciones filogenéticas es indispensable contar con un modelo que describa cómo las secuencias de ADN o proteínas evolucionan a través de mutaciones, deriva génica, selección, recombinación, etcétera (Huson *et al.,* 2010). En palabras más formales, un modelo de evolución es un conjunto de parámetros que se utilizan para describir el patrón de sustituciones que caracterizan a una matriz de secuencias de ADN o proteínas (Strimmer & von Haeseler, 2009). En este sentido, cuando se trabaja con secuencias, la selección del modelo es un paso crítico, pues tanto la subparametrización como la sobreparametrización pueden alterar los resultados del análisis (Lemmon & Moriarty, 2004).

Actualmente existen más de 88 modelos que reflejan distintos grados de complejidad en la evolución del ADN (Posada, 2008); entre ellos, los más importantes en orden de complejidad son Jukes y Cantor (JC), Felsenstein 81 (F81), Kimura dos parámetros (K2P), Hasegawa-Kishino-Yano (HKY), Kimura tres parámetros (K3P), Tamura-Nei (TrN), Simétrico (SYM) y el modelo general de tiempo reversible (GTR, por "*general time reversible*"). La diferencia entre cada uno de estos modelos es el número y tipo de parámetros usados; por ejemplo, JC sólo considera la tasa de sustitución, mientras que F81 añade a esto la proporción en la que están las bases y así se van incrementando los parámetros hasta llegar al más complejo, el GTR, que toma en cuenta todos los parámetros posibles como

[4] https://www.kuleuven.be/aidslab/phylogenybook/Table3.1.html
[5] https://www.ebi.ac.uk/Tools/msa/

sustituciones, frecuencias de las bases, transiciones, entre otros, y que puede hacerse más complejo al incluir la probabilidad de sitios invariables (I) y Gamma (γ); este último es una medida de la heterogeneidad entre los sitios que componen el alineamiento (Huson *et al.,* 2010).

Así, para un principiante quizás lo más práctico es dejar que un programa de computadora analice la matriz de datos y le sugiera tanto el modelo más apropiado, como los parámetros que mejor se ajustan a los datos; mientras que para un experto esto es una ayuda importante en el proceso, aunque pueden ir más allá si así lo desean. Para la selección de modelos hay varios programas, aunque uno de los más populares es Modeltest o su actualización Modeltest-NG (Posada, 2008; Strimmer y von Haeseler, 2009; Darriba *et al.,* 2019). Si se desea aprender más sobre los modelos de evolución, un par de lecturas excelentes son el capítulo 3 de Huson et al. (2010) y Strimmer & von Haeseler (2009).

Estrategias de análisis. La máxima parsimonia (MP) es un método que busca el árbol más corto que mejor explica los datos observados (Whelan, 2008; Swofford & Sullivan, 2009). La idea de esta búsqueda se basa en el principio conocido como la navaja de Ockham (*Occam's razor*), que sugiere que cuando hay dos o más hipótesis que proveen una respuesta igualmente probable, la más simple suele ser la mejor. En este sentido, en la MP el árbol más simple es aquel que minimiza la cantidad de cambios evolutivos que se requieren para explicar los datos (Swofford & Sullivan, 2009). Minimizar la cantidad de cambios evolutivos implica seleccionar el árbol con menor número de homoplasias, que son aquellos caracteres que se adquieren de forma independiente por grupos no relacionados desde el punto de vista evolutivo, este término también se conoce como evolución paralela (Sanderson & Hufford, 1996, Swofford & Sullivan, 2009). Ejemplo de caracteres homoplásicos son la partenogénesis en reptiles o la forma esférica de algunas *Cactaceae y Euphorbia* obesa (*Euphorbiaceae*).

En general, la MP tiene la ventaja de ser bastante intuitiva y de requerir pocos recursos computacionales en comparación con métodos estadísticos más demandantes; sin embargo, su problema es que si las ramas son muy largas, puede sugerir relaciones incorrectas con un soporte elevado.

Para buscar el árbol más corto, la parsimonia puede intentar generar todos los árboles, lo cual se conoce como búsqueda exhaustiva (es viable sólo hasta 12 terminales) o puede muestrear la población de árboles a través de una búsqueda heurística tradicional o utilizando el algoritmo Ratchet (Nixon, 1999). Independiente de cómo se establezca la búsqueda heurística, el programa utilizado muestreará la población de árboles a través de algoritmos como el "intercambio con el vecino más cercano" (NNI), el "cortado y redibujado de árboles" (SPR) o la "bisección y reconexión de árboles" (TBR), que respectivamente representan de la búsqueda más simple a la más completa (Swofford & Sullivan, 2009).

En MP el Bootstrap es empleado con frecuencia para evaluar el soporte estadístico de los árboles. Este análisis es una técnica estadística de remuestreo de datos con réplica, en la que la mitad del conjunto de caracteres es eliminado, los datos que permanecen son duplicados de tal manera que el grupo de caracteres es del mismo tamaño que al inicio; a partir de éste se realiza el análisis filogenético y se obtiene la filogenia más parsimoniosa. El proceso descrito se repite tantas veces como sea necesario, lo usual es al menos mil ocasiones y luego se estima a partir de todos los árboles generados con el conjunto de datos parcialmente simulados la proporción en la que se presenta cada clado, incluido en la filogenia basada en datos reales (Schmidt, 2009). Algunos programas populares para construir árboles utilizando máxima parsimonia son PAUP* (Swofford, 2002), Phylip (Felsenstein, 1989), TNT (Goloboff, Farris & Nixon, 2008), Nona (Goloboff, 1999) y WinClada (Nixon, 2002).

El método de máxima verosimilitud (ML, por sus siglas en inglés) toma un modelo de evolución de secuencias de ADN o proteínas y busca la combinación de valores de los parámetros que componen el modelo que genere el

árbol filogenético capaz de representar con la mayor probabilidad las secuencias que integran la matriz utilizada. En otras palabras, es la probabilidad de los datos dado un árbol (Lewis, 1998; Bromham & Penny, 2003).

Algo interesante de la ML es que entre los factores que se toman en cuenta al estimar un árbol filogenético están la longitud de las ramas y la topología. Por ello es posible que un árbol tenga una topología "correcta", pero si su longitud de ramas está mal calculada, no necesariamente tendrá una buena probabilidad. La precisión en la estimación de la longitud de ramas convierte a este método en una herramienta indispensable cuando se quiere evaluar hipótesis de evolución de caracteres o estimar tasas de diversificación. Algunos programas populares para construir árboles utilizando ML son RAxML-NG (Kozlov, Darriba, Flouri, Morel & Stamatakis, 2019), PhyML (Guindon *et al.,* 2010), entre otros. Para leer más al respecto, una introducción excelente es Lewis (1998).

La inferencia bayesiana (IB) es un método que, con base en un modelo específico de sustitución, dado los datos y las probabilidades previas, selecciona el árbol que tiene la mayor probabilidad posterior (Bromham & Penny, 2003). Este método utiliza modelos evolutivos acoplados a Cadenas de Markov Monte Carlo y a otros algoritmos como Metrópolis para poder hacer sus estimaciones (Ronquist, van den Mark & Huelsenbeck, 2009).

La IB funciona haciendo un muestreo a través de todas las poblaciones de árboles que se "cruzan" en el camino de la cadena de Markov. Generalmente el análisis se deja correr por un número de generaciones que sea suficientemente grande, como para que las cadenas alcancen la estabilidad y converjan. Se recomienda hacer varias corridas, a fin de asegurarse de que las cadenas realmente convergieron y que se muestrearon todas las poblaciones de árboles (Ronquist, van den Mark & Huelsenbeck, 2009). En este tipo de análisis es conveniente tener en cuenta que los árboles son escogidos conforme a una función de probabilidad y que todos están conectados (Whelan, 2008).

La probabilidad posterior (PP) puede considerarse como la probabilidad, valga la redundancia, de que un nodo de la filogenia obtenida bajo un determinado modelo de evolución sea cierto. De hecho, estudios han demostrado que estos valores pueden interpretarse directamente (Alfaro, Zoller & Lutzoni, 2003), por lo que una probabilidad posterior de 0.95 equivale a 95% de probabilidad de que sea cierto. El punto débil de la IB y sus probabilidades posteriores es la sensibilidad a parámetros como α (que indica la forma de la distribución) y a otros componentes intrínsecos del modelo, entre los que se encuentran la probabilidad de transversión/inversión (κ) y sitios invariables (I). De hecho, una sobreparametrización del modelo puede "inflar" los valores de las PP, aunque esto no es tan grave como la subparametrización, que sí puede alterar la topología y ocasionar un alto soporte a clados falsos (Alfaro *et al.,* 2003; Ronquist *et al.,* 2009). Algunos programas populares para construir árboles utilizando IB son MrBayes (Ronquist *et al.,* 2012), BEAST (Bouckaert *et al.,* 2019), entre otros.

Redes filogenéticas (*Phylogenetic networks*)

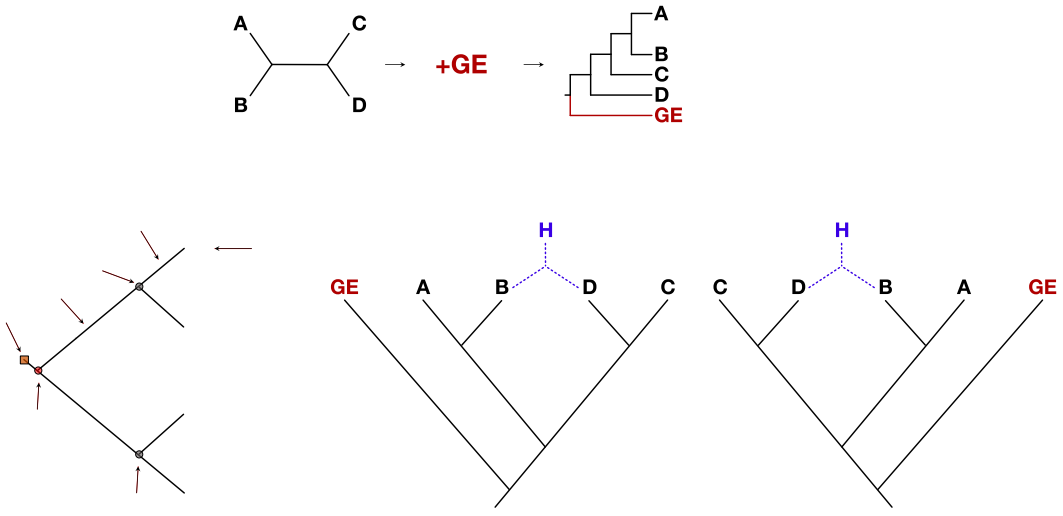
En la mayor parte de los trabajos de sistemática que involucran el uso de filogenias, la representación de las relaciones ancestro-descendiente se hace a través de dicotomías (Huson *et al.,* 2010). Este tipo de representación tiene una suposición subyacente: cada vez que ocurre un proceso de cladogénesis en un linaje, éste se divide sólo en dos nuevos linajes que nunca más interaccionan (Figura 1); no obstante, existe evidencia de que linajes diferentes pueden interactuar (Levin, 2004; Mallet, 2005; Via & West, 2008). Un ejemplo claro de este tipo patrón evolutivo son los híbridos (Soltis & Soltis, 2009; Russell *et al.,* 2010). Se han documentado procesos de recombinación o transferencia horizontal de genes entre especies (Richardson & Palmer, 2006; Czislowski *et al.,* 2018) y también hay casos en que las relaciones son mucho más complicadas que una dicotomía, como las radiaciones evolutivas que caracterizan a virus, como al de Inmunodeficiencia Humana (VIH) o al de la influenza

(Chan, Carlsson & Rabadan, 2013). En los anteriores y otros similares las relaciones reconstruidas que usan el enfoque dicotómico aportan poca o ninguna información sobre los procesos evolutivos subyacentes al grupo de estudio.

Por ello, cuando se sospecha que los procesos evolutivos no siguen un patrón dicotómico, es conveniente explorar los juegos de evidencia con herramientas que permitan el análisis de las reticulaciones y su representación a través de diagramas multifurcados, conocidos como redes filogenéticas (Figura 5). Para diferenciar una filogenia estándar de una red evolutiva, lo primero que debe quedar claro es que un árbol filogenético es un diagrama en el que se representan las relaciones ancestro-descendiente entre los taxa, a través de una serie de líneas (ramas) y conexiones (nodos) en los que se puede sugerir (árbol enraizado) o no (árbol no enraizado) una dirección en los procesos evolutivos (Figuras 1 y 3). Una característica importante de este diagrama es que no se forman ciclos y es estrictamente bifurcado cuando está completamente resuelto. Una red filogenética, en cambio, es cualquier gráfico utilizado para representar las relaciones evolutivas entre un grupo de organismos (sea abstracta o explícitamente), en el que se da nombre a algunos nodos del diagrama mientras que otros funcionan como conectores (*edges*) (Huson *et al.,* 2010). Dependiendo de si la hipótesis es explícita o abstracta, los nodos pueden representar entidades ancestrales o no.

Figura 5.

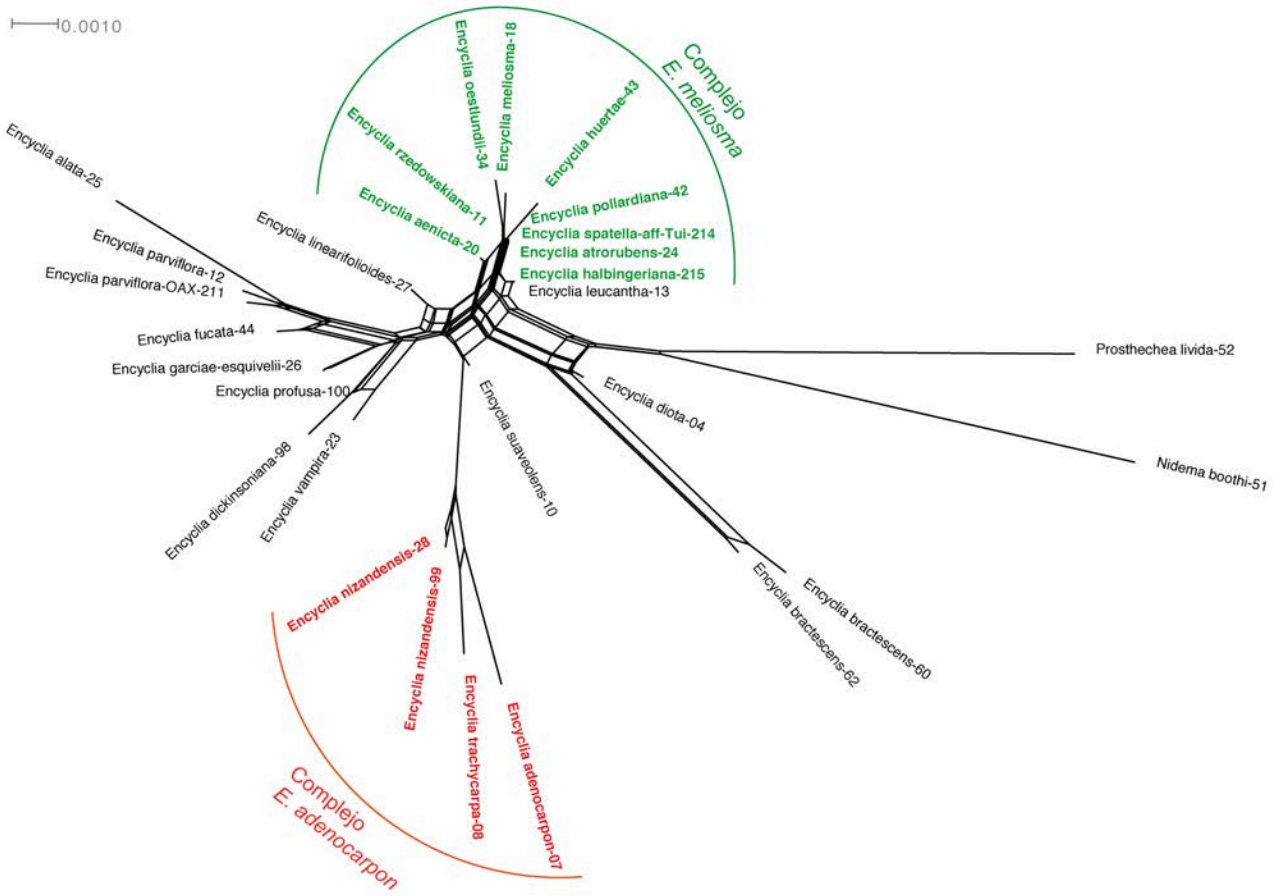
Filogenia hipotética en la que se muestra la posición de un taxón híbrido (H).



Las redes evolutivas pueden construirse utilizando como bloques de datos secuencias, distancias o árboles; asimismo, es posible procesar los bloques de datos empleando cualquier modelo evolutivo o paradigma de análisis (MP, ML, IB) y pueden ser enraizadas o no. Uno de los métodos más usados para elaborar redes es el conocido como redes divisivas *(splits networks*) (Figura 6). Para el caso de los esquemas enraizados, éstos pueden ser construidos utilizando varias herramientas, una de las más interesantes son las filogenias reticuladas, llamadas en inglés *galled networks*, aunque hay redes enraizadas para casos específicos como las redes de hibridación (*hybridization networks*), de recombinación (*recombination networks*), entre otras (Huson *et al.,* 2010).

Figura 6.

Ejemplo de una red filogenética tipo *split network*. En la figura se aprecian las relaciones entre especies del género *Encyclia (Orchidaceae: Laeliinae*) basadas en el gen *Phψ C*. Figura elaborada utilizando datos propios.



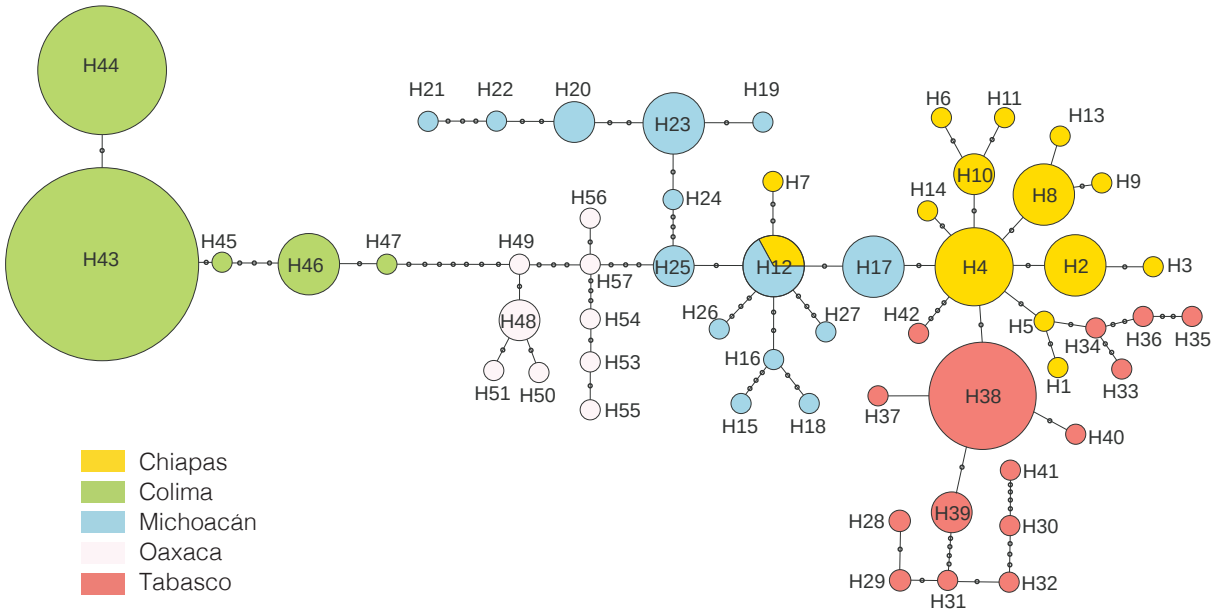
Las redes filogenéticas son especialmente útiles cuando se trata de entender historias evolutivas complejas en las que hay un alto intercambio de información entre linajes, por ejemplo, Stavrinides & Guttman (2004) estudiaron la evolución del coronavirus causante del Síndrome Respiratorio Agudo (SARS) en varios linajes de aves y mamíferos, con el fin de dilucidar su origen. La información base que utilizaron fueron secuencias de los genes que codifican para la replicasa, las espinas superficiales de la cápside, la matriz y las proteínas de la nucleocápside. Los análisis muestran que los linajes de estos virus relacionados con aves y mamíferos interaccionan generando reticulaciones que podrían deberse a recombinación entre éstos. En un trabajo relacionado, Forster, Forster, Renfrew & Forster (2020) utilizaron redes filogenéticas para analizar el genoma del SARS-CoV-2 y encontraron que la forma inicial de la red es consistente con el primer patrón de contagio, pero se hace difuso a medida que se acumulan efectos fundadores y el virus va mutando.

El enfoque de redes evolutivas también es útil por debajo del nivel de especie, en donde los linajes suelen intercambiar información genética. Por ejemplo, el uso de redes en forma de árboles de expansión mínima permite

comprender el flujo de genes entre poblaciones de una misma especie. Estas herramientas no distinguen entre grupos, por lo que pueden servir lo mismo para elucidar los patrones de domesticación de cerdos (Wu *et al.,* 2007) que para comprender la diversidad genética de un hongo fitopatógeno (Figura 7) (Manzo-Sánchez *et al.,* 2019).

Figura 7.

Una red de expansión mínima de los 57 haplotipos encontrados en cinco poblaciones de *Pseudocercospora fijiensis* (M. Morelet) Deighton. Los círculos negros pequeños representan los haplotipos faltantes, los círculos coloreados de diferente tamaño son haplotipos; los colores indican las poblaciones a las que pertenecen y el tamaño representa la frecuencia del haplotipo. Figura reproducida de Manzo-Sánchez *et al.,* 2019.



Conclusión

El estudio de patrones evolutivos es un área fascinante de la biología comparada. Las herramientas disponibles son variadas y permiten utilizar casi cualquier fuente de datos disponibles, desde la morfología hasta datos genómicos. Con esta variedad de datos también han ido cambiando las herramientas, iniciando por la máxima parsimonia, máxima verosimilitud, hasta los métodos que usan inferencia bayesiana y que son los más recientes. Cada una de estas formas de análisis tiene sus fortalezas y debilidades. De igual manera, se han flexibilizado las opciones de representar las relaciones y ya no sólo se piensa en dicotomías, sino que hay maneras de ver patrones más complejos de evolución como las reticulaciones. Sin embargo, los usos de las filogenias son muy diversos, como las estimaciones de edad que permiten los relojes moleculares o las comparaciones ecológicas en un contexto evolutivo.

Referencias

Alfaro, M. E., Zoller, S. & Lutzoni, F. (2003). Bayes or bootstrap? a simulation study comparing the performance of Bayesian Markov Chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular Biology and Evolution*, *20*(2), 255-266. doi: 10.1093/molbev/msg028

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., ... Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Computational Biology, 15*(4), e1006650. doi: 10.1371/journal.pcbi.1006650

Bromham, L. & Penny, D. (2003). The modern molecular clock. *Nature Reviews Genetics*, *4*(3), 216-224. doi: 10.1038/nrg1020

Bryan, H. N. (2001). Character polarity and the rooting of cladograms. En G. P. Wagner (Ed.), *The Character Concept in Evolutionary Biology* (319-341). San Diego, Estados Unidos: Academic Press.

Chan, J. M., Carlsson, G. & Rabadan, R. (2013). Topology of viral evolution. *Proceedings of the National Academy of Sciences*, *110*(46), 18566-18571. doi: 10.1073/pnas.1313480110

Czislowski, E., Fraser-Smith, S., Zander, M., O'Neill, W. T., Meldrum, R. A., Tran-Nguyen, L. T. T., Batley, J. & Aitken, E. A. B. (2018). Investigation of the diversity of effector genes in the banana pathogen, *Fusarium oxysporum* f. sp. *cubense*, reveals evidence of horizontal gene transfer. *Molecular Plant Pathology*, *19*(5), 1155-1171. doi: 10.1111/mpp.12594

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B. & Flouri, T. (2019). ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. *Molecular Biology and Evolution*, *37*(1), 291-294. doi: 10.1101/612903

Edgar, R. C. & Batzoglou, S. (2006). Multiple sequence alignment. *Current Opinion in Structural Biology 16* (3), 368–373. DOI: 10.1016/j.sbi.2006.04.004

Eldredge, N. & Cracraft, J. (1980). *Phylogenetic patterns and the evolutionary process, methods and theory in comparative biology*. New York: Columbia University Press.

Felsenstein, J. (1989). PHYLIP-Phylogeny Inference Package (Version 3.2). *Cladistics 5*(2), 164-166. doi: 10.1111/j.1096-0031.1989.tb00562.x

Forster, P., Forster, L., Renfrew, C. & Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences, 117*(17), 9241-9243. doi: 10.1073/pnas.2004999117

García-Cruz, J. & Sosa, V. (2006). Coding quantitative character data for phylogenetic analysis: A comparison of five methods. *Systematic Botany*, *31*(2), 302-309. doi: 10.1600/036364406777585739

Givnish, T. J., Barfuss, M. H. J., Ee, B. V., Riina, R., Schulte, K., Horres, R., Gonsiska, P. A., Jabaily, R. S., Crayn, D. M., Smith, A. C., Winter, K., Brown, G. K., Evans, T. M., Holst, B. K., Luther, H., Till, W., Zizka, G., Berry, P. E. & Sytsma, K. J. (2011). Phylogeny, adaptive radiation, and historical biogeography in *Bromeliaceae*: Insights from an eight-locus plastid phylogeny. *American Journal of Botany*, *98*(5), 872-895. doi: 10.3732/ajb.1000059

Goloboff, P. (1999). *NONA (NO NAME)* (Version 2). Tucumán, Argentina: autor.

Goloboff, P. A., Farris, J. S. & Nixon, K. C. (2008). TNT, a free program for phylogenetic analysis. *Cladistics, 24*(5), 774-786. doi: 10.1111/j.1096-0031.2008.00217.x

Goloboff, P. A., Mattoni, C. I. & Quinteros, A. S. (2006). Continuous characters analyzed as such. *Cladistics*, *22*(6), 589-601. doi: 10.1111/j.1096-0031.2006.00122.x

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology, 59*(3), 307-321. doi: 10.1093/sysbio/syq010

Harvey, P. H. & Pagel, M. D. (1991). *The Comparative Method in Evolutionary Biology*. Oxford, Reino Unido: Oxford University.

Hayward, A. C., Tollenaere, R., Dalton-Morgan, J. & Batley, J. (2015). Molecular marker applications in plants. En Batley (Ed.), *Plant genotyping: Methods and protocols* (13-27). New York: Springer. doi: 10.1007/978-1-4939-1966-6_2

Herron, J. C. & Freeman, S. (2014). *Evolutionary analysis*. Glenview, Estados Unidos: Pearson.

Huson, D., Rupp, R. & Scornavacca, C. (2010). *Phylogenetic Networks. Concepts, Algorithms and Applications*. New York: Cambridge University Press.

Jiang, W., Zhu, J., Song, C., Li, X., Yang, Y. & Yu, W. (2014). Molecular phylogeny of the butterfly genus *Polytremis* (Hesperiidae, Hesperiinae, Baorini) in China. *PLOS ONE*, *8*(12), 1-15. doi: 10.1371/journal.pone.0084098

Kim, Y.-K., Jo, S., Cheon, S.-H., Joo, M.-J., Hong, J.-R., Kwak, M. & Kim, K. J. (2020). Plastome evolution and phylogeny of Orchidaceae, with 24 new sequences. *Frontiers in Plant Science*, 11, 22. doi: 10.3389/fpls.2020.00022

Knowles, L. L. & Carstens, B. C. (2007). Delimiting species without monophyletic gene trees. *Systematic Biology*, *56*(6), 887-895. doi: 10.1080/10635150701701091

Koenen, E. J. M., Ojeda, D. I., Steeves, R., Migliore, J., Bakker, F. T., Wieringa, J. J., Kidner, C., Hardy, O. J., Pennington, R. T., Bruneau, A. & Hughes, C. E. (2020). Large-scale genomic sequence data resolve the deepest divergences in the legume phylogeny and support a near-simultaneous evolutionary origin of all six subfamilies. *New Phytologist*, *225*(3), 1355-1369. doi: 10.1111/nph.16290

Kozlov, A. M., Darriba D., Flouri, T., Morel, B. & Stamatakis A. (2019). RAxML-NG: A fast, scalable, and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics*, *35*(21), 4453-4455. doi: 10.1093/bioinformatics/btz305

Lemey, P., Rambaut, A., Welch, J. J. & Suchard, M. A. (2010). Phylogeography takes a relaxed random walk in continuous space and time. *Molecular Biology and Evolution*, *27*(8), 1877-1885. doi: 10.1093/molbev/msq067

Lemey, P., Salemi M. & Vandamme A. (Eds.) (2009). *The phylogenetic handbook: A practical approach to phylogenetic analysis and hypothesis testing*. New York: Cambridge University Press.

Lemmon, A. R. & Moriarty, E. C. (2004). The importance of proper model assumption in bayesian phylogenetics. *Systematic Biology*, *53*(2), 265-277. doi: 10.1080/10635150490423520

Leopardi-Verde, C. L., Carnevali, G. & Romero-González, G. A. (2017). A phylogeny of the genus *Encyclia* (*Orchidaceae: Laeliinae*), with emphasis on the species of the Northern Hemisphere. *Journal of Systematics and Evolution*, *55*(2), 110-123. doi: 10.1111/jse.12225

Levin, D. (2004). Ecological speciation: Crossing the divide. *Systematic Botany*, *29*(4), 807-816. doi: 10.1600/0363644042451134

Lewis, P. O. (1998). Maximun likelihood as an alternative to parsimony for inferring phylogeny using nucleotide sequence data. En D. E. Soltis, P. S. Soltis & J. J. Doyle (Eds.), *Molecular Systematics of Plants II-DNA Sequencing* (132-163). New York: Springer.

Mallet, J. (2005). Hybridization as an invasion of the genome. *TRENDS in Ecology and Evolution*, *20*(5), 229-237. doi: 10.1016/j.tree.2005.02.010

Manzo-Sánchez, G., Orozco-Santos, M., Islas-Flores, I., Martínez-Bolaños, L., Guzmán- González, S., Leopardi-Verde, C. L. & Canto-Canché, B. (2019). Genetic variability of *Pseudocercospora fijiensis*, the black Sigatoka pathogen of banana (*Musa* spp.) in Mexico. *Plant Pathology*, *68*(3), 513-522. doi: 10.1111/ppa.12965

Morrison, L. (2006). Multiple sequence alignment for phylogenetic purposes. *Australian Systematic Botany*, *19*(6), 479-539. doi: 10.1071/sb06020

Mount, D. (2001). *Bioinformatics sequence and genome analysis*. New York: Cold Spring Laboratory Press.

Nixon, K. C. (1999). The parsimony Ratchet, a new method for rapid parsimony analysis. *Cladistics*, *15*(4), 407-414. doi: 10.1111/j.1096-0031.1999.tb00277.x

Nixon, K. C. (2002). *WinClada* (Versión 1.0). Ithaca, Estados Unidos: autor.

Pivorano, W. & Heringa, J. (2008). Multiple sequence alignment. En J. Keith. (Ed.), *Bioinformatics, vol. I: data, sequence analysis, and evolution* (143-161). Berlín: Springer.

Poe, S. & Wiens, J. J. (2000). Character selection and the methodology of morphological phyogenetics. En J. J. Wien (Ed.), *Phylogenetic Analysis of Morphological Data* (20-36). Washington: Smithsonian Institution Press.

Posada, D. (2008). JModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution*, *25*(7), 1253-1256. doi: 10.1093/molbev/msn083

Prado, B. R., Pozo, C., Valdez-Moreno, M. & Hebert, P. D. N. (2011). Beyond the colours: Discovering hidden diversity in the Nymphalidae of the Yucatan Peninsula in Mexico through DNA barcoding. *PLoS ONE, 6*(11), e27776. doi: 10.1371/journal.pone.0027776

Richardson, A. O. & Palmer, J. D. (2006). Horizontal gene transfer in plants. *Journal of Experimental Botany*, *58*(1), 1-9. doi: 10.1093/jxb/erl148

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology, 61*(3), 539-542. doi: 10.1093/sysbio/sys029

Ronquist, F., van den Mark, P. & Huelsenbeck, J. P. (2009). Bayesian phylogenetics analysis using mrbayes. En P. Lemey, M. Salem & A. Vandamme (Ed.), *Phylogenetic Hanbook: A Practical Approach to Phylogenetic and Hypothesis Testing* (210-236). Oxford, Reino Unido: Cambridge University Press.

Russell, A., Samuel, R., Klejna, V., Barfuss, M. H. J., Rupp, B. & Chase, M. W. (2010). Reticulate evolution in diploid and tetraploid species of *Polystachya* (*Orchidaceae*) as shown by plastid DNA sequences and low-copy nuclear genes. *Annals of Botany*, *106*(1), 37-56. doi: 10.1093/aob/mcq092

Sanderson, M. J. & Hufford, L. (Eds.). (1996). *Homoplasy, the recurrence of similarity in evolution*. San Diego: Academic Press.

Schmidt, H. A. (2009). Testing tree topologies. En P. Lemey, M. Salemi & A. Vandamme (Eds.), *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, (381-396). Cambridge, Reino Unido: Cambridge University Press.

Simpson, M. G. (2019). *Plant systematics*. San Diego: Academic Press-Elsevier.

Smith, U. & Hendricks, J. (2013). Geometric morphometric character suites as phylogenetic data: Extracting phylogenetic signal from gastropod shells. *Systematic Biology*, *62*(3), 366-385. doi: 10.1093/sysbio/syt002

Soltis, D. E. & Soltis, P. S. (1999). Choosing an approach and an appropriate gene for phylogenetic analysis. En D. E. Soltis, P. S. Soltis & J. J. Doyle. (Ed.), *Molecular Systematics of Plants II-DNA Sequencing* (1-42). New York: Springer.

Soltis, P. S. & Soltis, D. E. (2009). The role of hybridization in plant speciation. *Annual Review of Plant Biology*, *60*(1), 561-588. doi: 10.1146/annurev.arplant.043008.092039

Stavrinides, J. & Guttman, D. S. (2004). Mosaic evolution of the severe acute respiratory syndrome Coronavirus. *Journal of Virology*, *78*(1), 76-82. doi: 10.1128/jvi.78.1.76-82.2004

Strimmer, K. & von Haeseler, A. (2009). Genetic distances and nucleotide substitution models. En P. Lemey, M. Salem & A. Vandamm. (Ed.), *The Phylogenetic Hanbook: A Practical Approach to Phylogenetic and Hypothesis Testing* (111-141). Cambridge, Reino Unido: Cambridge University Press.

Swofford, D. L. (2002). PAUP*. *Phylogenetic Analysis Using Parsimony (*and Other Methods)* (Version 4). Sunderland, Estados Unidos de América: Sinauer Associates.

Swofford, D. & Sullivan, J. (2009). Phylogeny inference based on parsimony and other methods using PAUP*. En P. Lemey, M. Salemi & A. Vandamme (Eds.), *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* (267-312). Cambridge, Reino Unido: Cambridge University Press.

Templeton, A. R. (2006). *Population genetics and microevolutionary theory*. Danvers, Estados Unidos de América: Wiley.

Turland, N. J., Wiersema, J. H., Barrie, F. R., Greuter, W., Hawksworth, D. L., Herendeen, P. S., Knapp, S., Kusber, W.-H., Li, D.-Z., Marhold, K., May, T. W., McNeill, J., Monro, A. M., Prado, J., Price, M. J. & Smith, G. F. (Eds.). (2018). *International Code of Nomenclature for algae, fungi, and plants (Shenzhen Code) adopted by*

the Nineteenth International Botanical Congress Shenzhen, China, July 2017. Regnum Vegetabile 159. Glashütten: Koeltz Botanical Books. doi: 10.12705/Code.2018

Via, S. & West, J. (2008). The genetic mosaic suggests a new role for hitchhiking in ecological speciation. *Molecular Ecology*, *17*(19), 4334-4345. doi: 10.1111/j.1365-294x.2008.03921.x

Wagner, G. P. (2001). *The Character Concept in Evolutionary Biology*. San Diego: Academic Press.

Whelan, S. (2008). Inferring trees. En Keith, J. M. (Ed.), *Bioinformatics, vol I: Data, sequence analysis, and evolution* (287-309). New York: Humana Press.

Wiley, E. O. & Lieberman, B. S. (2011). *Phylogenetics: Theory and Practice of Phylogenetic Systematics*. Hoboken, Estados Unidos de América: Willey-Blackwell.

Wu, G., Yao, Y., Qu, K, Ding Z., Li, H., Palanichamy M., Duan, Z., Li, N., Chen, Y. S. & Zhang, Y.-P. (2007). Population phylogenomic analysis of mitochondrial DNA in wild boars and domestic pigs revealed multiple domestication events in east Asia. *Genome Biology*, *8*(11), R245. doi: 10.1186/gb-2007-8-11-r245

"Mandala
(Capullus)"
Raku
16.5 x 25 x 25 cm
2014

"Estructura"
Cerámica/esmalte
42 x 30 x 13 cm
2011

"Flor cósmica"
Raku
19 x 14.5 x 19 cm
2016

Tequio *4*(11), 2021: 27-40
ISSN: 2594-0546

On some polytopes in phylogenetics
Politopos en filogenética

Linard Hoessly[1]

Fecha de recepción: 6 de noviembre de 2020
Fecha de aceptación: 23 de diciembre de 2020

Resumen - Presentamos las nociones matemáticas utilizadas en filogenética y tres clases de politopos de la filogenética. El Tight span y el politopo de Lipschitz se asocian a espacios métricos finitos y pueden conectarse a incrustaciones que conservan la distancia, mientras el politopo de evolución mínima balanceada (BME) se asocia con números naturales.

Palabras clave: Filogenética, politopo, espacio métrico finito, politopo fundamental, tight span.

Abstract - We introduce mathematical notions used in phylogenetics and three sorts of phylogenetics polytopes. The Tight span and the Lipschitz polytope are both associated to finite metric spaces and can be connected to distance-preserving embeddings, while the balanced minimum evolution (BME) polytope is associated to natural numbers.

Keywords: Phylogenetics, Polytope, Finite metric space, Fundamental polytope, Tight span.

1. Introduction

Phylogenetics studies the methods and the practice of identifying evolutionary relationships among biological species. Finding such relationships is a current focus of research, and is usually performed via phylogenetic inference based on mathematical models of evolution (Semple & Steel, 2003; Steel, 2016), which are represented as phylogenetic trees or networks (Huson, Rupp, & Scornavacca, 2010). Usually, genetic material is transferred from parents to offspring, resulting in tree-like representations. However, different biological species can transfer genetic information between otherwise unrelated organisms. Horizontal gene transfer e.g. is a mechanism where genetic material from one species is moved to another one which is relevant in how bacteria acquire antibiotic resistance. This suggests the possibility that corresponding parts of the evolutionary history might not be tree-like, and such relationships are often represented via phylogenetic networks. There are different approaches to phylogenetic reconstruction. We briefly introduce and elaborate on distance-based and likelihood-based methods. Distance-based techniques first compute a pairwise distance-like function between the taxa to construct a phylogenetic tree T (or structure) that best represents the distances obtained, usually via some optimality criterion. Distance-based methods are popular as they tend to be fast. Concerning likelihood-based methods there are two main paradigms: maximum likelihood (ML) and bayesian methods. In both, evolution is described through probabilistic model of sequence evolution, enabling in principle computations of likelihoods of observing the data given the model and its parameters. While these methods are assumed to be more correct from a foundational level, corresponding computations can be slow.

[1] Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark;
email: hoessly@math.ku.dk ORCID-ID: 0000-0002-2745-2141

Many fascinating objects originated from methods and structures used to understand the evolutionary history of species (Dress, Huber, Koolen, Moulton, & Spillner, 2011; Semple & Steel, 2003). We will first introduce objects from discrete mathematics and then focus on three polytopes which can be related to objects of interest in phylogenetics. Our treatment does not aim to be comprehensive in its scope, as these are fairly developed fields. In our exposition and treatment we mostly focus on phylogenetic trees, tree-like metric spaces and corresponding polytopes.

2. Introduction to some discrete objects

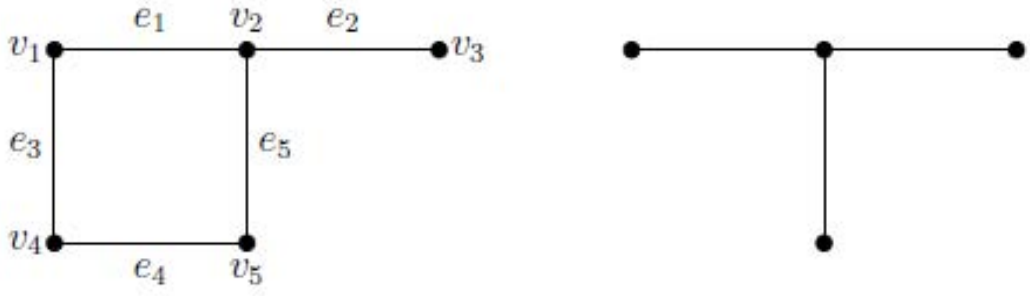
We introduce notions related to the combinatorics of phylogenetics, i.e. graphs in ξ 2.1, polytopes in ξ 2.2, finite metric spaces and splits in ξ 2.3 and phylogenetic trees in ξ 2.4.

2.1 Graphs and trees

In phylogenetics, graphs and trees are used to represent evolutionary relationships between species. We will focus on undirected graphs,[2] since this is our main setting of interest.

A finite undirected graph $G = (V, E)$ consists of vertices $V = V(G)$ and edges $E \subseteq V^2$, written as $E = E(G)$. A *path* is a sequence $e_0, e_1, \cdots, e_n$ of edges which join a sequence of distinct vertices. Graphs in which two arbitrary vertices are connected by exactly one path are called *trees*, as an example consider Figure 1. A graph G is *connected* if there is a path between any two vertices. The *degree* $deg(v)$ of a vertex $v \in V$ is the number of edges incident to v.

Figure 1.
Only the right graph is a tree.

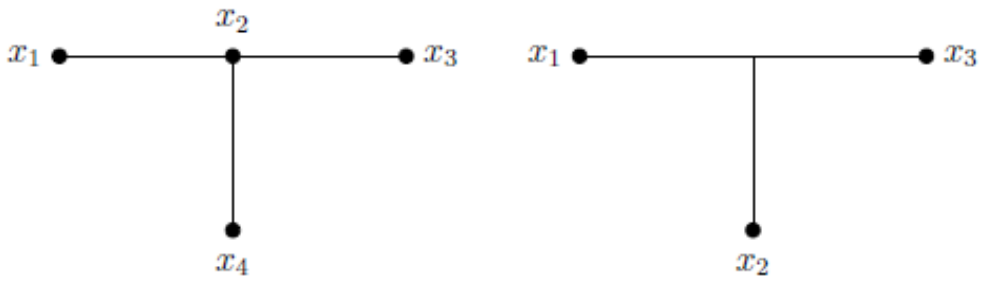


Lemma 2.1 ((Bollobas, 1998, equation (1), p.4 and ξ 1.2)) Let $G = (V, E)$ be a connected graph. Then $\sum_{v \in V} d\,eg(v) = 2|E|$, and furthermore $|E| = |V| - 1$ if and only if G is a tree.

Next we introduce a particular notion of a tree used in phylogenetics. Let X be a set. An X-tree is a labelling of some of the vertices of a tree T, where every leaf of T is labelled. We denote an X-tree by (T, ϕ), where $\phi: X \to V(T)$ is the labelling map. When all internal vertices are of degree 3, we call it binary X-tree.

[2] I.e. graphs where the edges are not directed.

Figure 2.
Two different binary X-trees on the tree T of figure 1.



2.2 Polytopes

Polytopes are seemingly simple geometric objects with flat sides. They appear as convex hulls of a finite set of points in Euclidean space (like, e.g., the plane $\mathbb{R}^2$ or 3-dimensional space $\mathbb{R}^3$), and exhibit a rich variety of combinatorial structures (Ziegler, 1995). The convex hull of a set of points $\{a_1, \cdots, a_m\} \subset \mathbb{R}^n$ is defined as

$$\operatorname{conv}\{a_1, \cdots, a_m\} := \{x \in \mathbb{R}^n \mid x = \sum_{i=1}^{m} \lambda_i a_i, \sum_{i=1}^{m} \lambda_i = 1, \lambda_i \geq 0\}$$

A *polytope* is a convex hull of a finite set of points. Well-known examples include two-dimensional polytopes that are convex polygons like the square (cf. Figure 3). The *dimension* of a polytope P is the dimension of the smallest Euclidean space which could contain it. As an example, the square of figure [fig_octa] has dimension two. A *face* of a polytope P is any intersection of the polytope with a half-space such that none of the interior points of the polytope lie on the boundary of the half-space. Any face of a polytope is a polytope itself. Some faces have a special name, faces of dimension $0,1$ and $\dim(P) - 1$ are called *vertices, edges and facets*. Moreover, the faces of polytopes can be ordered by inclusion, giving the poset of faces. A rougher invariant are its *face numbers* $f_0^P, \ldots, f_{\dim(P)}^P$, which are defined as

$$f_i^P = \#\{i - \text{dimensional faces of } P\}.$$

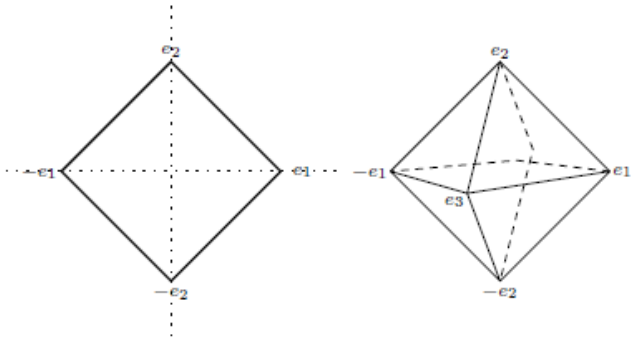
Putting all the face numbers together gives a convenient way of writing them as the so-called f-vector $(f_0^P, \ldots, f_{m-1}^P)$, where $m = \dim(P)$. Note that convex polytopes may equivalently be defined as an intersection of a finite number of half-spaces, corresponding to the so-called *hyperplane description*, see, e.g., (Ziegler, 1995, §2.4).

Example 2.2. Consider the *d-crosspolytope*, which is defined as

$$\beta_d := \operatorname{conv}\{e_1, -e_1, \cdots, e_d, -e_d\} \subseteq \mathbb{R}^d,$$

where e_1 is the unit vector with entry one in the first coordinate and zeros otherwise, i.e., e_i is the vector with the only nonzero entry one in the i-th[3] coordinate.

Figure 3.
A square (β_2) and an octahedron (β_3) with f-vectors (4; 4) and (6; 8; 8).



Another interesting class of polytopes are zonotopes, which are Minkowski sums of lines. Their combinatorial structure connects to hyperplane arrangements, tilings or oriented matroids (Ziegler, 1995, ξ 7). As an example consider the square of figure 3 as the sum of the lines $[e_1, e_2]$, $[e_1, -e_2]$.

2.3 Finite metric spaces and splits

Let X be a set. A metric (or distance function) on X is a symmetric function $d: X \times X \to \mathbb{R}_{\geq 0}$ such that
(1) For all $x, y \in X$, $d(x, y) = 0$ implies $x = y$.
(2) For all $x, y, z \in X$, $d(x, z) \leq d(x, y) + d(y, z)$ ("triangle inequality").
If condition (1) is dropped, then d is called a pseudometric. In the following we will focus on *finite metric spaces* with $|X| < \infty$.

Example 2.3 (Metric spaces from weighted graphs) A *weighting* of a graph G is any function $w: E(G) \to \mathbb{R}_{>0}$, and the pair (G, w) is called a *weighted graph*. Set

$$d_w(v, v') := \min\{w(e_1) + \cdots + w(e_k) \mid v, e_1, v_1 \ldots, e_k, v' \text{ is a path joining } v \text{ with } v'\}$$

such that the pair $(V(G), d_w)$ is a metric space.

If (G, w) represents (X, d) it is called a *graph realisation* of the metric space. Note that any finite metric space has a graph realisation from the complete graph[4] by setting the weight of the edge $e_{i,j}$ between i, j to $d(i, j)$. Next, we introduce metric spaces coming from X-trees.

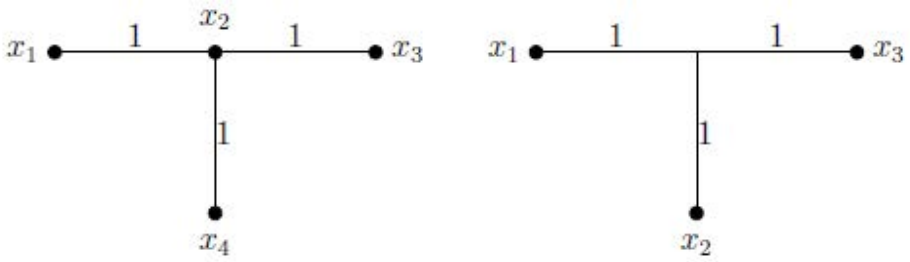
[3] I.e. i stands for any of the elements i∈{1,⋯,d}.
[4] The complete graph on a set of vertices is the graph where any two vertices are connected to each other through an edge.

Definition 2.4 (Tree-like metrics) *A (pseudo)metric d on a set X is called a* tree-like (pseudo)metric *if there exists an X-tree (T, ϕ) and a weighting w of T such that for all $x, y \in X$*

$$d(x, y) = d_w(\phi(x), \phi(y)).$$

The pseudometric d is a metric if and only if ϕ is injective.

Figure 4.
Two *X*-trees with edge weight one for each edge.



Next we consider splits. Let X be a finite set.

- A *split* of X is a bipartition of X, i.e., a pair of disjoint subsets $A, B \subseteq X$ such that the union[5] $A \cup B = X$, which is written as $A|B$.
- Two splits $A|B$ and $C|D$ are *compatible* if at least one of the four intersections[6] $A \cap C, A \cap D, B \cap C, B \cap D$ is empty.
- A *system of splits* on X is just a set of splits of X; the system is called [compatible]*compatible* if its elements are pairwise compatible.

There are more general definitions for split systems, e.g. weakly compatible or circular splits (Semple & Steel, 2003, x 3.8 or x 7.4). Next we consider weightings on splits.

Definition 2.5 A *weighted split system is a pair $(\mathcal{S}, \alpha)$ where $\mathcal{S}$ is a system of splits on X and $\alpha \in (\mathbb{R}_{\geq 0})^{\mathcal{S}}$ is any weighting. Any such weighted split system defines a nonnegative function $d_\alpha: X \times X \to \mathbb{R}$* via *$d_\alpha(x, y) = \sum_{\sigma \in \mathcal{S}} \alpha_\sigma \, \delta_\sigma(x, y)$ where δ_σ is defined for $\sigma = A|B$ as*

$$\delta_\sigma(i, j) = \begin{cases} 0 & i, j \in A \text{ or } i, j \in B \\ 1 & \text{otherwise.} \end{cases}$$

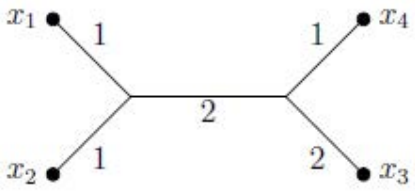
The functions of the form d_α are called *split-decomposable (pseudo)metrics* associated to $\mathcal{S}$, where (X, d_α) is a pseudometric space. A *positively weighted* split system is one where $\alpha_\sigma > 0$ for all $\sigma \in \mathcal{S}$.
For metric spaces from weighted trees we have the following.

[5] The union of two sets A,B which is denoted as A∪B is the set containing all the elements that are either in A or in B.
[6] The intersection of two sets A,C which is denoted A∩C is the set of all elements that are both in A and in C.

Theorem 2.6 ((Semple & Steel, 2003, Theorems 3.1.4, 7.1.8, 7.3.2)) *Let (X, d) be a pseudometric space. The following are equivalent:*
(i) d is a tree-like pseudo-metric on X (in the sense of Definition [df:tm2]).
(ii) d is a split-decomposable pseudometric associated to a positively weighted system of compatible splits. Moreover, this system is unique.
Under the equivalence of (I) with (II), splits in the decomposition of the metric correspond bijectively[7] to edges in the tree.

Example 2.7 *Consider the metric on $X = \{x_1, x_2, x_3, x_4\}$ given as follows*

$d(x_i, x_j)$	x_1	x_2	x_3	x_4
x_1	0	2	5	4
x_2	2	0	5	4
x_3	5	5	0	3
x_4	4	4	3	0



The metric is tree-like, where the underlying tree can be illustrated in the sense of Definition [df:tm2] as above. With Theorem [tree], the corresponding splits can be read off the graph leading to the decomposition of the distance as

- $x_1|x_2, x_3, x_4, \quad x_2|x_1, x_3, x_4, \quad x_3|x_1, x_2, x_4, \quad x_4|x_1, x_2, x_3, \quad x_1, x_2|x_3, x_4$
- $d(\cdot,\cdot) = \delta_{x_1|x_2,x_3,x_4} + \delta_{x_2|x_1,x_3,x_4} + \delta_{x_3|x_1,x_2,x_4} + \delta_{x_4|x_1,x_2,x_3} + 2 \cdot \delta_{x_1,x_2|x_3,x_4}$

Remark 2.8 *For a finite metric space (X, d), it is often of interest to obtain a decomposition of the metric into a sum of more elementary parts. One possible family of functions are the δ_σ from splits of Definition 2.5.*

Remark 2.9 *There is a more general theory for decompositions into weighted split systems. (Bandelt & Dress, 1992, Theorem 2) says that any metric (X, d) can be uniquely decomposed into $d = d_0 + \sum_{\sigma \in \mathcal{S}} \alpha_\sigma \delta_\sigma$, where d_0 is split prime and $\mathcal{S}$ is a (unique) weakly compatible system of splits.[8] Furthermore if in this decomposition $d_0 = 0$, then the metric is called totally split decomposable.*

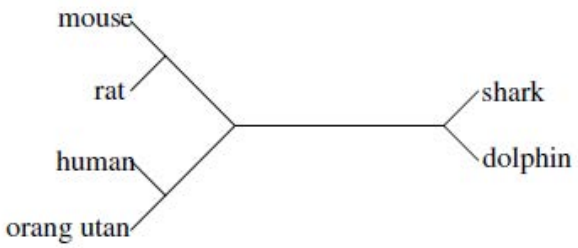
2.4 Phylogenetic trees

Phylogenetic trees describe evolutionary relationships, and we will mostly focus on undirected phylogenetic trees. However, both directed versions and networks are also used in phylogenetics, see, e.g., (Huson *et al.*, 2010).

[7] I.e. in a one-to-one relationship.
[8] In (Bandelt & Dress, 1992), a split prime metric is such that it is not further decomposable with respect to split metrics.

A *phylogenetic tree* is an X-tree where only the leaves are labelled and all internal vertices have a degree of at least 3. As in the case of X-trees, a phylogenetic tree is binary if all internal vertices have degree three. As a more concrete example of a binary phylogenetic tree consider

Figure 6.
A binary phylogenetic tree.



For $n = |X| \geq 3$ denote by $\mathcal{T}_n$ the set of all binary X-trees with n leaves. If the context is clear we will also simply say binary tree for binary phylogenetic X-trees.

Given an X-tree, there is an associated system of splits on X obtained by considering the two connected components obtained by the removal of e in T for each edge $e \in E(T)$. Denote the so-obtained set of splits by $\sum(T)$. For an example we refer to example [ex_dist]. On the other hand, by Theorem [tree], we get that for $\sum$ a system of splits, there is an X-tree T such that $\sum = \sum(T)$ if and only if the system of splits is compatible.

More general split systems are employed for generalizations of unrooted phylogenetic trees, where graphs in such split networks are not necessarily trees, and one or more edges in the graph are used to represent a split (Dress *et al.,* 2011; Huson *et al.,* 2010).

3. Polytopes in phylogenetics

Both the Tight span and the Lipschitz polytope are associated to a (finite) metric space and relate to a distance-preserving embedding in a bigger space. The minimum evolution polytope on the other hand is associated to natural numbers $n \in \mathbb{N}_{\geq 3}$. In the following, we aim to introduce and motivate the main objects. However, the topics are mature research directions and we restrict to a non-exhaustive treatment.

3.1 Tight span

Isbell studied the tight span in his investigation of injectivity for metric spaces (Isbell, 1964). In phylogenetics, it appeared in relation to reconstruction of phylogenetic trees from finite metric spaces (Dress, 1984). Representations of distances of phylogenetic trees can be seen as a connected one-dimensional polytope. Distances between vertices correspond to the sum of the edge lengths of the shortest paths. Hence it is natural to ask whether we can embed a given finite metric space distance-preserving into a low-dimensional compact polytope. One such possibility is the so-called Tight span.

The Tight span often helped to establish properties of classes of metrics, particularly in relation to decompositions that are of interest in phylogenetics. Furthermore, the 1-skeleton[9] of the Tight span is a graph realisations of the metric (Dress, 1984). For more on the motivation and connection of the study of Tight span

[9] I.e. the one dimensional faces.

to phylogenetics we refer to, e.g., (Dress *et al.,* 2011) or (Huson *et al.,* 2010), and for a concrete algorithmic application to, e.g., First we consider an unbounded polytope $U_{(X,d)} := \{z \in \mathbb{R}^X \mid z_i + z_j \geq d(i,j)\ \forall i,j \in X\}$.

Definition 3.1 The *Tight span* of (X,d) is given by the minimal points of $U_{(X,d)}$, which are defined as $T_{(X,d)} := \{z \in U_{(X,d)} \mid y \in U_{(X,d)} \text{ and } y \leq z \text{ implies } z = y\}$.

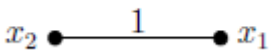
Note that the Tight span $T_{(X,d)}$ corresponds to the bounded faces of the polyhedron $U_{(X,d)}$. The Tight span is a polytopal complex that is associated to any finite metric space (X,d), whose structure often catches features of (X,d). The Kuratowski embedding is a map $f_{(X,d)}: X \to \mathbb{R}^X$ that sends elements of $X = \{x_1, \cdots, x_n\}$ to its Tight span, while preserving their pairwise distance. It is defined as

$$\begin{aligned} f_{(X,d)}: \quad & X \to \mathbb{R}^X \\ & x_i \mapsto f_{(X,d)}(x_i) := (d(x_i,x_j))_{j \in X} \end{aligned}$$

We have the following.

Lemma 3.2 The function $f_{(X,d)}: (X,d) \to (T_{(X,d)}, ||\cdot||_\infty)$ (where $T_{(X,d)} \subseteq \mathbb{R}^X$) is an isometric map into the tight span $T_{(X,d)}$, where for $z \in \mathbb{R}^X$, $||z||_\infty := \max_{x_i \in X}\{|z_i|\}$.

Example 3.3 *Consider the metric on $X = \{x_1, x_2\}$ with $d(x_1,x_2) = 1$. As a tree-like metric, it can be illustrated as*

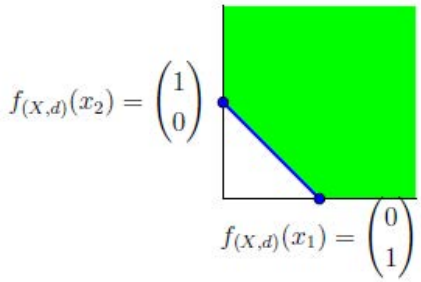
The upper map sends x_1 to 

$\begin{pmatrix} d(x_1,x_1) \\ d(x_1,x_2) \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, x_2 to $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$, hence

$$||f_{(X,d)}(x_1) - f_{(X,d)}(x_2)||_\infty = ||\begin{pmatrix} 1 \\ 0 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix}||_\infty = 1 = d(x_1,x_2),$$

and the Tight span looks as follows.

Figure 7.
The Tight-span as the blue line.



Example 3.3 generalises as follows to tree-like metric spaces, which can be classified via their Tight span

Theorem 3.4 *(Dress, 1984, Theorem 8)* The metric space (X, d) is tree-like if and only if the tight span $T_{(X,d)}$ is an $\mathbb{R}$-tree.[10]

This has been generalised to show that for (X, d) a finite metric that is totally decomposable, the Buneman graph D representing the split decomposition is contained in the 1-skeleton of the tight span $T_{(X,d)}$ (cf., i.e., (Huson *et al.,* 2010, x 5.12)). Injectivity of metrics corresponds to some factorisation property, where Isbell showed that $(T_{(X,d)}, ||\cdot||_\infty)$ is injective and that every metric can be isometrically embedded into this metric space (Isbell, 1964).

3.2 Lipschitz polytope

Studying fundamental polytopes was proposed by Vershik (Vershik, 2015) as an approach to a combinatorial classification of metric spaces. It also relates to an isometric embedding of the metric space, however, through optimal transport. As in the case of the Tight span it can be expected that properties of metric spaces can be connected to properties of the fundamental polytope.

The polar dual of the fundamental polytope consists of the real-valued functions with Lipschitz constant at most 1, called Lipschitz polytope. As polar duality preserves all combinatorial data, it is enough to classify the combinatorial structure of Lipschitz polytopes. We will mostly focus on Lipschitz polytopes in the following.

The structure of fundamental polytopes of tree-like metric spaces were studied via associated hyperplane arrangements and corresponding decompositions of the matroid in (Delucchi & Hoessly, 2020), enabling explicit formulas for face numbers of tree-like finite metric spaces. Values of the f-vectors as well as concrete values for f-vectors for "generic"[11] metrics were given in (Gordon & Petrov, 2017). For more on connections, terminology, history and further context around fundamental polytopes we refer to, e.g., (Ostrovska & Ostrovskii, 2019, ξ 1.6) or (Delucchi & Hoessly, 2020), where we further remark that direct applications to phylogenetics are still outstanding.

Definition 3.5 *The* Lipschitz polytope *of* (X, d) *is given as an intersection of halfspaces by*

$$LIP(X,d) := \left\{ x \in \mathbb{R}^X \mid \textstyle\sum_i x_i = 0,\ x_i - x_j \leq d(i,j)\ \forall i,j \in X \right\}. \quad (1)$$

Next we concentrate on the case of tree-like metric spaces and their Lipschitz polytopes as in (Delucchi & Hoessly, 2020). Let X be a finite set and consider a split $\sigma = A|B$ of X, where $|X| = n$. To σ we associate the

$$S_\sigma := \operatorname{conv}\left\{ \frac{|B|}{n} \cdot \mathbb{1}_A - \frac{|A|}{n} \cdot \mathbb{1}_B, \frac{|A|}{n} \cdot \mathbb{1}_B - \frac{|B|}{n} \cdot \mathbb{1}_A \right\} \subseteq \mathbb{R}^X$$

line segment (one-dimensional polytope)

:::

where Accordingly, associated to a split system $\mathcal{S}$ we define the zonotope defined by the Minkowski sum $Z(\mathcal{S}) := \sum_{\sigma \in \mathcal{S}} S_\sigma$.

[10] An R-tree (also called real trees) in some R^n corresponds to the points of a graph-theoretical embedding of a tree.
[11] In (Gordon & Petrov, 2017), finite metric spaces are called generic if the triangle inequality is always strict and the fundamental polytope is simplicial.

Then the form of Lipschitz polytopes of finite tree-like spaces can be given as follows.

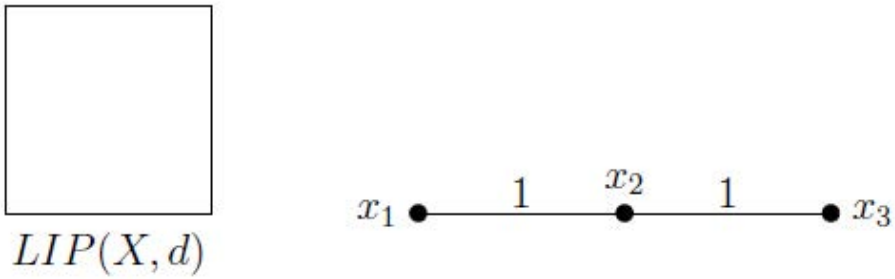
Theorem 3.6 *(Delucchi & Hoessly, 2020, Theorem 3.1) Let (X,d) be a tree-like pseudometric space. Then, $LIP(X,d) = \sum_{\sigma\in\mathcal{S}} \alpha_\sigma S_\sigma$ where $(\mathcal{S},\alpha)$ is the unique weighted system of compatible splits of X such that $d = d_\alpha$ (cf. Theorem 2.6).*

$$\mathbb{1}_A := \sum_{x\in A} \mathbb{1}_x.$$

We next go through an example.

Example 3.7 (Points in $\mathbb{R}^1$) *Distances defined by a set of n points in $\mathbb{R}^1$ come from a metric graph in a line. The associated set of splits from the split-metric are compatible, as such distances are tree-like. Consider the following metric on*
$X = \{x_1, x_2, x_3\}$.

Figure 8.
Lipschitz polytope as a square and graph realisation.



3.3 Minimal evolution polytope

The minimal evolution polytope (BME polytope) originates from the distance based approach to phylogenetic reconstruction. We will first give an intuitive description and then give the definition.

Assume we are given a distance function on the set of taxa, and we are looking for a corresponding phylogenetic representation. Assuming tree-likeness, we look for the best distance from a tree in order to represent the data at hand. One such method is the Balanced Minimum Evolution (BME) principle, that builds on a tree length calculation from (Pauplin, 2000) where the total tree length for phylogenetic trees can be computed via pairwise distances and the number of edges between the leaves. This is in contrast to simply summing all edge lengths in the tree.

Assume we are looking for a tree-like phylogenetic representation while only knowing distances obtained from data. Then, if the distance is from a tree, the correct tree topology minimises the total tree length. Applying this minimisation procedure is the BME method.

The tree with minimal tree length can be found by computing the tree lengths over all possible phylogenetic trees, or equivalently by minimizing over the BME polytope, which allows to reformulate the BME problem as a linear programming problem[12] (Haws, Hodge, & Yoshida, 2011).

While the BME method is just a heuristic, the Neighbor joining method[13] was shown to be a greedy version for the BME method (Gascuel & Steel, 2006).

The combinatorial structure of the BME polytope is of interest for the application in algorithms and as a basic mathematical object in phylogenetics. Some properties of the structure of the BME polytope are in (Eickmeyer, Huggins, Pachter, & Yoshida, 2008; Haws *et al.,* 2011), which were extended to the study of facets in (Forcey, Keefe, & Sands, 2016), whereas a direct algorithmic application is, e.g., in (Lefort, Desper, & Gascuel, 2015).

We represent distance functions $d: X \times X \to \mathbb{R}$ by a vector $D \in \mathbb{R}^{\binom{n}{2}}$, where we index entries of any $v \in \mathbb{R}^{\binom{n}{2}}$ by $\{i,j\} \subset X$ via lexicographic order, so we write v as $v = (v_{1,2}, v_{1,3}, \cdots, v_{n-1,n})$. For every labelled binary tree T on n vertices we consider the associated vector $w^T \in \mathbb{R}^{\binom{n}{2}}$ defined by the entries $w^T_{i,j} := 2^{n-l-2}$ where l is the number of interior nodes of the shortest path between i,j in T. Note that these vectors w^T depend only on the tree topology.

The balanced tree length estimation $l(T)$ of Pauplin is given by

$$l(T) := \sum_{i,j;i<j} w^T_{i,j}\, d(i,j).$$

Note that this is just the dot-product of the vector w^T and and the pairwise distances D, i.e. $l(T) = w^T \cdot D$. The BME principle aims at finding the tree T that minimises the above balanced tree length estimation. In (Haws *et al.,* 2011) they showed that minimising over all trees in $\mathcal{T}_n$ is equivalent to minimising over the convex hull of all the vectors w^T, where $T \in \mathcal{T}_n$.

BME polytopes are associated to natural numbers $n \in \mathbb{N}_{\geq 3}$, and not to distances (i.e. metric spaces) as in the case of the polytopes of ξ 3.1 and ξ 3.2. We define the BME(n) polytope as follows.

Definition 3.8 *The BME(n) polytope $\mathcal{P}_n$ for $n \geq 3$ is defined as*

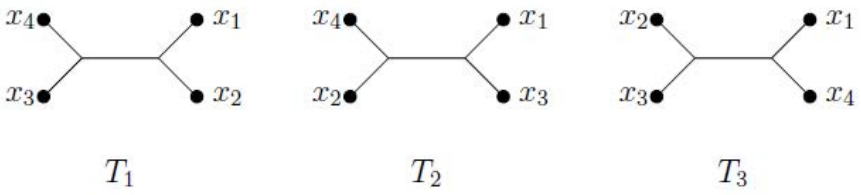
$$\mathcal{P}_n := \operatorname{conv}\{w^T \mid T \in \mathcal{T}_n\}.$$

As an example consider

Example 3.9 (*Eickmeyer et al., 2008*) *Consider the case $n = 4$. Then first we look at $\mathcal{T}_4$, which consists of the following binary trees:*

[12] Linear programming or LP is a method to find a maximum (or a minimum) of a linear objective function over a feasible region given by a convex polytope.
[13] A popular distance-based reconstruction method.

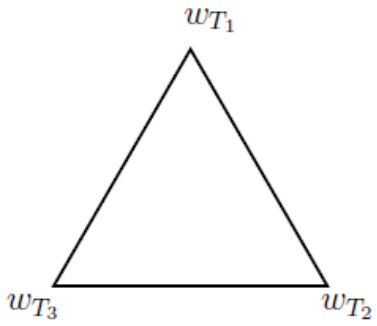
Figure 9.
The binary trees on X = fx1; x2; x3; x4g



The corresponding vectors $w_{T_i} \in \mathbb{R}^6$ *with coordinates lexicographic order have the form* $(w_{12}, w_{13}, w_{14}, w_{23}, w_{24}, w_{34})$ *and are given by*

$$w_{T_1} = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{2}), \quad w_{T_1} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}), \quad w_{T_1} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}).$$

Figure 10.
The BME(4) polytope is given by a triangle in $\mathbb{R}^6$



The BME(n) polytope $\mathcal{P}_n \subseteq \mathbb{R}^{\binom{n}{2}}$ has dimension $\binom{n}{2} - n$, as there are exactly n linear independent[14] equations obeyed by $\mathcal{P}_n$ (Eickmeyer *et al.,* 2008). Furthermore it has $(2n-5)!!$ vertices, which is $|\mathcal{T}_n|$, the cardinality of the set $\mathcal{T}_n$ (see, e.g., (Semple & Steel, 2003)). Some known results are summarized in the following table.

n	$dim(\mathcal{P}_n)$	# of vertices of $\mathcal{P}_n$	# of facets of $\mathcal{P}_n$
3	0	1	0
4	2	3	3
5	5	15	52
6	9	105	90262
n	$\binom{n}{2} - n$	$(2n-5)!!$	?

[14] A set of vectors is linearly independent if none of the vectors in the set can be defined as a linear combination of the others.

It is interesting to note that each NNI-move[15] on $\mathcal{T}_n$ corresponds to an edge of $\mathcal{P}_n$ (Haws *et al.,* 2011). Furthermore, partial results on facet inequalities exist, i.e., e.g. some facets from cherries (Forcey *et al.,* 2016) were characterised.

4. Conclusion and Outlook

We introduced notions from phylogenetics and mathematics that mostly relate to the distance-based approach to phylogenetic reconstruction. The three polytopes are associated to either distances or the number of species. While we focussed on tree-like metrics where tight span and fundamental polytope are well-understood, for more general classes of metrics we still have limited knowledge about their structure. The situation for BME polytopes is similar, where only small examples have complete characterisations. It will be interesting to see in what ways structural properties of the introduced objects relate to each other and to other notions from phylogenetics and mathematics in the future.

References

Bandelt, H.-J., & Dress, A. W. M. (1992). A canonical decomposition theory for metrics on a finite *set. Adv. Math.*, 92(1), 47-105.

Bollobas, B. (1998). *Modern graph theory* (corrected ed.). Heidelberg: Springer.

Delucchi, E., & Hoessly, L. (2020). Fundamental polytopes of metric trees via parallel connections of matroids. *European Journal of Combinatorics*, 87, 103098.

Dress, A. (1984). Trees, tight extensions of metric spaces, and the cohomological dimension of certain groups: A note on combinatorial properties of metric spaces. *Advances in Mathematics*, 53, 321-402.

Dress, A., Huber, K. T., Koolen, J., Moulton, V., & Spillner, A. (2011). *Basic phylogenetic combinatorics*. Cambridge University Press.

Eickmeyer, K., Huggins, P., Pachter, L., & Yoshida, R. (2008). On the optimality of the neighbor-joining algorithm. *Algorithms for Molecular Biology*, 3(1), 5.

Forcey, S., Keefe, L., & Sands, W. (2016). Facets of the balanced minimal evolution polytope. *Journal of Mathematical Biology*, *73*(2), 447-468.

Gascuel, O., & Steel, M. (2006). Neighbor-Joining Revealed. *Molecular Biology and Evolution*, 23(11), 1997-2000. Retrieved from https://doi.org/10.1093/molbev/msl072

Gordon, J., & Petrov, F. (2017). Combinatorics of the Lipschitz polytope. *Arnold Math. J.*, 3(2), 205–218.

Haws, D. C., Hodge, T. L., & Yoshida, R. (2011). Optimality of the neighbor joining algorithm and faces of the balanced minimum evolution polytope. *Bulletin of Mathematical Biology*, 73(11), 2627–2648.

Huson, D. H., Rupp, R., & Scornavacca, C. (2010). *Phylogenetic networks: Concepts, algorithms and applications*. Cambridge University Press.

Isbell, J. R. (1964). Six theorems about injective metric spaces. *Commentarii Mathematici Helvetici*, 39(1), 65–76.

Lefort, V., Desper, R., & Gascuel, O. (2015). FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Molecular Biology and Evolution*, 32(10), 2798-2800.

[15] A nearest-neighbour interchange (NNI) move on a phylogenetic tree rearranges the tree. Such moves are e.g. used in algorithms in order to optimise over the set of trees.

Ostrovska, S., & Ostrovskii, M. (2019). Generalized transportation cost spaces. *arXiv*.

Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 51(1), 41-47.

Semple, C., & Steel, M. (2003). *Phylogenetics*. Oxford University Press.

Steel, M. (2016). Phylogeny: *Discrete and random processes in evolution*. Society for Industrial and Applied Mathematics.

Vershik, A. M. (2015). Classification of finite metric spaces and combinatorics of convex polytopes. *Arnold Math. J.*, 1(1), 75-81.

Ziegler, G. M. (1995). *Lectures on polytopes* (Vol. 152). Springer-Verlag, New York.

"Maíz MON 810"
Papel de algodón
30 x 13 x 7 cm
2014

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.

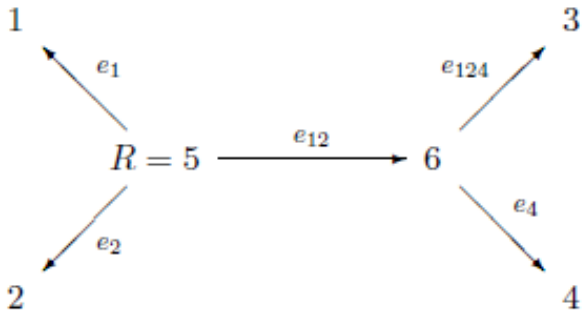
Para establecer dicha relación es necesario construir dos matrices espectrales: la matriz "Espectro de longitud de borde" y la matriz "Espectro de secuencia espectral". La primera incluye los pesos del árbol filogenético; la segunda contiene las probabilidades de todos los patrones de sustitución asociados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

Tequio *4*(11), 2021: 53-68
ISSN: 2594-0546

Indagando aspectos evolutivos con filogenias: reloj molecular y otras técnicas útiles en biología comparada

Researching evolutionary aspects with phylogenies: molecular clock and other useful techniques in comparative biology

Carlos Luis Leopardi-Verde[1] y Guadalupe Jeanett Escobedo-Sarti[1,2]

Fecha de recepción: 4 de diciembre de 2020
Fecha de aceptación: 6 de enero de 2021

Resumen - El primer paso para los estudios evolutivos suele ser establecer una hipótesis de relaciones entre los miembros del grupo de estudio. Luego de esto hay una amplia gama de posibilidades, según el interés del investigador. Esta contribución presenta las generalidades de algunas de las metodologías más utilizadas en los estudios macroevolutivos: el reloj molecular y la reconstrucción de caracteres ancestrales. También se ofrece información sobre otras técnicas útiles para estudios comparados que utilizan un marco de trabajo filogenético, como son la estimación de señal filogenética, los contrastes independientes, la descomposición ortonormal y el análisis de componentes principales filogenético.

Palabras clave: Evolución de caracteres, señal filogenética, macroevolución, máxima parsimonia, máxima verosimilitud.

Abstract - The first step for evolutionary studies is usually to establish a hypothesis of relationships between members of the study group. After this, there is a wide range of possibilities depending on the researcher's interest. This contribution presents the generalities of some of the methodologies most commonly used in macroevolutionary studies: the molecular clock and the reconstruction of ancestral characters. Information on other useful techniques for comparative studies that use a phylogenetic framework is also presented, such as phylogenetic signal estimation, independent contrasts, orthonormal decomposition, and phylogenetic principal component analysis.

Keywords: Character evolution, phylogenetic signal, macroevolution, maximum parsimony, maximum likelihood.

[1] Facultad de Ciencias Biológicas y Agropecuarias, Universidad de Colima. Km. 40 Autopista Colima-Manzanillo, crucero de Tecomán, Tecomán, Colima, México, C.P. 28930. *email: cleopardi@ucol.mx ORCID: https://orcid.org/0000-0001-5172-5114
[2] ORCID: https://orcid.org/0000-0002-4901-971X

Introducción

Las comparaciones en temas de evolución, entre otras cosas, procuran elucidar los patrones que ha seguido la vida a través de su historia (Harvey & Pagel, 1991). La base para este tipo de investigaciones es el establecimiento de una hipótesis de ancestría entre los taxa investigados, una filogenia. Sin embargo, esto es sólo el inicio, pues a partir de la hipótesis, y con el uso de metodologías auxiliares, se pueden plantear preguntas más profundas o fijar metas más ambiciosas, como pueden ser generar un marco temporal de trabajo (p. e. Sauquet *et al.,* 2017; Schneider *et al.,* 2004).

Incluso es posible ir más allá e investigar cómo ha sido la evolución de un carácter (p. e. Gómez-Acevedo, Rico-Arce, Delgado-Salinas, Magallón & Eguiarte, 2010; Silvera, Santiago, Cushman & Winter, 2009), qué relación tiene dicha evolución con variaciones climáticas o lo que se desee (p. e. Evans, Smith, Flynn & Donoghue, 2009; Yesson & Culham, 2006). También se pueden abordar temas de investigación que planteen elucidar si hay varios caracteres que están evolucionando de manera coordinada (Adams, 2010; Polly, Lawing, Fabre & Goswami, 2013; Serb, Alejandrino, Otárola-Castillo & Adams, 2011), entre muchas más alternativas. Por eso, en esta contribución se presentan los aspectos generales de algunas de las metodologías complementarias en el estudio de la evolución: el reloj molecular y la reconstrucción de estados ancestrales de caracteres. Del mismo modo, se incluyen algunos de los métodos que se han desarrollado recientemente para la búsqueda de patrones evolutivos, como los contrastes independientes, el análisis filogenético de componentes principales y la descomposición ortonormal. En otro manuscrito se exponen las generalidades acerca de las filogenias y los métodos para construirlas.

El reloj molecular

La idea original del reloj molecular supone que los cambios en el ADN (las mutaciones) se acumulan a una tasa aproximadamente constante en el tiempo evolutivo (Battistuzzi, Filipski & Kumar, 2011). La propuesta de que la evolución ocurre a una tasa aproximadamente constante a lo largo del tiempo fue planteada a principios de los años 60, de manera independiente por Margoliash (1963), utilizando la enzima Citocromo C, y por Zuckerkandl & Pauling (1965) usando globinas.

Motoo Kimura propuso que en el ADN hay mutaciones que pueden tener un efecto funcional ventajoso, deletéreo (mortal) o neutral para la población; por ello, desde un punto de vista selectivo, las mutaciones no tienen el mismo comportamiento. Aquellas mutaciones con efectos deletéreos son eliminadas rápidamente por selección, mientras que las que son ventajosas pueden perderse por deriva génica o ser fijadas por selección. Las mutaciones que no tienen un impacto funcional (las neutrales o casi neutrales), al estar libres de la acción de la selección pueden variar libremente (incrementando o disminuyendo su frecuencia) por la deriva génica. Según la teoría de Kimura, aunque muchas de estas mutaciones neutrales se pierden, también muchas otras se fijan en el genoma y por lo mismo tienden a ser las que más influyen en la tasa de evolución (Herron & Freeman, 2014; Ohta, 2013).

La velocidad a la que evoluciona cada región del ADN no es igual. Aquellas regiones que tienen fuertes presiones selectivas, como por ejemplo la que codifica para la enzima que fija el carbono en el ciclo de Calvin, la

ribulosa-1,5-bifosfato-carboxilasa/oxigenasa (RuBisCo), tenderán a acumular pocas mutaciones detectables; mientras las regiones que suelen tener menores presiones selectivas, como los espaciadores intergénicos (ITS, ETS) pueden acumular grandes cantidades de cambios en menor tiempo (Soltis & Soltis, 1999). Por ello, como la tasa de evolución no es constante a lo largo de todo el genoma y entre todos los grupos, el reloj molecular ha evolucionado y se ha diversificado.

De manera general, para fechar un evento de divergencia con los métodos modernos de reloj molecular se necesita: (i) obtener las distancias genéticas entre los taxa bajo análisis; para uno o varios de ellos debe existir un estimado de la edad aproximada, información que normalmente se obtiene de fósiles; (ii) se debe calcular la tasa de sustitución; (iii) se debe utilizar la tasa calculada en el paso anterior para convertir las distancias genéticas entre los taxa de interés en estimados de sus edades (Renner, 2005). Las sustituciones son mutaciones del ADN en las que una base es remplazada por otra diferente; a la velocidad con la que ocurre este proceso se le denomina tasa de sustitución y se calcula en función del modelo de sustitución seleccionado (Vandamme, 2009).

Como cualquier técnica, el reloj molecular no es perfecto. Según Sanderson, Thorne, Wikström & Bremer (2004), Renner (2005) y Rutschmann (2006), algunas fuentes de error que pueden incidir en los resultados son: (i) especificaciones erróneas de las distancias genéticas; (ii) topologías incorrectas; (iii) modelo inadecuado de evolución; (iv) uso de pocos puntos de calibración; (v) uso del tipo de reloj incorrecto; (vi) presencia de parálogos que pueden introducir sesgo en la reconstrucción de las relaciones históricas. Los parálogos son genes relacionados vía duplicación, pero que normalmente evolucionan de manera independiente entre sí y por lo tanto no son útiles para rehacer la historia de un linaje. En oposición existen los ortólogos, que son aquellas copias de un gen que se originan de un único gen ancestral en el antepasado común de los genomas comparados y cuya evolución se mantiene vinculada, por lo que permiten establecer hipótesis de relaciones ancestro-descendiente entre linajes (Koonin, 2005).

Entre todos los elementos previos, uno de los pasos clave en el fechaje de las relaciones ancestro-descendiente es seleccionar el tipo más apropiado de reloj molecular. Existen dos tipos básicos: los que suponen una tasa global de sustitución (reloj molecular estricto) y los que asumen heterogeneidad en las tasas de substitución (reloj molecular relajado). En el Cuadro 1 se presenta una comparación de los principales métodos, agrupados en función del tipo de datos de entrada y el método de optimización. En Welch & Bromham (2005), así como en Rutschmann (2006) hay otros métodos que no fueron incluidos aquí.

Cuadro 1.

Comparación de los diferentes métodos utilizados para fechar con reloj molecular utilizando como base el tipo de datos de entrada. Las técnicas se organizaron conforme a los datos de entrada y el método de optimización.

DATOS DE ENTRADA	MÉTODO DE OPTIMIZACIÓN	NOMBRE DEL MÉTODO Y AUTORES	OBSERVACIONES
Matriz de distancia		Regresión lineal (Nei, 1987; Li & Graur, 1991)	Presupone una tasa fija de evolución, por lo que es útil sólo si se cumple la hipótesis de reloj molecular estricto.

DATOS DE ENTRADA	MÉTODO DE OPTIMIZACIÓN	NOMBRE DEL MÉTODO Y AUTORES	OBSERVACIONES
Filograma		Media de la longitud del camino (Bremer & Gustafsson, 1997)	El algoritmo original sólo permite un punto de calibración ubicado en la raíz del árbol. La generalización de este método (PATHd8) permite el uso de múltiples puntos de calibración y la variación de las tasas de sustitución entre los nodos. Por la forma como trabaja es útil en filogenias grandes. Está implementado en un software del mismo nombre PATHd8 (Britton, Anderson, Jacquet, Lundqvist & Bremer, 2007).
Secuencias + topología	Máxima verosimilitud (ML).	Hay varios, algunos de los más populares son: Suavizado de tasa no paramétrico (NPRS por sus siglas en inglés) (Sanderson, 1997), verosimilitud penalizada (PL, por sus siglas en inglés) (Sanderson, 2002)	En estos métodos se utiliza un suavizado estadístico y la autocorrelación para determinar las edades de los nodos. Están implementados en el software r8s (Sanderson, 2003).
	Máxima verosimilitud (ML).	ML con reloj (Felsenstein, 1981)	Es un método lento, pero popular, puede ser muy eficiente si cuenta con una topología idónea y la longitud de ramas es correcta. Está implementado en software como PAUP*, PhyML, entre otros.
	Inferencia bayesiana (IB)	Estimación bayesiana con reloj relajados (Aris-Brosou & Yang, 2002).	Aunque no es un método popular, es interesante porque conjuga tasas de autocorrelación (como el PL) con un modelo explícito de especiación y extinción de linajes, algo que es desarrollado mejor en otros algoritmos. Está implementado en el software PhyBayes (Aris-Brosou & Yang, 2002).
Secuencias de ADN	Análisis bayesianos	Hay varios: reloj molecular local (Drummond *et al.,* 2006), así como los modelos de tasa variable (Suchard *et al.,* 2018).	En estos relojes la filogenia es calculada de manera simultánea con el reloj, por lo que se disminuyen los errores asociados a la topología del árbol. En general son métodos flexibles en los que se implementa una amplia variedad de modelos de evolución. Están implementados en el programa BEAST y actualmente son los más utilizados.

Modificado y actualizado de Rutschmann, 2006.

Entre todos los métodos que se presentan en el Cuadro 1, el más simple y de alguna manera el que expresa la idea más elemental del reloj molecular es el método basado en la matriz de distancias genéticas, el cual ocupa como algoritmo para estimar los tiempos de divergencia una regresión lineal y la idea subyacente es que si el árbol es ultramétrico, las distancias nodales pueden ser convertidas fácilmente en tiempos de divergencia (Rutschmann, 2006). Debido a que estos cálculos pueden hacerse con cualquier software estadístico, esta aproximación –que fue una de las primeras– gozó de gran aceptación.

También se advertirá en el Cuadro 1 que las técnicas para la estimación del reloj molecular se han vuelto cada vez más complejas y demandantes de equipos de cómputo. Esto está relacionado con el hecho de que en la mayoría de las matrices con datos moleculares lo común es la heterogeneidad de las tasas de evolución entre los linajes incluidos (Lartillot, Phillips & Ronquist, 2016).

Así, de acuerdo con Rutschmann (2006), si el grupo de organismos con el que se está trabajando tiene tasas heterogéneas de evolución, no es conveniente utilizar métodos basados en el reloj estricto (como la regresión lineal, Cuadro 1), sino que es más adecuado el uso de métodos que corrigen o incorporan la heterogeneidad de las tasas de sustitución a los cálculos. De hecho, este tipo de métodos actualmente son los que han cobrado mayor fuerza y se están desarrollando muy activamente.

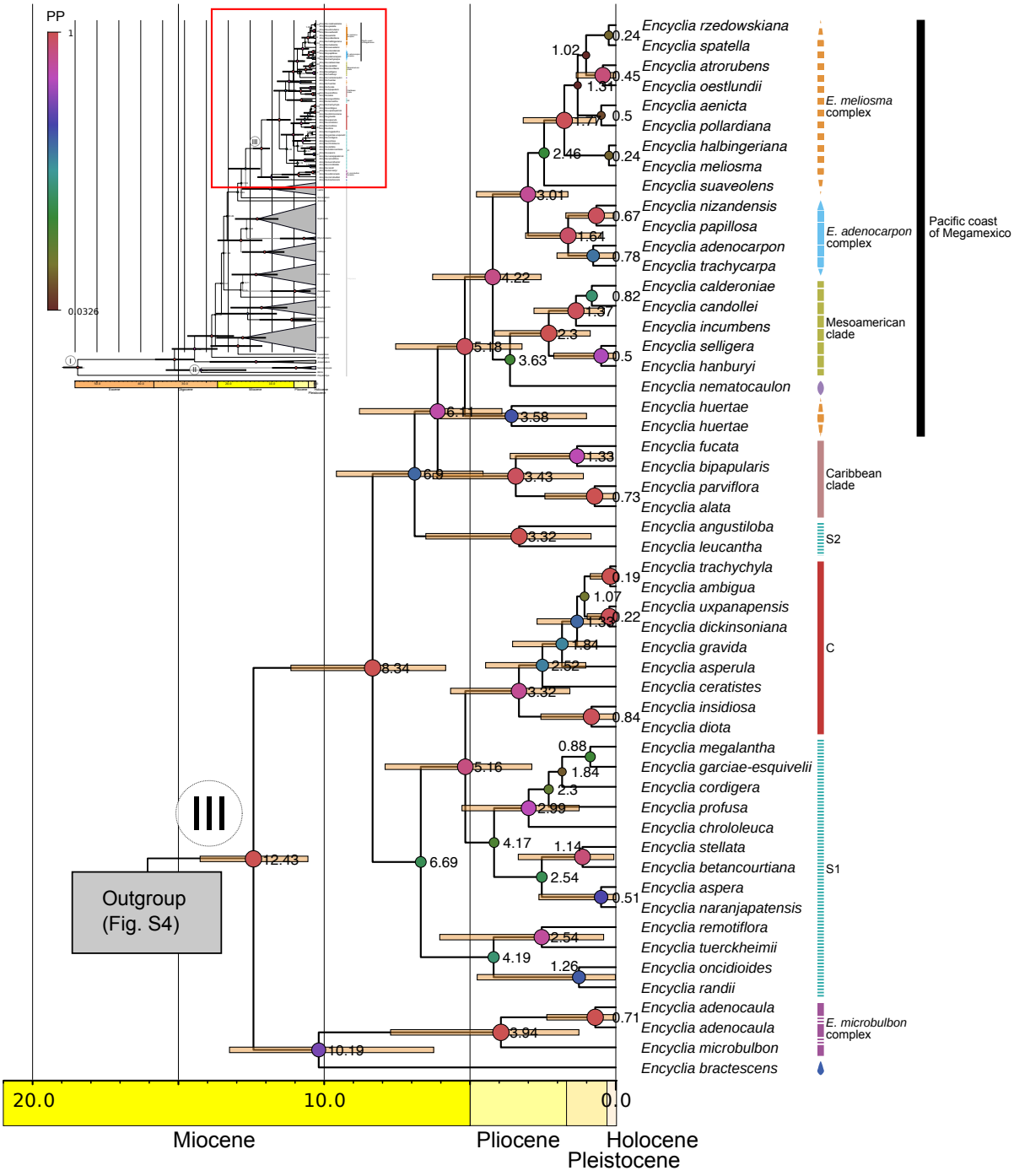
Los nuevos métodos acomodan la heterogeneidad basándose en especificaciones que indican cómo las tasas cambian entre los linajes (Drummond, Ho, Phillips & Rambaut, 2006). Una de las maneras de lograr esto es a través de las tasas autocorrelación temporal; otra es con el uso de distribuciones compuestas de Poisson (Rutschmann, 2006). La primera aproximación es más popular que la segunda en el diseño de los métodos; sin embargo, la segunda, al estar enmascarada en los distintos algoritmos de fechaje incluidos en el programa "Bayesian Evolutionary Analysis Sampling Trees" (BEAST), ha crecido mucho en uso. Un elemento común a todos los métodos de fechaje que no asumen una tasa de cambio constante es que intentan minimizar las discrepancias entre las longitudes de rama y las tasas de cambio en las ramas (Rutschmann, 2006).

En los métodos mencionados en el Cuadro 1, los que emplean tasas variables normalmente se asocian a una implementación que utiliza para las estimaciones inferencia bayesiana y el método de Monte Carlo, basado en cadenas de Markov (MCMC). En estos métodos se calculan las probabilidades posteriores junto con las tasas y tiempos (Drummond & Suchard, 2010). Las innovaciones de este método con respecto a los previos incluyen características como la capacidad de estimación de parámetros (no necesariamente todos deben ser no especificados), la correlación entre las tasas de ramas adyacentes puede ser probada; además, no se requiere una topología previa, lo que permite manejar la incertidumbre filogenética. Por otro lado, sobre esta base se han desarrollado variantes para incrementar la flexibilidad en el análisis: acorde con las especificaciones que se proporcione al modelo se puede probar sobre cada rama del árbol un reloj molecular estricto, relajado e incluso uno mixto (Drummond & Suchard, 2010; Lartillot *et al.,* 2016).

El reloj molecular tiene múltiples aplicaciones que van desde el fechaje de eventos, como por ejemplo la aparición de las angiospermas u otros linajes (Figura 1) (p. e. Barba-Montoya, dos Reis, Schneider, Donoghue & Yang, 2018; Leopardi-Verde, Carnevali & Romero-González, 2017; Magallón, Gómez-Acevedo, Sánchez-Reyes & Hernández-Hernández, 2015), estudios de evolución de caracteres de diversa índole (Schneider *et al.,* 2004; Schultz & Brady, 2008), estudios biogeográficos (Renner, 2005; Sanderson *et al.,* 2004), entre otros. Si bien es necesario tener en cuenta que ninguno de estos métodos es perfecto, no cabe duda de que permiten inferir sucesos del pasado y que tienen la capacidad de ayudar a entender la historia.

Figura 1.

Extracto de un cronograma de los Laeliinae (Orchidaceae) con énfasis en Encyclia Hook. El número romano indica uno de los puntos de calibración (III, MRCA Encyclia). Los números en los nodos señalan las edades estimadas en millones de años, las barras muestran el intervalo de confianza de 95 % y los círculos de color indican la probabilidad posterior de acuerdo con la escala que se encuentra en la esquina superior derecha. Tomado de Leopardi-Verde *et al.,* 2017.



Evolución de caracteres

Un carácter o atributo puede definirse como cualquier diferencia entre dos grupos de organismos, que puede ser utilizada para "caracterizar" o distinguirlos (Wagner, 2001). Este concepto es intuitivo a cualquiera de las actividades de la vida cotidiana y es una parte esencial de la biología comparada, que tiene por función analizar y capturar los patrones biológicos y elaborar teorías sobre los procesos que podrían explicarlos (Eldredge & Cracraft, 1980).

Una herramienta importante para comprender la evolución de un grupo cualquiera de organismos es estudiar la evolución de aquellos caracteres que podrían ser innovaciones clave o que se suponga que puedan tener alguna importancia evolutiva. La evolución de un carácter es el proceso por el cual un atributo evoluciona a lo largo de las ramas de una filogenia (Gupta, Maňuch, Stacho & Zhu, 2004).

Actualmente hay tres vías para reconstruir la historia evolutiva de un carácter. La primera es con el uso de máxima parsimonia, la cual sugiere que la mejor explicación de los datos es la que involucra la menor cantidad de cambios evolutivos (Swofford & Sullivan, 2009). Como características positivas pueden mencionarse el que suele ser una buena aproximación a los estados ancestrales, es intuitivo, está implementado en un número importante de programas y trabaja con caracteres polimórficos; no obstante, funciona bien sólo a tasas bajas de evolución, subestima el número total de cambios y no ofrece información sobre la incertidumbre de los procesos de transición de estados de carácter. En la Figura 2 se muestra un ejemplo de una reconstrucción de caracteres ancestrales utilizando como modelo sistemas sexuales en la familia *Bromeliaceae*, note que las ramas tienen sólo el color que corresponde al estado más parsimonioso. Este método está implementado en el programa Mesquite (Maddison & Maddison, 2019) (ver Figura 2).

La segunda alternativa es con el uso de métodos probabilísticos, que en general asumen que los cambios entre los estados de carácter siguen un proceso markoviano, como el descrito por Lewis (2001). Entre los aspectos positivos de este modelo destaca el que la probabilidad de cambio en cualquier punto a lo largo de alguna rama en un árbol filogenético depende sólo del estado de carácter actual y no de estados anteriores, se pueden asignar o calcular la tasa de cambio de un carácter y combinarla con información como la longitud de ramas.

La mayor debilidad de estos modelos es que dan por sentado que las longitudes de rama del árbol filogenético y la topología son conocidas con certeza (Paradis, 2012). En la Figura 3 se muestra el ejemplo de una reconstrucción de caracteres ancestrales utilizando un método probabilístico con los mismos datos de sistemas sexuales en la familia *Bromeliaceae*, note que en los nodos se incluye un diagrama de pastel en el que se representa la probabilidad de cada estado de carácter en cada nodo y por lo mismo ofrece una medida de la incertidumbre en la reconstrucción de los estados ancestrales. Si se comparan las Figuras 2 y 3 se observará que, aunque reflejan un patrón similar, al contar con una medida de incertidumbre en la reconstrucción de los estados ancestrales la Figura 3 brinda un panorama más completo para hacer la interpretación del resultado.

Figura 2.

Filogenia de la familia *Bromeliaceae* sobre la que se reconstruye la evolución del síndrome de polinización utilizando máxima parsimonia. Figura elaborada en Mesquite con datos de Escobedo-Sarti *et al.* (2013) y Aguilar-Rodríguez *et al.* (2019).

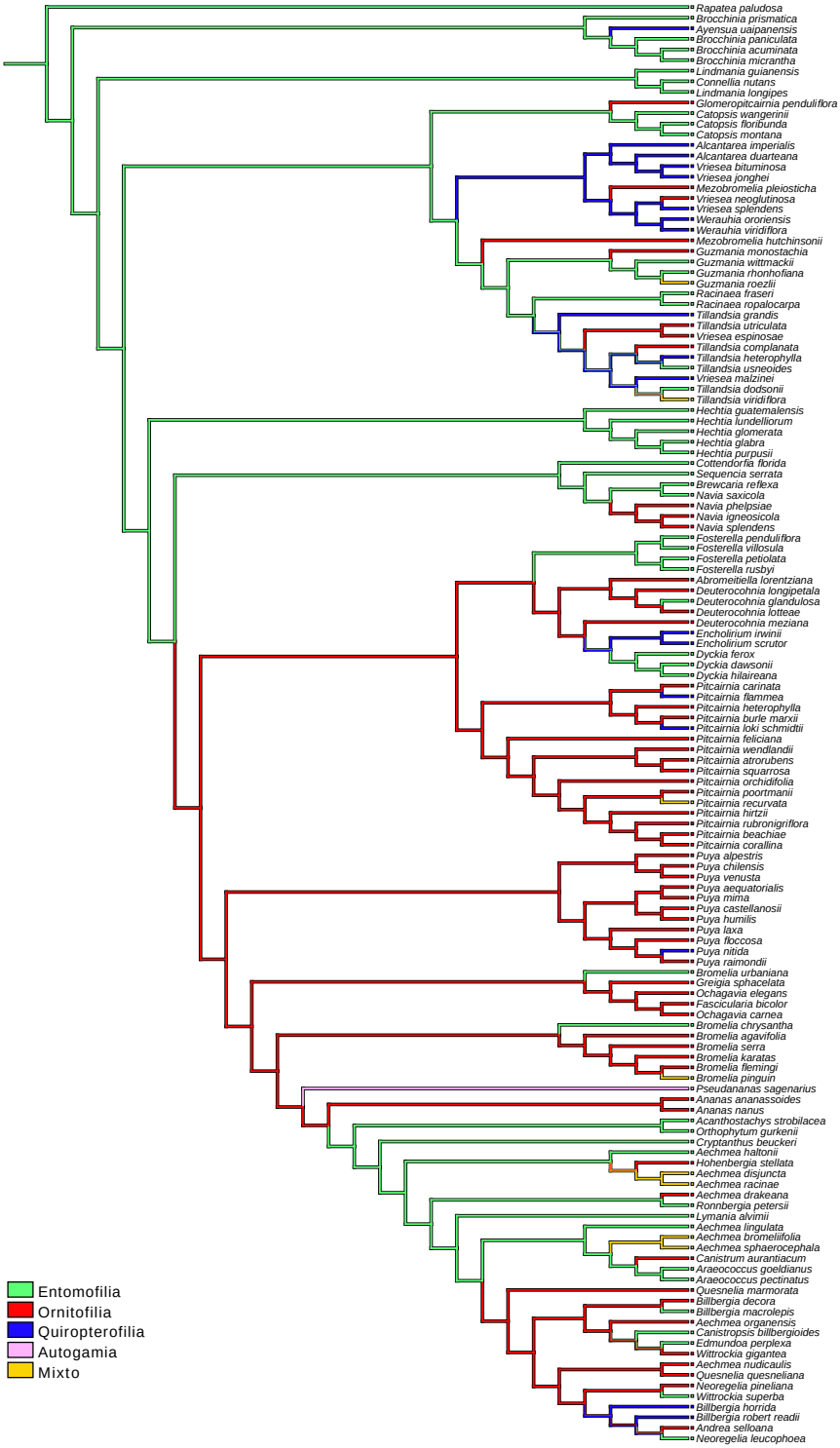
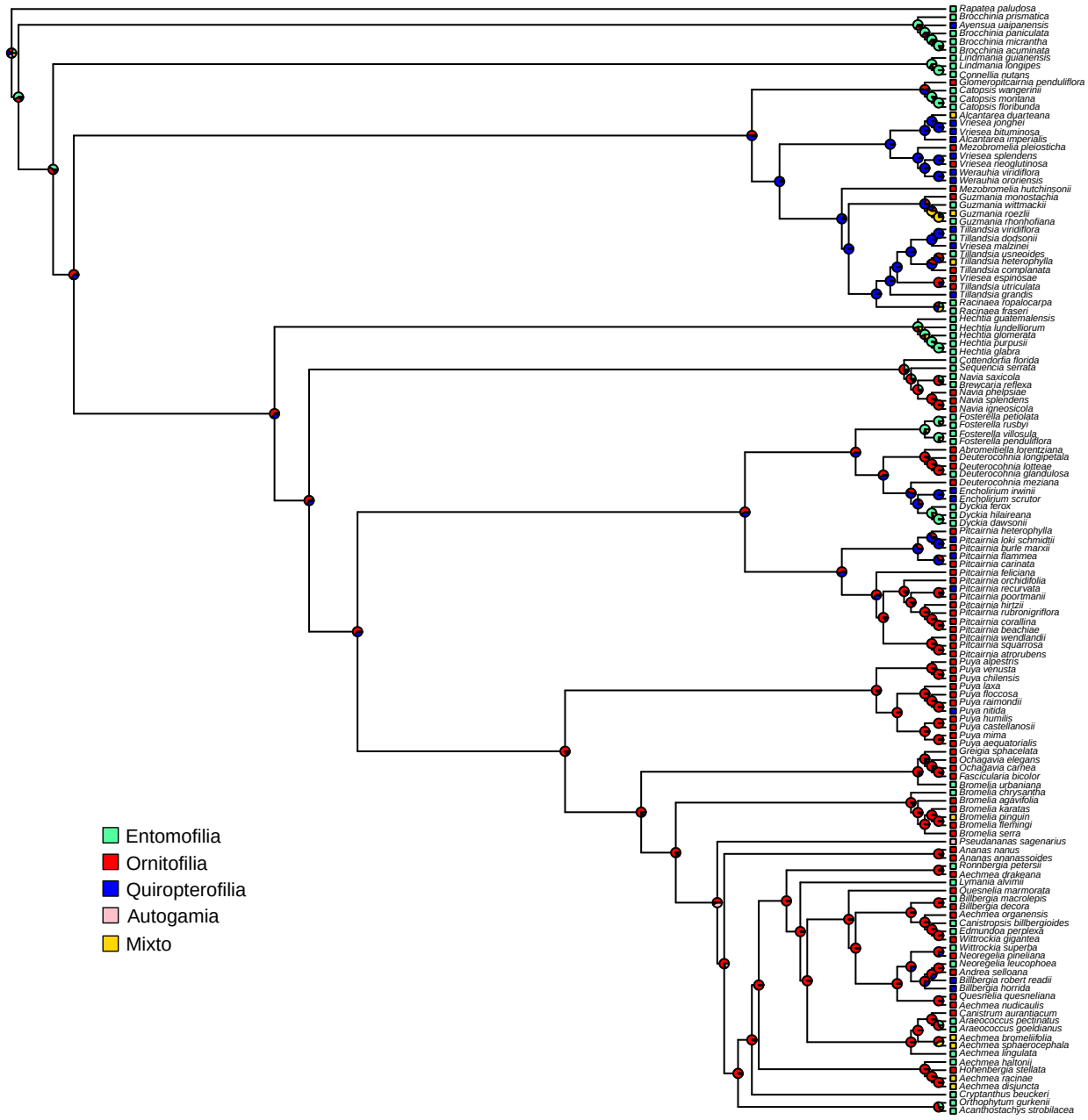


Figura 3.

Filogenia de la familia *Bromeliaceae* sobre la que se reconstruye la evolución del síndrome de polinización utilizando máxima parsimonia. Figura elaborada en Mesquite con datos de Escobedo-Sarti *et al.* (2013) y Aguilar-Rodríguez *et al.* (2019).



En los modelos probabilísticos hay dos implementaciones básicas: Mk1 y AsymMk. El Mk1 es un modelo que aplica sólo para caracteres discretos y puede trabajar con caracteres polimórficos. En este modelo todas las transiciones son igualmente probables y sólo tiene un parámetro: la tasa de cambio (Maddison & Maddison, 2019; Paradis, 2012).

El modelo AsymmMk es una modificación del Mk1 que incluye un parámetro de aleatoriedad. Este modelo sólo puede ser aplicado a caracteres continuos, aunque puede trabajar con caracteres binarios. Se caracteriza por sus tasas de cambio diferenciales: una hacia adelante y otra para las reversiones. Tanto el Mk1 como el AsymmMk están implementados en los programas Mesquite y en paquetes de R (Harmon, Weir, Brock, Glor & Challenger, 2008; Maddison & Maddison, 2019; Paradis, Claude & Strimmer, 2004; R Core Team, 2020).

Adicionalmente, se ha propuesto el uso de métodos bayesianos para reconstruir la historia de un carácter; la mayor diferencia con los modelos de máxima probabilidad es que se calculan las probabilidades posteriores para la transición de los caracteres. La implementación más conocida de este método está en BayesTraits.[3]

La tercera alternativa para hacer una reconstrucción de caracteres es con el uso del mapeo estocástico, que puede considerarse como una representación de la historia del carácter en la filogenia (Bollback, 2006). La principal característica de este método es que no considera fijo un estado a lo largo de una rama; por el contrario, explora las historias posibles, lo que permite tener una hipótesis de los cambios a lo largo de las ramas. Por el momento este método únicamente se puede utilizar con caracteres discretos. Si se desea explorar, este método se encuentra en el programa SIMMAP (Bollback, 2006) y en el paquete Phylogenetic Tools for comparative biology (and other things) (phytools) de R (Revell, 2012).

Los métodos disponibles para hacer reconstrucciones de caracteres ancestrales suelen ser intuitivos y para que la interpretación tenga sentido hay cuatro aspectos a considerar: (i) es necesario conocer toda la información disponible sobre la historia de vida de los organismos a estudiar, de lo contrario es complicado interpretar el resultado de la reconstrucción de la evolución del carácter de interés; (ii) el carácter estudiado debe tener señal filogenética, que es la dependencia de los caracteres y los grupos de especies en que se presentan (Paradis, 2012). En otras palabras, se dice que si un carácter se presenta aleatoriamente en grupos no relacionados carece de señal filogenética; (iii) es conveniente conocer un estimado de la tasa de evolución del carácter; (iv) la elección del método puede ser importante. De hecho, Xiang & Thomas (2008) encontraron que el impacto que el método puede tener sobre la reconstrucción de un carácter depende precisamente de la naturaleza de dicho carácter, pues para aquellos sin homoplasia, sin polimorfismos y sin datos faltantes, la reconstrucción de los estados ancestrales es consistente entre todos los métodos. Pero cuando los caracteres no son "perfectos", cada método trata con la incertidumbre de una forma diferente, por lo que el patrón general es que en los nodos de la filogenia con poco soporte la reconstrucción tenderá a ser incongruente entre los métodos.

Se debe considerar que, para hacer comparaciones adecuadas entre caracteres utilizando un marco de trabajo filogenético, en primera instancia es necesario realizar una diagnosis de los caracteres y sólo luego es que se establecen las comparaciones a que haya lugar. La diagnosis consiste en determinar qué asociación tiene la filogenia con las diferencias observadas entre los taxa (señal filogenética). En este sentido, el problema de la asociación de los caracteres con la filogenia tiene que ver con la independencia de los datos, que a su vez está en íntima relación con la jerarquía que hay en las filogenias y con el supuesto de que los organismos que son filogenéticamente cercanos suelen ser similares (Gittleman & Luh, 1992; Harvey & Pagel, 1991).

[3] Disponible en http://www.evolution.rdg.ac.uk/

La señal filogenética de un carácter es una consecuencia directa de la evolución de los caracteres y su forma dependerá de los mecanismos evolutivos involucrados en la historia de los que se hallen bajo análisis. Estadísticamente, la señal filogenética es la presencia de covarianza entre especies o, en otras palabras, es la no independencia de caracteres (Paradis, 2012). En la misma publicación se menciona que intuitivamente es posible conocer si un carácter tiene o no señal filogenética, por lo que hacer un cálculo de esta índole puede parecer trivial. Aunque los investigadores deben estar conscientes que cuantificar esta señal es importante debido a que no es lo mismo la señal de un carácter con una tasa de evolución rápida a la de uno con una tasa de evolución lenta. Existen dos formas de calcular la señal filogenética utilizando R, una es la propuesta por Blomberg, Garland & Ives (2003) y otra es a través de la descomposición ortonormal (ver más adelante).

Una vez que se ha determinado si hay señal filogenética, existen varios métodos para el desarrollo de análisis comparados filogenéticos. La gran mayoría de ellos implementados en R, sólo unos pocos como el de los contrastes independientes de Felsenstein (1985), están en programas como Mesquite. Enseguida se mencionarán algunos de los métodos y se explicará brevemente en qué consisten. Pero antes de continuar es necesario hacer eco de dos advertencias formuladas por Paradis (2012): (i) Hay que tener en cuenta que la distribución de los estados de carácter depende de la filogenia y de la manera en que los caracteres bajo estudio evolucionan; (ii) como sucede con cualquier método estadístico, un uso inadecuado puede dar origen a resultados sin sentido (*garbage in, garbage out*).

El método conocido como contrastes independientes fue creado por Felsenstein (1985) y hace comparaciones entre pares de taxa de cada una de las bifurcaciones de la filogenia. Así, si los nodos ancestrales de la filogenia son conocidos, entonces se pueden calcular las diferencias entre los dos taxa que comparten un ancestro común inmediato, sin que se confunda la comparación con el efecto de la filogenia. Este método se basa en un modelo browniano de evolución de caracteres, por lo que asume que cada nodo es independiente; para utilizarlo se requiere que la filogenia sea perfectamente dicotómica y el resultado es una gráfica similar a la de una regresión en la que los caracteres comparados, si están asociados, se acercarán a la línea, de lo contrario estarán dispersos.

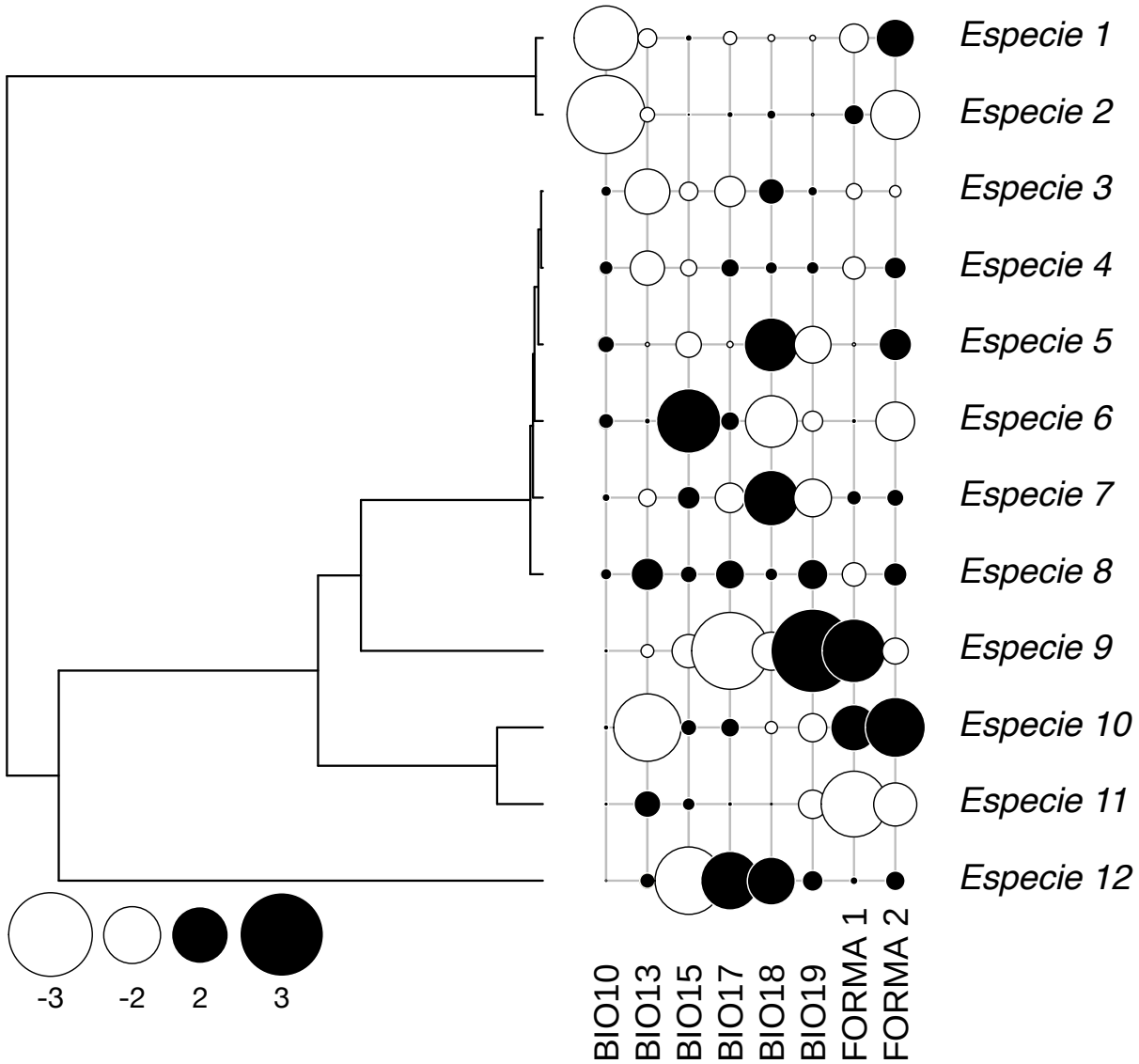
La descomposición ortonormal es un procedimiento canónico que permite descomponer la varianza de los caracteres de historia de vida (el método sólo trabaja con caracteres continuos) con respecto a la estructura del árbol filogenético. Al hacer esto es posible cuantificar hasta qué punto la historia evolutiva ha moldeado los estados que vemos hoy de un carácter. Los creadores de esta prueba sugieren que antes de aplicarla se haga un análisis de autocorrelación filogenética y si efectivamente existe, entonces se debe utilizar para cuantificar el grado de influencia de la filogenia. Este análisis da como resultado un gráfico llamado ortograma (Ollier, Couteron & Chessel, 2006).

Hay métodos de análisis multivariados, como el análisis de componentes principales filogenéticos (pPCA), propuesto por Jombart, Pavoine, Devillard & Pontier (2010). La idea de esta técnica es resumir un set de caracteres a unas pocas variables sintéticas que exhiban los patrones globales y locales que existen en la estructura filogenética. Los patrones globales son las autocorrelaciones positivas que forman taxa cercanamente vinculados que comparten los valores de un carácter, esto normalmente se asocia a la señal filogenética. Los patrones locales son las autocorrelaciones negativas que se forman a partir de las disimilitudes que se puede alcanzar en sectores de la filogenia; generalmente esto ocurre cuando especies cercanamente relacionadas tienen valores muy dispares para un carácter. El resultado de esta comparación

es un gráfico en el que se conjuga una filogenia con un análisis de componentes principales (PCA) para mostrar los patrones globales, locales, así como la proporción de valores propios (*eingenvalues*) y cuáles son los caracteres con mayor peso en conjunto con su asociación. En la Figura 4 se puede ver un pPCA generado con datos hipotéticos de nicho climático y forma del labelo de un grupo de orquídeas, en el que se aprecia cuáles de los caracteres incluidos tienen algún tipo de relación con la filogenia y cuál es su intensidad.

Figura 4.

Análisis de componentes principales filogenético para un grupo hipotético de especies en el que se muestra la relación entre la filogenia, variables bioclimáticas y dos variables relacionadas con la forma del labelo. Los círculos negros representan evolución asociada a la filogenia y los blancos, lo contrario; el tamaño de los círculos habla de la intensidad de la asociación. Figura elaborada con datos propios.



Conclusión

Los métodos descritos no son los únicos que hay y de hecho constantemente se añaden más posibilidades que permiten explorar distintos aspectos de la dinámica evolutiva. Por ejemplo, es posible estudiar la incertidumbre filogenética, hacer comparación directa entre topologías o de éstas con matrices, estimar tasas de diversificación, hacer pruebas para determinar si existen saltos en los patrones de diversificación. Incluso se pueden hacer análisis de ecología evolutiva como explorar si existe conservadurismo de nicho, si hay evolución relacionada con la exploración de nuevos espacios bioclimáticos, evaluar el desplazamiento de caracteres mediado por presiones selectivas, origen y cambio temporal de caracteres, influencia del cambio climático sobre la evolución y diversificación de seres vivos, entre otras cosas. El tipo y cantidad de datos que pueden utilizar los análisis macroevolutivos es variado y se pueden derivar prácticamente de cualquier fuente imaginable, por lo que vale la pena indagar, siempre y cuando se hagan los ajustes necesarios para que puedan ser procesados por los métodos disponibles. Así, los análisis macroevolutivos son una ventana que posibilita estudiar el pasado, comprender el presente y tal vez permitan inferir eventos del futuro.

Referencias

Adams, D. C. (2010). Parallel evolution of character displacement driven by competitive selection in terrestrial salamanders. *BMC evolutionary biology*, *10*, 72. doi: 10.1186/1471-2148-10-72

Aguilar-Rodríguez, P. A., Krömer, T., Tschapka, M., García-Franco, J. G., Escobedo-Sarti, J. & MacSwiney, C. (2019). Bat pollination in Bromeliaceae. *Plant Ecology and Diversity*, *12*(1), 1-19. doi: 10.1080/17550874.2019.1566409

Aris-Brosou, S. & Yang, Z. (2002). Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Systematic Biology, 51*(5), 703-714. doi: 10.1080/10635150290102375

Barba-Montoya, J., dos Reis, M., Schneider, H., Donoghue, P. C. J. & Yang, Z. (2018). Constraining uncertainty in the timescale of angiosperm evolution and the veracity of a Cretaceous Terrestrial Revolution. *New Phytologist*, *218*(2), 819-834. doi: 10.1111/nph.15011

Battistuzzi, F. U., Filipski, A. J. & Kumar, S. (2011). Molecular Clock: Testing. *eLS*, 1-7. doi: 10.1002/9780470015902.a0001803.pub2

Blomberg, S. P., Garland, T. & Ives, A. R. (2003). Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. *Evolution*, *57*(4), 717. doi: 10.1554/0014-3820(2003)057[0717:TFPSIC]2.0.CO;2

Bollback, J. P. (2006). SIMMAP: stochastic character mapping of discrete traits on phylogenies. *BMC bioinformatics*, *7*, 88. doi: 10.1186/1471-2105-7-88

Bremer, K. & Gustafsson M., H. G. (1997). East Gondwana ancestry of the sunflower alliance of families. *Proceedings of the National Academy of Sciences*, *94*(17), 9188-9190. doi: 10.1073/pnas.94.17.9188

Britton, T., Anderson, C. L., Jacquet, D., Lundqvist, S. & Bremer, K. (2007). Estimating divergence times in large phylogenetic trees. *Systematic Biology, 56*(5), 741-752. doi: 10.1080/10635150701613783

Drummond, A. J., Ho, S. Y. W., Phillips, M. J. & Rambaut, A. (2006). Relaxed phylogenetics and dating with confidence. *PLoS Biology*, *4*(5), 699-710. doi: 10.1371/journal.pbio.0040088

Drummond, A. J. & Suchard, M. A. (2010). Bayesian random local clocks, or one rate to rule them all. *BMC Biology*, *8*(1), 114. doi: 10.1186/1741-7007-8-114

Escobedo-Sarti, J., Ramírez, I., Leopardi, C., Carnevali, G., Magallón, S., Duno, R. & Mondragón, D. (2013). A phylogeny of Bromeliaceae (Poales, Monocotyledoneae) derived from an evaluation of nine supertree methods. *Journal of Systematics and Evolution*, *51*(6), 743-757. doi: 10.1111/jse.12044

Eldredge, N. & Cracraft, J. (1980). *Phylogenetic patterns and the evolutionary process, method and theory in comparative biology*. New York: Columbia University Press.

Evans, M. E. K., Smith, S. A., Flynn, R. S. & Donoghue, M. J. (2009). Climate, niche evolution, and diversification of the "Bird-Cage" evening primroses (*Oenothera*, sections *Anogra* and *Kleinia*). *The American Naturalist*, *173*(2), 225-240. doi: 10.1086/595757

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, *17*(6), 368-376. doi: 10.1007/BF01734359

Felsenstein, J. (1985). Phylogenies and the comparative method. *The American Naturalist*, *125*(1), 1-15. doi: 10.1086/284325

Gittleman, J. L. & Luh, H. (1992). On comparing comparative methods. *Annual Review of Ecology and Systematics*, *23*(1), 383-404. doi: 10.1146/annurev.es.23.110192.002123

Gómez-Acevedo, S., Rico-Arce, L., Delgado-Salinas, A., Magallón, S. & Eguiarte, L. E. (2010). Neotropical mutualism between *Acacia* and *Pseudomyrmex*: Phylogeny and divergence times. *Molecular Phylogenetics and Evolution*, *56*(1), 393-408. doi: 10.1016/j.ympev.2010.03.018

Gupta, A., Maňuch, J., Stacho, L. & Zhu, C. (2004). Small Phylogeny Problem: Character Evolution Trees. En Sahinalp, S.C., Muthukrishnan, S. & Dogrusoz, U. (Eds.), *Combinatorial Pattern Matching. CPM 2004. Lecture Notes in Computer Science* (23-243), vol 3109. Berlin: Springer. doi: 10.1007/978-3-540-27801-6_17

Harmon, L. J., Weir, J. T., Brock, C. D., Glor, R. E. & Challenger, W. (2008). GEIGER: investigating evolutionary radiations. *Bioinformatics*, *24*(1), 129-131. doi: 10.1093/bioinformatics/btm538

Harvey, P. & Pagel, M. D. (1991). *The comparative method in evolutionary biology*. Oxford, Reino Unido: Oxford University Press.

Herron, J. C. & Freeman, S. (2014). *Evolutionary analysis*. Glenview, Estados Unidos: Pearson.

Jombart, T., Pavoine, S., Devillard, S. & Pontier, D. (2010). Putting phylogeny into the analysis of biological traits: A methodological approach. *Journal of Theoretical Biology*, *264*(3), 693-701. doi: 10.1016/j.jtbi.2010.03.038

Koonin, E.V. (2012). A half-century after the molecular clock: new dimensions of molecular evolution. *EMBO reports*, 13, 664-666. doi: 10.1038/embor.2012.103

Lartillot, N., Phillips, M. J. & Ronquist, F. (2016). A mixed relaxed clock model. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1699), 20150132. doi: 10.1098/rstb.2015.0132

Leopardi-Verde, C. L., Carnevali, G. & Romero-González, G. A. (2017). A phylogeny of the genus *Encyclia* (*Orchidaceae: Laeliinae*), with emphasis on the species of the Northern Hemisphere. *Journal of Systematics and Evolution*, *55*(2), 110-123. doi: 10.1111/jse.12225

Lewis, P. O. (2001). A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology*, *50*(6), 913-925. doi: 10.1080/106351501753462876

Li, W.-H. & Graur, D. (1991). *Fundamentals of molecular evolution*. Sunderland, Estados Unidos: Sinauer Associates.

Maddison, W. P. & Maddison, D. R. (2019). Mesquite: A modular system for evolutionary analysis. Recuperado de https://www.mesquiteproject.org

Magallón, S., Gómez-Acevedo, S., Sánchez-Reyes, L. L. & Hernández-Hernández, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytologist*, *207*(2), 437-453. DOI: 10.1111/nph.13264

Margoliash, E. (1963). Primary structure and evolution of *Cytochrome* C. *Proceedings of the National Academy of Sciences*, *50*(4), 672-679. doi: 10.1073/pnas.50.4.672

Nei, M. (1987). *Molecular evolutionary genetics*. New York: Columbia University Press.

Ohta, T. (2013). Molecular Evolution: Nearly Neutral Theory. *eLS*, 1-6. doi: 10.1002/9780470015902.a0001801.pub4

Ollier, S., Couteron, P. & Chessel, D. (2006). Orthonormal transform to decompose the variance of a life-history trait across a phylogenetic tree. *Biometrics*, *62*(2), 471-477. doi: 10.1111/j.1541-0420.2005.00497.x

Paradis, E. (2012). *Analysis of phylogenetics and evolution with R*. New York: Springer. doi: 10.1007/978-1-4614-1743-9

Paradis, E., Claude, J. & Strimmer, K. (2004). APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, *20*(2), 289-290. doi: 10.1093/bioinformatics/btg412

Polly, D., Lawing, M., Fabre, A. C. & Goswami, A. (2013). Phylogenetic principal components analysis and geometric morphometrics. *Hystrix*, *24*(1), 33-41. doi: 10.4404/hystrix-24.1-6383

R Core Team. (2020). R: A language and environment for statistical computing. Recuperado de https://www.r-project.org

Renner, S. S. (2005). Relaxed molecular clocks for dating historical plant dispersal events. *Trends in Plant Science*, *10*(11), 550-558. doi: 10.1016/j.tplants.2005.09.010

Revell, L. J. (2012). phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution, 3*(2), 217-223. doi: 10.1111/j.2041-210X.2011.00169.x

Rutschmann, F. (2006). Molecular dating of phylogenetic trees: A brief review of current methods that estimate divergence times. *Diversity and Distributions*, *12*(1), 35-48. doi: 10.1111/j.1366-9516.2006.00210.x

Sanderson, M. J. (1997). A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution*, *14*(12), 1218-1231. doi: 10.1093/oxfordjournals.molbev.a025731

Sanderson, M. J. (2002). Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular Biology and Evolution*, *19*(1), 101-109. doi: 10.1093/oxfordjournals.molbev.a003974

Sanderson, M. J. (2003). r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics*, *19*(2), 301-302. doi: 10.1093/bioinformatics/19.2.301

Sanderson, M., Thorne, J., Wikström, N. & Bremer, K. (2004). Molecular evidence on plant divergence times. *American Journal of Botany*, *91*(10), 1656-1665. doi: 10.3732/ajb.91.10.1656

Sauquet, H., Von Balthazar, M., Magallón, S., Doyle, J. A., Endress, P. K., Bailes, E. J., ... & Schönenberger, J. (2017). The ancestral flower of angiosperms and its early diversification. *Nature Communications*, *8*, 16047. doi: 10.1038/ncomms16047

Schneider, H., Schuettpelz, E., Pryer, K. M., Cranfill, R., Magallón, S. & Lupia, R. (2004). Ferns diversified in the shadow of angiosperms. *Nature*, *428*(6982), 553-557. doi: 10.1038/nature02361

Schultz, T. R. & Brady, S. G. (2008). Major evolutionary transitions in ant agriculture. *Proceedings of the National Academy of Sciences*, *105*(14), 5435-5440. doi: 10.1073/pnas.0711024105

Serb, J. M., Alejandrino, A., Otárola-Castillo, E. & Adams, D. C. (2011). Morphological convergence of shell shape in distantly related scallop species (*Mollusca: Pectinidae*). *Zoological Journal of the Linnean Society*, *163*(2), 571-584. doi: 10.1111/j.1096-3642.2011.00707.x

Silvera, K., Santiago, L. S., Cushman, J. C. & Winter, K. (2009). Crassulacean acid metabolism and epiphytism linked to adaptive radiations in the Orchidaceae. *Plant Physiology*, *149*(4), 1838-1847. doi: 10.1104/pp.108.132555

Soltis, D. E. & Soltis, P. S. (1999). Choosing an approach and an appropriate gene for phylogenetic analysis. En D. E. Soltis, P. S. Soltis & J. J. Doyle (Eds.), *Molecular Systematics of Plants II-DNA Sequencing* (pp. 1-42). New York: Springer.

Suchard, M. A., Lemey, P., Baele, G., Ayres, D. L., Drummond, A. J. & Rambaut, A. (2018). Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evolution*, *4*(1), vey016. doi: 10.1093/ve/vey016

Swofford, D. L. & Sullivan, J. (2009). Phylogeny inference based on parsimony and other methods using PAUP*. En P. Lemey, M. Salemi & A. Vandamme (Eds.), *The Phylogenetic Handbook* (pp. 267-312). Cambridge, Reino Unido: Cambridge University Press. doi: 10.1017/CBO9780511819049.010

Vandamme, A. (2009). Basic concepts of molecular evolution. En P. Lemey, M. Salem & A. Vandamme (Ed.), *The Phylogenetic Hanbook: A Practical Approach to Phylogenetic and Hypothesis Testing* (pp. 3-29). Cambridge, Reino Unido: Cambridge University Press. doi: 10.1017/CBO9780511819049.003

Wagner, G. (Ed.). (2001). *The character concept in evolutionary biology*. Florida, Estados Unidos: Academic Press.

Welch, J. J. & Bromham, L. (2005). Molecular dating when rates vary. *TRENDS in Ecology and Evolution, 20*(6), 320-327. doi: 10.1016/j.tree.2005.02.007

Xiang, Q. Y. & Thomas, D. T. (2008). Tracking character evolution and biogeographic history through time in Cornaceae-Does choice of methods matter? *Journal of Systematics and Evolution*, *46*(3), 349-374. doi: 10.3724/SP.J.1002.2008.08056

Yesson, C. & Culham, A. (2006). Phyloclimatic modeling: Combining phylogenetics and bioclimatic modeling. *Systematic Biology*, *55*(5), 785-802. doi: 10.1080/1063515060081570

Zuckerkandl, E. & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. En V. Bryson & H. J.Vogel (Eds.), *Evolving genes and proteins* (pp. 97-166). Londres: Academic Press. doi: 10.1016/B978-1-4832-2734-4.50017-6

"Pentapetalae"
Cerámica/esmalte
16 x 21.5 x 21.5 cm
2016

Tequio *4*(11), 2021: 69-79
ISSN: 2594-0546

Métodos de reconstrucción filogenética I: máxima verosimilitud

Methods for phylogenetic reconstruction I: maximum likelihood

Pablo Duchen[1*]

Fecha de recepción: 6 de noviembre de 2020
Fecha de aceptación: 22 de enero de 2021

Resumen - La inferencia filogenética es ampliamente utilizada en biología evolutiva, la cual tiene el objetivo de encontrar las relaciones evolutivas entre diferentes especies y representarlas en la forma de un árbol filogenético (o filogenia). Existen varios métodos estadísticos para la inferencia filogenética. En esta revisión se presenta la máxima verosimilitud como modelo de reconstrucción filogenética, método que consiste en calcular la verosimilitud de múltiples filogenias candidatas y reportar aquella con el valor máximo como la filogenia representativa de un grupo de organismos. En la presente revisión se explica cómo se calcula la verosimilitud de una filogenia a partir de secuencias de ADN provenientes de varias especies. También se presentan modelos de mutación de ADN para calcular probabilidades de transición entre nucleótidos, los cuales son usados en la estimación de la verosimilitud. Se muestra también un ejemplo ilustrativo sencillo que aplica los pasos necesarios para inferir una filogenia y se explica el software más usado para inferencia bajo máxima verosimilitud para alineamientos de ADN más grandes.

Palabras clave: Prunning, Jukes-Cantor, inferencia filogenética, alineamiento, bootstrap, modelos de mutación de ADN.

Abstract - Phylogenetic inference is widely used in evolutionary biology, aiming to find evolutionary relationships between different species and report the result in the form of a phylogenetic tree (phylogeny). There are several statistical methods used for phylogenetic inference. In this review, the method of maximum likelihood for phylogenetic reconstruction is presented. This technique consists of finding the likelihood of multiple candidate phylogenies, and report the one with the highest likelihood as a representative of the evolutionary relationships of a group of species. In this paper, the likelihood calculation of a phylogeny from multiple-species DNA sequences is reviewed. Also, some key DNA mutation models to calculate transition probabilities between nucleotides are presented. Such transition probabilities are used in the likelihood calculation of a given phylogeny. A simple example is shown to illustrate the necessary steps to infer a phylogeny, as well as the most common software for maximum likelihood inference for larger DNA alignments.

Keywords: Prunning, Jukes-Cantor, phylogenetic inference, alignment, bootstrap, DNA mutation models.

[1] Departamento de Biología Computacional, Universidad de Lausana, Suiza. *Correos electrónicos: pablo.duchenbocangel@unil.ch, pduchen@gmail.com
ORCID: 0000-0002-9318-5002

Introducción

La sistemática filogenética tiene como objetivo encontrar las relaciones evolutivas o de parentesco entre diferentes especies o distintos taxones supraespecíficos. Dichas conexiones evolutivas se representan comúnmente en la forma de un árbol filogenético, donde organismos emparentados están unidos por medio de líneas que simbolizan las ramas del árbol, y donde los nodos significan ancestros comunes entre especies o clados. Dada la relevancia biológica de conocer las relaciones evolutivas entre distintas especies, inferir una filogenia a partir de datos morfológicos o moleculares es una práctica muy efectuada en biología evolutiva y taxonomía. Por otro lado, las filogenias también se usan para reconstruir caracteres ancestrales (Pagel, 1999), establecer relojes moleculares (Bronham & Penny, 2003), o para estudiar la evolución de caracteres morfológicos (Duchen, Alfaro, Rolland, Salamin & Silvestro, 2020).

Existen varios métodos estadísticos para la inferencia o reconstrucción filogenética (Brocchieri, 2001); entre ellos, la máxima verosimilitud (ML) y la inferencia bayesiana son probablemente los más usados en la actualidad. A diferencia de procedimientos basados en distancias genéticas entre secuencias (como el *neighbor joining* o el UPGMA), o basados en máxima parsimonia (Edwards & Cavalli-Sforza, 1963; Peña, 2011), ML y el bayesiano utilizan las verosimilitudes de cada posición (o columna) en un alineamiento de secuencias de ADN para inferir una filogenia. En este trabajo se van a desarrollar primero los pasos necesarios para calcular una filogenia basada en ML y en el siguiente artículo se plantearán los pasos para realizar una inferencia bayesiana. Se comenzará con el algoritmo necesario para calcular la verosimilitud de un alineamiento (dada una filogenia particular); luego se describirán dos ejemplos de modelos de mutación de ADN, los cuales son usados en el cálculo de verosimilitudes; se concluye con un ejemplo simple de inferencia filogenética, mencionando además el software usado para alineamientos más grandes.

Cálculo de la verosimilitud de una topología

Es importante comenzar por definir nuestros datos y parámetros a estimar. Dado un alineamiento D de secuencias de ADN para un número n de especies, el objetivo es encontrar la filogenia, árbol o topología T que mejor describa dicho alineamiento. A lo largo de este documento se usarán los vocablos filogenia y árbol indistintamente para referirse a T, al igual que los términos alineamiento o datos para referirse a D.

Dado un alineamiento D y asumiendo que la evolución en cada posición de D y cada rama de T es independiente, la verosimilitud está determinada por:

$$P(D|T) = \prod_{k=1}^{m} P\left(D^{(k)}|T\right), \quad (1)$$

Donde $D^{(k)}$ corresponde al alineamiento en la posición k (con un total de m posiciones). Esto significa que si podemos calcular la verosimilitud en cada posición del alineamiento separadamente, la verosimilitud total será simplemente el producto de todas las posiciones de alineamiento. Para fines prácticos y evitar problemas numéricos es mejor utilizar la versión logarítmica de dicha ecuación:

De esta manera, en vez de multiplicar todas las verosimilitudes, se suman los logaritmos de $P(D^{(k)}|T)$ para cada posición. El problema con multiplicar verosimilitudes es que éstas representan probabilidades con valores entre 0 y 1. Al multiplicar valores menores a 1 varias veces, los ceros decimales aumentan y se pierden rápidamente las cifras no periódicas durante la multiplicación. Por este motivo se debe usar la versión logarítmica.

Para calcular la verosimilitud en una posición se emplea frecuentemente el método o algoritmo de "pruning" de Felsenstein (1973), el cual se describirá a continuación.

Algoritmo "prunning"

Este algoritmo es muy eficiente para calcular la verosimilitud de una filogenia y está basado en verosimilitudes condicionales para cada clado de *T*. Se llama "condicional" a dicha verosimilitud porque su valor depende de los nucleótidos que estén en el extremo de cada clado. Aquí se denomina $V^{(k)}$ a la verosimilitud condicional de cada clado en una filogenia en la posición *k* de un alineamiento. Para no sobrecargar la notación vamos a dejar momentáneamente de lado la indicación de la posición *k*; por ejemplo, el árbol de la Figura 1 tiene las especies E_1, E_2 y E_3, las cuales, en esa posición del alineamiento, muestran las bases A, G y G, respectivamente. Por el contrario, las bases de las especies ancestrales E_{12} y E_{123} se desconocen y pueden tomar cualquier valor entre los nucleótidos A, C, G, T.

Las bases de ADN en los extremos del árbol son datos observados, por tanto, los valores de *V* en los extremos serían: V_{E1}(A,C,G,T)= (1,0,0,0), ya que se observa al nucleótido A en la especie E_1; V_{E2}(A,C,G,T)= (0,0,1,0), porque se observa al nucleótido G en la especie E_2; y V_{E3}(A,C,G,T)= (0,0,1,0). Una vez calculadas las *V*s en los extremos del árbol, se van a calcular las *V*s en los nodos internos.

La verosimilitud condicional del nodo interno E_{12} está dada por dos posibilidades: la de cambiar del estado de E_{12} a A, y de E_{12} a G, a lo largo de la rama t_{12} que separa a ambos nodos. Por tanto,

$$V_{\mathrm{E}_{12}} = (P(A|\mathrm{E}_{12}, t_{12}) \times 1)(P(G|\mathrm{E}_{12}, t_{12}) \times 1).$$

En esta ecuación, el primer factor corresponde a la probabilidad de cambiar del estado de E_{12} a A, y el segundo factor a la probabilidad de cambiar del estado de E_{12} a G. Nótese aquí que multiplicamos ambos factores por 1, que corresponden a las probabilidades de observar las bases A y G en las puntas de dicho árbol, respectivamente (estas probabilidades corresponden a V_{E1} y V_{E2}, calculados anteriormente).

Se procede ahora a calcular la verosimilitud condicional en la raíz del árbol. Si se define *x* como el nucleótido que corresponde a E_{12}, entonces la verosimilitud condicional en la raíz del árbol de la Figura 1 está dada por:

$$V_{\mathrm{E}_{12}} = \left(\sum_x P(x|\mathrm{E}_{123}, t_{123} - t_{12}) V_{\mathrm{E}_{12}}\right) P(G|\mathrm{E}_{123}, t_{123}) \times 1.$$

El primer factor muestra la probabilidad de cambiar del estado de E_{123} al nucleótido *x*, a lo largo de la rama de longitud t_{123}-t_{12} dada la probabilidad V_{E12} (calculada en el paso anterior). Nótese que aquí se suman todas las probabilidades de *x* ($x \in$ {A,C,G,T}), ya que se desconoce el nucleótido correspondiente a E_{12}. El segundo factor atañe a la probabilidad de cambiar del estado de E_{123} a G a lo largo de la rama t_{123}. Dicho factor igualmente se multiplica por 1, que refiere a la probabilidad de observar una G en el extremo del árbol.

En general, y para cualquier filogenia, asumiendo que el clado en cuestión tiene como nucleótido *s* ($s \in$ {A,C,G,T}) y tiene dos descendientes con nucleótidos *x* ($x \in$ {A,C,G,T}) y *y* ($y \in$ {A,C,G,T}), con longitudes de rama t_x y t_y, entonces cada V_E está dada por:

$$V_E = \left(\sum_x P(x|s,t_x)V_x\right)\left(\sum_y P(y|s,t_y)V_y\right)$$

De esta manera, comenzando por la punta del árbol, se calculan las verosimilitudes condicionales descendiendo por cada nodo hasta llegar a la raíz. Al final, la verosimilitud total de la filogenia *T* en la posición *k* (retomando la notación original) es:

$$V^{(k)} = P\left(D^{(k)}\middle|T\right) = \sum_x \pi_x V_{E_{raíz}}^{(k)}(x),$$

donde π_x es la probabilidad *a priori* del nucleótido *x* –la cual se puede estimar por su frecuencia en el alineamiento– y $E_{raíz}$ es el nucleótido en la raíz del árbol, correspondiente a E_{123} en el ejemplo de la Figura 1. Finalmente, todas las probabilidades *P(x|s,t)* o *P(y|s,t)* se calculan con diversos modelos de mutación de ADN, los cuales se describen en la sección "Modelos de mutación de ADN".

Inferencia filogenética

La inferencia por ML funciona de la siguiente manera: dadas las topologías candidatas para un alineamiento *D* particular, se pueden calcular las verosimilitudes de cada una (utilizando el algoritmo "prunning"). Luego, la topología con la mayor verosimilitud será la filogenia correspondiente a *D*. Hay dos aspectos importantes para tomar en cuenta al realizar la inferencia por ML: el *bootstrapping* y la búsqueda de topologías.

Bootstrapping. En inferencia por ML se recurre al *bootstrap* para obtener una medida de incertidumbre para el árbol con la máxima verosimilitud. El *bootstrap* en filogenética consiste en: 1) tomar muestras con reemplazo de las columnas de un alineamiento, 2) formar un nuevo alineamiento con dichas columnas, y 3) volver a inferir la topología con el nuevo alineamiento. Para ser más preciso, si *D* tiene *m* columnas, entonces se toman *m* muestras con reemplazo y se infiere la filogenia. Este proceso se repite múltiples veces. Estudios que utilizan hasta 2 500 secuencias muestran que 100 a 500 repeticiones de *bootstrap* son suficientes, pero para criterios más conservadores se llegan a hacer varios miles de repeticiones (Pattengale, Alipour, Bininda-Emonds, Moret & Stamatakis, 2010). En la filogenia final (aquella con la máxima verosimilitud) se reporta el porcentaje de ocasiones que cada clado se mantiene en las repeticiones del *bootstrapping*. Clados con valores de *bootstrap* mayores a 75% se consideran con buen soporte estadístico.

Búsqueda de topologías. Otro aspecto importante constituye la búsqueda de topologías. Para alineamientos con pocas especies es posible calcular la verosimilitud de todas las topologías posibles, lo que se conoce como una búsqueda exhaustiva. Sin embargo, para alineamientos con muchas especies la cantidad de topologías para analizar es muy grande, por lo que se emplea la búsqueda heurística (aproximada). Para dar un ejemplo, un alineamiento con tres especies tiene tres topologías posibles; cuatro especies, 15 topologías posibles (Fig. 2);

cinco especies tienen 105 topologías, y si hablamos de un alineamiento de 50 especies (muy común en estudios biológicos) tendríamos $2{,}75\times10^{76}$ topologías posibles. Por tanto, computacionalmente no es realista calcular la verosimilitud de tal cantidad de árboles. Para solucionar esto, existen algoritmos que hacen una búsqueda de topologías solamente entre aquellas con mayor verosimilitud (Stamatakis, 2014).

Modelos de mutación de ADN

Se continúa ahora con la descripción de modelos de mutación de ADN, los cuales se usan para calcular verosimilitudes. Existen distintos modelos para estimar la probabilidad de cambiar de un nucleótido a otro a lo largo de una rama. La complejidad de dichos modelos radica en la reversibilidad de la sustitución de nucleótidos, o en la inclusión de eventos denominados "transiciones" y "transversiones". Desde el punto de vista genético-molecular, una transición es una sustitución entre purinas (nucleótidos A y G) o pirimidinas (C y T), mientras que una transversión es un cambio de una purina a una pirimidina, o viceversa.

Jukes-Cantor

El modelo más simple es el de Jukes-Cantor (Jukes & Cantor, 1969), donde se asume que la probabilidad de que un nucleótido modifique a los otros tres es la misma. Partiendo de que la tasa instantánea de cambio de un nucleótido específico es $\mu/3$, y que la frecuencia de cada nucleótido es 1/4 ($\pi_A = \pi_C = \pi_G = \pi_T = ¼$), entonces la tasa de sustitución total de un nucleótido cualquiera es $\mu/3 + \mu/3 + \mu/3 + \mu/3 = 4\mu/3$.

Para calcular la probabilidad de un evento de sustitución a lo largo de una rama de longitud t se emplea la distribución de Poisson. En este caso, la probabilidad de una "no" sustitución es $e^{-4\mu t/3}$, donde $4\mu/3$ corresponde a la tasa de sustitución calculada en el párrafo anterior y la probabilidad de al menos un evento de sustitución es $1\text{-}e^{-4\mu t/3}$. Por tanto, la probabilidad de cambio de un nucleótido s a otro x a lo largo de una rama de longitud t viene dada por:

$$P(x|s,\mu,t) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}\mu t}\right),$$

donde el factor 1/4 indica la probabilidad de que el último evento de sustitución resulte en el nucleótido x.

El modelo de Kimura (1980) también asume que las frecuencias nucleotídicas π_X ($X \in \{A,C,G,T\}$) son las mismas. Sin embargo, a diferencia del modelo de Jukes-Cantor, el de Kimura distingue entre transiciones y transversiones. No se desarrollará aquí este modelo, pero se presentará uno más general que incluye al de Kimura, además de otros como casos especiales.

Tamura-Nei

Este modelo descrito originalmente por Tamura & Nei (1993) hace diferencia entre transiciones y transversiones, y también permite que las frecuencias nucleotídicas π_X sean distintas entre sí. Para tomar en cuenta transiciones y transversiones –y siguiendo la notación en Felsenstein (2004)– se definirá como α_R a la probabilidad de escoger una purina a partir una purina, y α_Y a la probabilidad de escoger una pirimidina a partir de una pirimidina. Además, se definirá como β a la probabilidad de elegir un nucleótido cualquiera a partir de las cuatro bases A, C, G y T.

Con base en estas definiciones la tasa instantánea de cambio entre las purinas e.g. G a A sería $\alpha_R \frac{\pi_A}{\pi_A+\pi_G} + \beta\pi_A$. Aquí, el primer término representa la probabilidad de escoger otra purina (A) dado que se parte de una purina (G), mientras que el segundo término representa la probabilidad de escoger al nucleótido A a partir de

cualquier otro. Para el caso de una transversión (por ejemplo de C a A) la tasa instantánea de cambio es simplemente $\beta\pi_A$. De acuerdo con estos ejemplos, es sencillo escribir las tasas de cambio para el resto de nucleótidos.

Para calcular la probabilidad de un evento de sustitución a lo largo de una rama de longitud t también se utilizará la distribución de Poisson. Si se comienza con una purina, la probabilidad de que no haya ningún evento de transición a lo largo de t es $e^{-\alpha_R t}$, la probabilidad de transversiones es $(1-e^{-\beta t})$, la de transiciones pero no transversiones es $(1-e^{-\alpha_R t})e^{-\beta t}$, la de que no ocurra ningún evento es $e^{-(\alpha_R+\beta)t}$, etcétera. Estos mismos criterios para cálculo de probabilidades aplican para las pirimidinas. Ahora sí, la probabilidad de sustitución total entre e.g. A y G a lo largo de t es:

$$P(G|A,t) = e^{-\beta t}(1-e^{-\alpha_R t})\frac{\pi_G}{\pi_A+\pi_G} + \left(1-e^{-\beta t}\right)\pi_G.$$

En otras palabras, para obtener una G a partir de una A existen dos posibilidades: 1) no puede haber ninguna transversión $(e^{-\beta t})$ y tiene que haber una transición $(1-e^{-\alpha_R t})$ que resulte en una G, o 2) pueden existir transversiones $(1-e^{-\beta t})$ siempre y cuando la última resulte en una G. De esta misma manera pueden escribirse expresiones para las otras sustituciones entre los distintos pares de nucleótidos.

Se finaliza esta sección mencionando los modelos F84 (Felsenstein & Churchill, 1996), HKY (Hasegawa, Kishino & Yano, 1985) y el modelo *General time reversible* (GTR) (Lanave, Preparata, Sacone & Serio, 1984; Tavaré, 1986). En el modelo F84 $\alpha_R = \alpha_Y$. Por otro lado, si $\frac{\alpha_R}{\alpha_Y} = \frac{\pi_A+\pi_C}{\pi_G+\pi_T}$ entonces obtenemos el modelo HKY. Finalmente, el modelo GTR generaliza los modelos vistos anteriormente, ya que incluye seis tasas de sustitución (una para cada par de nucleótidos). Si bien no es necesaria su descripción a detalle en la presente revisión, este modelo está implementado en los programas clásicos de inferencia filogenética (ver sección "Software para inferencia filogenética con ML").

Ejemplo de algoritmo para inferencia por ML

Ahora se va a desarrollar un ejemplo muy sencillo para inferir una filogenia usando ML. Se utilizará el alineamiento del ejemplo 1.1.1 del artículo "Modelo de estimación de pesos de árbol filogenético para un cuartet, aplicando conjugación de Hadamard" (publicado también en este número de la revista). Dicho alineamiento contiene cuatro especies y 16 posiciones; se convierte a formato FASTA y se guarda en un archivo denominado "alineamiento.fas":

```
>E1
CCATCAAACGTGTGAC
>E2
ACAGCAATGTTATCTC
>E3
CCATTGAAGATGCGTT
>E4
ACAGTAGTGTTACCAG
```

Posteriormente, se consideran posibles topologías para las especies E1, E2, E3 y E4. En total existen 15 posibles topologías, las cuales se pueden escribir en formato NEWICK y se guardan en un archivo denominado "topologias.tre" (visualizadas en la Figura 2):

```
( ( E1 : 1, E2 : 1 ) : 1, ( E3 : 1, E4 : 1 ) : 1 ) : 1;
( ( E1 : 1, E4 : 1 ) : 1, ( E2 : 1, E3 : 1 ) : 1 ) : 1;
( ( ( E1 : 1, E2 : 1 ) : 1, E3 : 1 ) : 1, E4 : 1 ) : 1;
( ( ( E1 : 1, E2 : 1 ) : 1, E4 : 1 ) : 1, E3 : 1 ) : 1;
( ( ( E1 : 1, E3 : 1 ) : 1, E2 : 1 ) : 1, E4 : 1 ) : 1;
( ( ( E1 : 1, E3 : 1 ) : 1, E4 : 1 ) : 1, E2 : 1 ) : 1;
( ( E1 : 1, E3 : 1 ) : 1, ( E2 : 1, E4 : 1 ) : 1 ) : 1;
( ( ( E1 : 1, E4 : 1 ) : 1, E2 : 1 ) : 1, E3 : 1 ) : 1;
( ( ( E1 : 1, E4 : 1 ) : 1, E3 : 1 ) : 1, E2 : 1 ) : 1;
( ( ( E2 : 1, E3 : 1 ) : 1, E1 : 1 ) : 1, E4 : 1 ) : 1;
( ( ( E2 : 1, E3 : 1 ) : 1, E4 : 1 ) : 1, E1 : 1 ) :1;
( ( ( E2 : 1, E4 : 1 ) : 1, E1 : 1 ) : 1, E3 : 1 ) :1;
( ( ( E2 : 1, E4 : 1 ) : 1, E3 : 1 ) : 1, E1 : 1 ) : 1;
( ( ( E3 : 1, E4 : 1 ) : 1 , E2 : 1 ) : 1, E1 : 1 ) :1;
( ( ( E3 : 1, E4 : 1 ) : 1, E1 : 1 ) : 1, E2 : 1 ) :1;
```

Finalmente, utilizando los archivos que recién se crearon ("alineamiento.fas" y "topologias.tre") como input, se desarrolla a continuación un programa corto (escrito en el lenguaje de programación R) para ejemplificar el algoritmo de inferencia filogenética utilizando ML:

```
#######-INICIO DEL PROGRAMA-#######
library(ape)
library(phangorn)
#Alineamiento de ADN.
D <- phyDat(read.FASTA("alineamiento.fas")) #Posibles topologias.
T <- read.tree("topologias.tre")

##-------Maxima Verosimilitud-------##
cat ("\nEjemplo de inferencia con ML\n")

#El vector logV guardara las verosimilitudes de cada topología.
logV <- numeric(length(T))

#La función pml de la librería phangorn calcula la verosimilitud.
for (i in 1:length(T)) {
        logV[i] <- pml(T[[i]],D)$logLik
        print(paste("Log Verosimilitud topologia",i,"=",logV[i] ) )
}
```

```
#Aqui se escoge la topología con la máxima verosimilitud. print ( paste ( "Topologia con la maxima
verosimilitud =",
                    which ( logV==max(logV) ) ) )
#######-FIN DEL PROGRAMA-#######
```

El output de este programa es el siguiente:

```
Ejemplo de inferencia con ML
[1] "Log Verosimilitud topologia 1 = -87.1232112880638"
[1] "Log Verosimilitud topologia 2 = -87.8578012541171"
[1] "Log Verosimilitud topologia 3 = -87.0763445725479"
[1] "Log Verosimilitud topologia 4 = -87.0328449489164"
[1] "Log Verosimilitud topologia 5 = -86.6093614372388"
[1] "Log Verosimilitud topologia 6 = -86.7630718337047"
[1] "Log Verosimilitud topologia 7 = -85.6551750661665"
[1] "Log Verosimilitud topologia 8 = -87.61408828079"
[1] "Log Verosimilitud topologia 9 = -87.8112983008873"
[1] "Log Verosimilitud topologia 10 = -87.5197055341976"
[1] "Log Verosimilitud topologia 11 = -87.5197055341976"
[1] "Log Verosimilitud topologia 12 = -86.2876683750476"
[1] "Log Verosimilitud topologia 13 = -86.331167998679"
[1] "Log Verosimilitud topologia 14 = -87.3545380111076"
[1] "Log Verosimilitud topologia 15 = -87.5082484075735"
[1] "Topologia con la maxima verosimilitud = 7"
```

Como se puede corroborar, la topología con la máxima verosimilitud es la 7 (Figura 2). En la realidad, para un alineamiento tan reducido, otros métodos basados en distancias genéticas serán suficientes para inferir la filogenia. ML es más útil para alineamientos con más especies y más posiciones.

Software para inferencia filogenética con ML

Existen muchos programas que calculan filogenias utilizando ML. Históricamente el software PAUP (Swofford, 2002) también incluye un método ML. Sin embargo, se desarrollaron otros programas con mayor velocidad, aquí se nombrarán los más usados en la actualidad.

RAxML. Desarrollado por Stamatakis (2014), este programa toma un alineamiento como input y reporta la filogenia con la máxima verosimilitud. Es muy utilizado debido a su rapidez; Además, RAxML también realiza el *bootstrapping*.

PHYLIP. Desarrollado por Felsenstein y colaboradores (Felsenstein, 2019). La primera versión salió en 1980 y desde entonces fue desarrollándose y actualizándose continuamente. PHYLIP, además de calcular filogenias

empleando ML, también lleva a cabo *bootstrapping* e incluye varios métodos de análisis evolutivo a lo largo de filogenias.

PhyML. Desarrollado por Guindon *et al.* (2010), PhyML consiste en diversos programas que calculan una filogenia utilizando ML. Además, contiene herramientas para la calibración de fósiles en filogenias y para estimar tasas de dispersión.

MEGA. Desarrollado por Hall (2013). También incluye inferencia filogenética con ML, en conjunto con una multitud de herramientas útiles en genética evolutiva.

Figura 1.
Algoritmo "prunning" para calcular la verosimilitud de un árbol *T* en una posición (o columna) *k* del alineamiento *D*. Aquí observamos los nucleótidos A, G y G para las especies E1, E2 y E3, respectivamente. Los nucleótidos en los nodos interiores E_{12} y E_{123} se desconocen, pero eso no impide calcular su verosimilitud, ya que se pueden sumar las cuatro posibilidades correspondientes a los cuatro nucleótidos posibles.

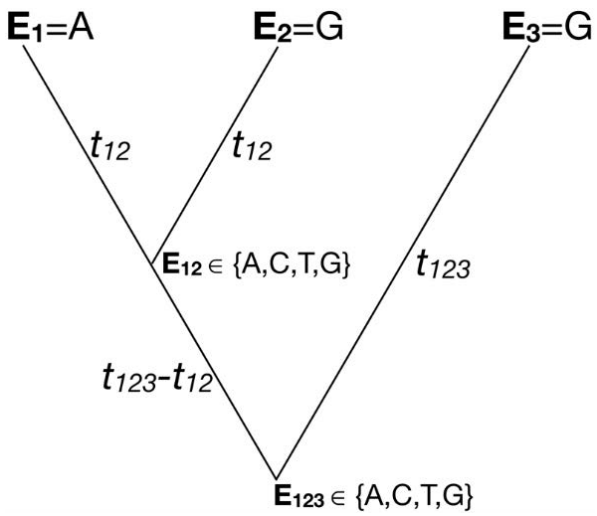
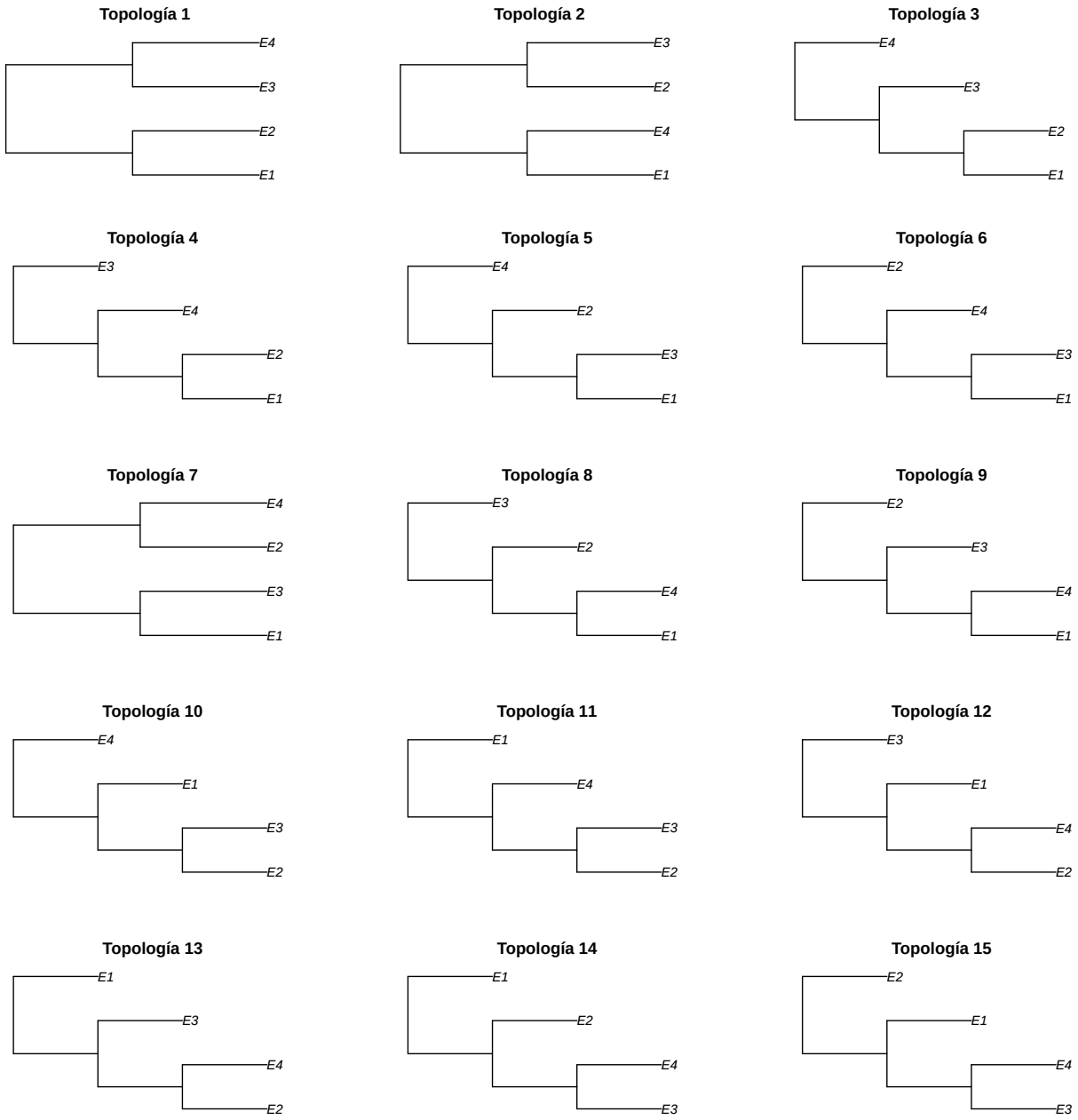


Figura 2.

Topologías posibles para el alineamiento *D* del ejemplo en la sección 4. La topología con la máxima verosimilitud es la 7.



Referencias

Brocchieri, L. (2001). Phylogenetic Inferences from Molecular Sequences: Review and Critique. *Theoretical Population Biology, 59*(1), 27-40. doi: 10.1006/tpbi.2000.1485

Bronham, L. & Penny, D. (2003). The modern molecular clock. *Nature Reviews Genetics,* 4, 216-224. doi: 10.1038/nrg1020

Duchen, P., Alfaro, M., Rolland, J., Salamin, N. & Silvestro, D. (2020). On the effect of asymmetrical trait inheritance on models of trait evolution. *Systematic Biology* (In press). doi: 10.1093/sysbio/syaa055

Edwards, A. & Cavalli-Sforza, L. (1963). The reconstruction of evolution. *Annals of Human Genetics,* 27, 106-106.

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology,* 22, 240-249. doi: 10.1093/sysbio/22.3.240

Felsenstein, J. (2004). *Inferring phylogenies*, vol. 2. Sunderland, Massachusetts: Sinauer associates.

Felsenstein, J. (2019). *PHYLIP (phylogeny inference package) version 3.698*. Recuperado de https://evolution.genetics.washington.edu/phylip.html

Felsenstein, J. & Churchill, G. A. (1996). A Hidden Markov Model approach to variation among sites in rate of evolution. *Molecular Biology and Evolution,* 13, 93-104. doi: 10.1093/oxfordjournals.molbev.a025575

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W. & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* 59, 307-321. doi: 10.1093/sysbio/syq010

Hall, B. G. (2013). Building phylogenetic trees from molecular data with MEGA. *Molecular Biology and Evolution*, 30, 1229-1235. doi: 10.1093/molbev/mst012

Hasegawa, M., Kishino, H. & Yano, T.-a. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22, 160-174. doi: 10.1007/BF02101694

Jukes, T. H. & Cantor, C. R. (1969). Evolution of protein molecules. En M. Munro (Ed.), *Mammalian protein metabolism* (pp. 21-132), vol. 3. New Yor: Academic Press.

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution,* 16, 111-120. doi: 10.1007/BF01731581

Lanave, C., Preparata, G., Sacone, C. & Serio, G. (1984). A new method for calculating evolutionary substitution rates. *Journal of Molecular Evolution,* 20, 86-93. doi: 10.1007/BF02101990

Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature,* 401, 877-884. doi: 10.1093/sysbio/syr124

Pattengale, N., Alipour, M., Bininda-Emonds, O., Moret, B. & Stamatakis, A. (2010). How many bootstrap replicates are necessary? *Journal of Computational Biology,* 17, 337-354. doi: 10.1089/cmb.2009.0179

Peña, C. (2011). Métodos de inferencia filogenética. *Revista Peruana de Biología*, 18, 265-267.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30, 1312-1313. doi: 10.1093/bioinformatics/btu033

Swofford, D. L. (2002). PAUP: phylogenetic analysis using parsimony, version 4.0 b10. Doi: 10.1111/j.0014-3820.2002.tb00191.x

Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular Biology and Evolution*, 10, 512-526. doi: 10.1093/oxfordjournals.molbev.a040023

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on Mathematics in the Life Sciences,* 17, 57-86.

"Ordo"
Cerámica/engobes y esmalte
15 x 25 x 25 cm
2016

"Mural Lisis"
Cerámica/esmaltes y engobes
120 x 45 x 5 cm
2010

Tequio *4*(11), 2021: 81-89
ISSN: 2594-0546

Métodos de reconstrucción filogenética II: inferencia bayesiana

Methods for phylogenetic reconstruction II: Bayesian inference

Pablo Duchen[1*]

Fecha de recepción: 6 de noviembre de 2020
Fecha de aceptación: 22 de enero de 2021

Resumen - La inferencia bayesiana como modelo de reconstrucción filogenética es muy utilizada en la actualidad. La ventaja de este método es la generación directa de probabilidades posteriores para cada clado en la filogenia final, por lo cual no se requiere de bootstrapping como medida de incertidumbre. Además, la inferencia bayesiana se presta perfectamente para la datación de filogenias por medio de relojes moleculares. En este trabajo se describen los principios de este método, comenzando por el teorema de Bayes; posteriormente se caracteriza el uso del algoritmo de Metropolis-Hastings para el muestreo de las topologías más probables y se le ilustra con un ejemplo sencillo. Se finaliza mencionando los programas más usados actualmente.

Palabras clave: Teorema de Bayes, Metropolis-Hastings, MCMC.

Abstract - Phylogenetic reconstruction through Bayesian inference is currently widely used. The main advantage of this method is the direct output of posterior probabilities for each clade on the final phylogeny. Thus, it does not require bootstrapping as a measure of uncertainty. Moreover, Bayesian inference is perfectly fit for dating phylogenies through molecular clocks. In this paper, the basics of Bayesian inference applied to phylogenetic reconstruction are described, starting with an explanation of Bayes' theorem. Then, the use of the Metropolis-Hastings algorithm to sample topologies from the posterior distribution is characterized and illustrated through a simple example. At the end, there is a mention of the software used for Bayesian phylogeny reconstruction.

Keywords: Bayes' theorem, Metropolis-Hastings, MCMC.

Introducción

Se presenta ahora el método bayesiano para reconstrucción filogenética. Como se verá a continuación, algunos elementos (como el cálculo de verosimilitudes y los modelos de mutación de ADN) también se utilizarán aquí; para no repetir su descripción, el lector deberá referirse a la primera parte de esta revisión.

En la inferencia bayesiana para la reconstrucción filogenética, el objetivo es encontrar el árbol con la mayor probabilidad posterior. Dicho cálculo va a depender de asumir una probabilidad a priori para cada árbol

[1] Departamento de Biología Computacional, Universidad de Lausana, Suiza.
Correos electrónicos: pablo.duchenbocangel@unil.ch, pduchen@gmail.com. ORCID: 0000-0002-9318-5002

y del uso de métodos markovianos (cadenas markovianas). De manera general, las cadenas markovianas son procesos estocásticos que describen una secuencia de eventos donde la probabilidad de un evento actual depende únicamente del anterior. En la inferencia bayesiana las cadenas markovianas se emplean para explorar el espacio de filogenias posibles.

La inferencia filogenética por medio de métodos bayesianos fue introducida por Rannala & Yang (1996), mientras que las extensiones markovianas fueron agregadas independientemente por Yang & Rannala (1997); Mau & Newton (1997) y Li, Pearl & Doss (2000). La base fundamental de toda inferencia bayesiana radica en el teorema de Bayes, el cual describimos a continuación.

Teorema de Bayes

Se debe iniciar por definir nuestros datos y parámetros a estimar. Dado un alineamiento *D* de secuencias de ADN para un número *n* de especies, el objetivo es encontrar el árbol *T* que mejor describa a dicho alineamiento. En esta revisión se usarán los términos filogenia y árbol indistintamente para referirse a *T*, al igual que los términos alineamiento o datos para aludir a *D*.

El teorema de Bayes en inferencia filogenética se presta fácilmente para calcular *T* dado un alineamiento *D*. Bajo este teorema, la probabilidad posterior de *T* es:

$$P(T|D) = \frac{P(T)P(D|T)}{P(D)}, \quad (1)$$

donde *P(T)* es la probabilidad *a priori* del árbol, *P(D|T)* es la verosimilitud (también conocida como *likelihood*) y *P(D)* es la probabilidad del alineamiento. Para fines prácticos, *P(D)* constituye la sumatoria del numerador *P(T)P(D|T)* sobre todas las posibles topologías *T*:

$$P(T|D) = \frac{P(T)P(D|T)}{\sum_T P(T)P(D|T)}. \quad (2)$$

En otras palabras, al sumar *P(T)P(D|T)* para todos los *T* posibles obtenemos *P(D)*.

Aplicación del teorema de Bayes en filogenética

En la práctica no es posible calcular el denominador de la ecuación (2), ya que la cantidad de topologías posibles (la manera en que las especies se agrupan) incrementa exponencialmente con el número de especies *n*. Repitiendo el mismo ejemplo de la primera parte, un alineamiento con tres especies tiene tres topologías posibles; cuatro especies, 15 topologías posibles; cinco especies tienen 105 topologías, y si hablamos de un alineamiento de 50 especies –muy común en estudios biológicos– tendríamos $2{,}75\times10^{76}$ topologías posibles. Por tanto, computacionalmente no es realista calcular la verosimilitud de tal cantidad de árboles.

Metropolis-Hastings

Para solucionar este problema se usa el algoritmo de Metropolis-Hastings (Metropolis *et al.*, 1953; Hastings, 1970), el cual se basa en una cadena markoviana de Monte Carlo (MCMC, por sus siglas en inglés). Dicho algoritmo explora en el espacio de topologías y toma una muestra representativa de la distribución *P(T|D)*, que es la distribución posterior de la cual queremos obtener *T*. En otras palabras, con Metropolis-Hastings se

analizan individualmente muchas topologías posibles y se toma una muestra de ellas; sin embargo, esta muestra no es aleatoria, más bien representativa, de la distribución posterior de topologías *P(T|D)*. Los pasos del algoritmo de Metropolis-Hastings son los siguientes:

1. Establecer un árbol inicial T_i, el cual constituye la topología "actual".
2. Modificar ligeramente la topología T_i y llamarla T_j (T_j ahora constituye el árbol "candidato").
3. Calcular la relación *A* entre las probabilidades posteriores de las topologías actual T_i y candidata T_j

$$A = \frac{P(T_j|D)}{P(T_i|D)}. \tag{3}$$

4. Si $A > 1$ aceptar T_j como el nuevo árbol actual. En caso contrario, aceptar T_j con probabilidad *A* y constituirlo en el nuevo T_i.
5. Volver al paso 2.

Está demostrado que repetir este algoritmo muchas veces resulta en una muestra significativa de árboles pertenecientes a *P(T|D)* (Metropolis *et al.*, 1953; Hastings, 1970). Ahora, aplicando el teorema de Bayes (1), la ecuación (3) se puede reescribir como:

$$A = \frac{\frac{P(T_j)P(D|T_j)}{P(D)}}{\frac{P(T_i)P(D|T_i)}{P(D)}}, \tag{4}$$

y simplificando los denominadores:

$$A = \frac{P(T_j)P(D|T_j)}{P(T_i)P(D|T_i)}. \tag{5}$$

Por ende, el tomar una muestra representativa de *T* a partir de la distribución objetivo *P(T|D)* se reduce a poder calcular la verosimilitud de topologías actuales $P(D|T_i)$ y candidatas $P(D|T_j)$, y de asumir probabilidades *a priori* para cada una de dichas topologías. En muchos casos se asume que la probabilidad *a priori* de cada topología es la misma, por lo que éstas también se simplificarían. No obstante, es igualmente posible asignar probabilidades *a priori* para las longitudes de rama de *T* a partir de una distribución exponencial, tal es el caso del programa MrBayes (Huelsenbeck & Ronquist, 2001). En cuanto a la verosimilitud *P(D|T)*, el cálculo se efectúa utilizando el algoritmo "prunning" (Felsenstein, 1973).

Pasos generales para inferir una filogenia bajo un modelo bayesiano

Como se estableció anteriormente, la inferencia bayesiana de filogenias requiere también de estimar verosimilitudes de distintas topologías. Combinando lo que se describió en la primera parte de esta revisión con los métodos descritos aquí, los pasos generales para la inferencia bayesiana de filogenias son los siguientes:

1. Proponer una topología inicial.
2. A partir del alineamiento observado *D* y la topología propuesta en el paso 1, calcular la verosimilitud para cada posición (o columna) de *D* utilizando el algoritmo "prunning". Para conocer las probabilidades de sustitución nucleotídica utilizadas en dicho algoritmo, referirse a la sección "Modelos de mutación de ADN" en la primera parte de esta revisión.
3. Una vez obtenida la verosimilitud en cada posición de *D*, calcular la verosimilitud total por medio de la ecuación (2) de la primera parte de esta revisión.
4. Teniendo la verosimilitud de *T*, proponer una topología candidata (similar a la topología actual), calcular su verosimilitud de forma similar siguiendo los pasos 2 y 3, y calcular la relación *A* con la ecuación (5).
5. Si $A > 1$, aceptar la topología candidata como nuevo árbol actual. Caso contrario, aceptar la topología candidata con probabilidad *A*.
6. Repetir los pasos 2 a 5 hasta haber obtenido una muestra representativa de árboles de *P(T|D)*.

Generación de la topología final a partir de la muestra de *P(T|D)*

Es importante describir un paso más para finalizar la reconstrucción de una filogenia con el método bayesiano descrito aquí. Hasta ahora hemos logrado una muestra representativa de árboles de la distribución *P(T|D)*, pero ¿cuál de todas esas topologías se reporta al final? Una forma de abordar este problema consiste en estimar distancias entre todos los árboles de la muestra y tomar como representante al que se encuentre al medio (Li *et al*., 2000; Critchlow, Pearl & Qian, 1996). Otra posibilidad consiste en observar la frecuencia de cada clado en la muestra total de topologías y reportar todas las especies en los clados donde estén más frecuentes (Huelsenbeck, Ronquist, Nielsen & Bollback, 2001; Larget & Simon, 1999).

Finalmente, en cuanto a las especificaciones del Metropolis-Hastings MCMC, es conveniente prestar atención a la frecuencia con que se toman las muestras de *P(T|D)*. Como primer punto, es bueno descartar la primera parte de muestras, ya que no todas ellas pertenecerán a *P(T|D)* (éstas corresponden al *burn-in*). En segundo lugar, no conviene mantener a todas las topologías candidatas aceptadas, ya que en muchos casos serán muy parecidas; es mejor guardar los árboles cada cierto número de repeticiones del algoritmo de Metropolis-Hastings, para así obtener una muestra más representativa de *P(T|D)* (Huelsenbeck *et al*., 2001; Felsenstein, 2004). Es importante notar que para la inferencia bayesiana de filogenias no es necesario utilizar *bootstrapping* como medida de incertidumbre, ya que la probabilidad posterior *P(T|D)* cumple con esta función. La filogenia final reportada contiene probabilidades posteriores para cada clado de la filogenia y cada uno de estos valores describe la probabilidad del clado en cuestión.

Ejemplo de algoritmo para inferencia bayesiana

Enseguida se desarrolla un ejemplo muy sencillo para inferir una filogenia usando el algoritmo de Metropolis-Hastings para inferencia bayesiana. Al igual que en la primera parte de esta revisión, utilizaremos el alineamiento del ejemplo 1.1.1 del artículo "Modelo de estimación de pesos de árbol filogenético para un cuartet, aplicando conjugación de Hadamard" (publicado también en este número). Dicho alineamiento contiene cuatro especies y 16 posiciones. Convertimos dicho alineamiento a formato FASTA y lo guardamos en un archivo denominado "alineamiento.fas":

```
>E1
CCATCAAACGTGTGAC
```

```
>E2
ACAGCAATGTTATCTC
>E3
CCATTGAAGATGCGTT
>E4
ACAGTAGTGTTACCAG
```

Posteriormente, consideramos posibles topologías para las especies E1, E2, E3 y E4. En total existen 15 posibles topologías, las cuales las escribimos en formato NEWICK y las guardamos en un archivo denominado "topologias.tre" (visualizadas en la Figura 1):

```
( ( E1 : 1, E2 : 1 ) : 1, ( E3 : 1, E4 : 1 ) : 1 ) : 1;
( ( E1 : 1, E4 : 1 ) : 1, ( E2 : 1, E3 : 1 ) : 1 ) : 1;
( ( ( E1 : 1, E2 : 1) : 1, E3 : 1 ) : 1, E4 : 1 ) : 1;
( ( ( E1 : 1 , E2 : 1 ) : 1, E4 : 1 ) : 1, E3 : 1 ) : 1;
( ( ( E1 : 1, E3 : 1 ) : 1, E2 : 1 ) : 1, E4 : 1 ) : 1;
( ( ( E1 : 1, E3 : 1 ) : 1, E4 : 1 ) : 1, E2 : 1 ) : 1;
( ( E1 : 1, E3 : 1 ) : 1, ( E2 : 1, E4 : 1 ) : 1 ) : 1;
( ( ( E1 : 1, E4 : 1 ) : 1, E2 : 1 ) : 1, E3 : 1 ) : 1;
( ( ( E1 : 1, E4 : 1 ) : 1, E3 : 1 ) : 1, E2 : 1 ) : 1;
( ( ( E2 : 1, E3 : 1 ) : 1, E1 : 1 ) : 1, E4 : 1 ) : 1;
( ( ( E2 : 1, E3 : 1 ) : 1, E4 : 1 ) : 1, E1 : 1 ) : 1;
( ( ( E2 : 1, E4 : 1 ) : 1, E1 : 1 ) : 1, E3 : 1 ) : 1;
( ( ( E2 : 1, E4 : 1 ) : 1, E3 : 1 ) : 1, E1 : 1 ) : 1;
( ( ( E3 : 1, E4 : 1 ) : 1, E2 : 1 ) : 1, E1 : 1 ) : 1;
( ( ( E3 : 1, E4 : 1 ) : 1, E1 : 1 ) : 1, E2 : 1 ) : 1;
```

Finalmente, usando los archivos que se acaban de crear ("alineamiento.fas" y "topologias.tre") como input, desarrollo a continuación un programa corto (escrito en el lenguaje de programación R) a fin de ejemplificar el algoritmo Metropolis-Hastings para la inferencia filogenética bayesiana (nótese que los pasos del algoritmo también están mencionados en el programa):

```
#######-INICIO DEL PROGRAMA-#######

library(ape)
library(phangorn)

#Alineamiento de ADN.
D <- phyDat(read.FASTA("alineamiento.fas")) #Posibles topologias.
T <- read.tree("topologias.tre")
```

```
##-------Metropolis-Hastings para inferencia bayesiana-------## cat("\nEjemplo Metropolis-Hastings para inferencia Bayesiana\n")
#Se comienza con la topología inicial (paso 1).
indice_actual <- 1
T_actual <- T[ [ indice_actual ] ]

for (i in 2:length(T) ) {
        #La topología candidata es la siguiente en la lista de T (paso 2). T_candidata <- T[[i]]

        #Se calcula la verosimilitud de la topología actual.
        logV_actual <- pml(T_actual,D)

        #Se calcula la verosimilitud de la topología candidata.
        logV_candidata <- pml(T_candidata,D)

        #Se calcula la relacion de verosimilitudes A (paso 3).
        A <- logV_actual$logLik/logV_candidata$logLik

        if (A>1) { #Si A>1 (paso 4 parte 1).
                T_actual <- T_candidata
                indice_actual <- i
                print(paste("T candidata =",i,", log V = ",logV_candidata$logLik))
        } else {
                #Si A<1 (paso 4 parte 2).
                nuevoIndice <- sample(c(indice_actual,i),1,prob=c(A,1-A)) T_actual <- T[[nuevoIndice]]

                if (nuevoIndice==indice_actual) {
                        print(paste("T actual =",indice_actual,", log V = ",logV_actual$logLik))
                } else {
                        print(paste("T candidata =",indice_actual,", log V = ",logV_candidata$logLik))
                }
                        indice_actual <- nuevoIndice
        }
}

#######-FIN DEL PROGRAMA-#######
```

El output de este programa es el siguiente:

```
Ejemplo Metropolis-Hastings para inferencia Bayesiana
[1]  "T actual = 1 , log Verosimilitud = -87.1232112880638"
[1]  "T candidata = 3 , log Verosimilitud = -87.0763445725479"
```

[1] "T candidata = 4 , log Verosimilitud = -87.0328449489164"
[1] "T candidata = 5 , log Verosimilitud = -86.6093614372388"
[1] "T actual = 5 , log Verosimilitud = -86.6093614372388"
[1] "T candidata = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"
[1] "T actual = 7 , log Verosimilitud = -85.6551750661665"

Como se puede corroborar, el algoritmo de Metropolis-Hastings converge en la topología 7; sin embargo, cabe recordar que aquí usamos Metropolis-Hastings solamente para desarrollar un ejemplo ilustrativo. En realidad, para un alineamiento tan reducido serán suficientes otros métodos basados en distancias genéticas para inferir la filogenia. La máxima verosimilitud y la inferencia bayesiana usando Metropolis-Hastings MCMC son más útiles para un mayor número de especies en alineamientos más grandes. En tales casos, no existirá solamente una topología con la mayor probabilidad posterior, sino varias, por lo cual se deberán utilizar los métodos descritos en la sección "Generación de la topología final a partir de la muestra de *P(T|D)*" para obtener el resultado final. Finalmente, dependiendo de la cantidad de especies y de la longitud del alineamiento se prefiere un método sobre otro (e. g. Inferencia bayesiana sobre máxima verosimilitud, o viceversa). Si se desea revisar más detalladamente los factores que influyen en la selección de un método de inferencia filogenética, remitirse a Peña (2011).

Software para inferencia bayesiana de filogenias

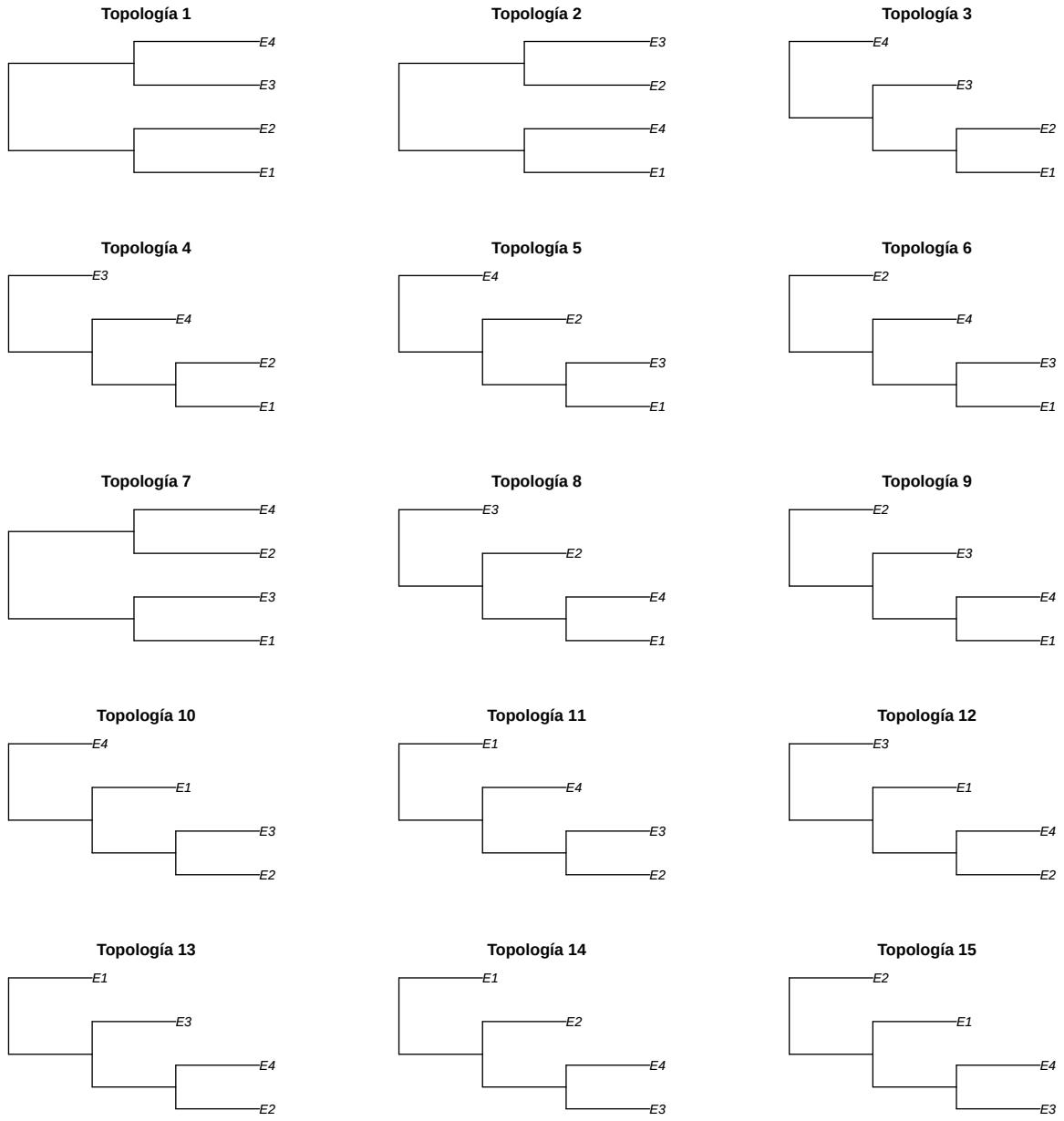
Existen muchos programas que hacen inferencia bayesiana de filogenias. Aquí se nombran los más usados en la actualidad:

MrBayes. Este programa fue desarrollado inicialmente por Huelsenbeck & Ronquist (2001), con una nueva versión publicada más recientemente (Ronquist *et al.*, 2012). Este es el programa clásico empleado en reconstrucción filogenética con un método bayesiano. Utiliza todos los elementos descritos aquí: probabilidades *a priori*, modelos de mutación de ADN y Metropolis-Hastings MCMC para encontrar el árbol con la mayor probabilidad posterior.

BEAST. Este programa fue originalmente introducido por Drummond & Rambaut (2007). Además de encontrar el árbol con la mayor probabilidad posterior, BEAST incluye rutinas que calculan un reloj molecular (o datación de filogenias) bajo distintos modelos.

RevBayes. Desarrollado por Höhna *et al.* (2016), sus funciones comprenden inferencia filogenética, MCMC, relojes moleculares, selección de modelos, estimación de tasas de diversificación, etcétera. Este programa incluye su propio intérprete, lo que lo hace particularmente útil a la hora de desarrollar tareas más complejas.

Figura 1.
Topologías posibles para el alineamiento *D* del ejemplo en la sección "Ejemplo de algoritmo para inferencia bayesiana". La topología con la máxima verosimilitud es la 7.



Referencias

Critchlow, D. E., Pearl, D. K. & Qian, C. (1996). The triples distance for rooted bifurcating phylogenetic trees. *Systematic Biology,* 45, 323-334. doi: 10.1093/sysbio/45.3.323

Drummond, A. J. & Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology,* 7, 214. doi: 10.1186/1471-2148-7-214

Felsenstein, J. (1973). Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Biology,* 22, 240-249. doi: 10.1093/sysbio/22.3.240

Felsenstein, J. (2004). *Inferring phylogenies*, vol. 2. Sunderland, Massachusetts: Sinauer associates.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97-109. doi: 10.1093/biomet/57.1.97

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., Huelsenbeck, J. P. & Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology,* 65, 726-736. doi: 10.1093/sysbio/syw021

Huelsenbeck, J. P. & Ronquist, F. (2001). MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17, 754-755. doi: 10.1093/bioinformatics/17.8.754

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, 294, 2310-2314. doi: 10.1126/science.1065889

Larget, B. & Simon, D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*, 16, 750-759. doi: 10.1093/oxfordjournals.molbev.a026160

Li, S., Pearl, D. K. & Doss, H. (2000). Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American statistical Association*, 95, 493-508. doi: 10.1080/01621459.2000.10474227

Mau, B. & Newton, M. A. (1997). Phylogenetic inference for binary data on dendograms using Markov chain Monte Carlo. *Journal of Computational and Graphical Statistics*, 6, 122-131. doi: 10.1080/10618600.1997.10474731

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics,* 21, 1087-1092. doi: 10.1063/1.1699114

Peña, C. (2011). Métodos de inferencia filogenética. *Revista Peruana de Biología* 18, 265-267.

Rannala, B. & Yang, Z. (1996). Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. *Journal of Molecular Evolution*, 43, 304-311. doi: 10.1007/BF02338839

Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A. & Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61, 539-542. doi: 10.1093/sysbio/sys029

Yang, Z. & Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences: A Markov chain Monte Carlo method. *Molecular Biology and Evolution,* 14, 717-724. doi: 10.1093/oxfordjournals.molbev.a025811

"Semilla voladora I"
Cerámica/Raku
28 x 13 x 15 cm
2017

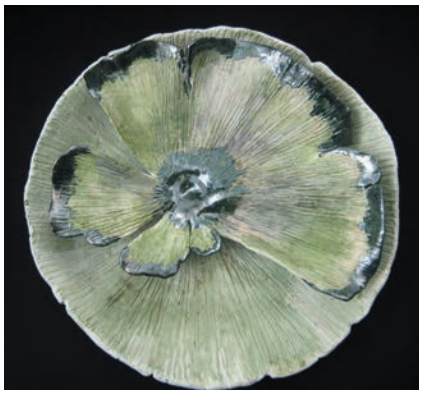
"Semilla voladora II"
Cerámica/Raku
27 x 10 x 10
2017

"Semillas voladoras"
Cerámica/Raku
Medidas varias
2017

OBRA GRÁFICA DE ALMALUZ GUZMÁN

Almaluz Guzmán es licenciada en Artes Plásticas y Visuales por la Universidad Autónoma "Benito Juárez" de Oaxaca. Se ha formado también en otras instituciones, como en el Taller Rufino Tamayo, Centro de las Artes San Agustín, Instituto de Artes Visuales de la Ciudad de Puebla y la Academia de San Carlos. Ha participado en múltiples exposiciones a nivel local, nacional e internacional. Es colaboradora del Colectivo de Ilustradores de la Ciencia y la Naturaleza de México (CICYNM), así como integrante del Colectivo Redox cerámica y alalimón.

Su trabajo se desarrolla principalmente en el área tridimensional en diferentes medios, entre ellos la escultura y la gráfica; usa e investiga diferentes materiales, como cerámica, papel, pigmentos naturales, etcétera. Su enfoque temático es diverso; sin embargo, recae esencialmente en la representación de elementos cósmicos y biológicos.



"Macrófita"
Cerámica/engobes y esmalte
4.5 x 28.5 x 30 cm
2016

"Diatomea II"
Raku
18 x 17 x 15 cm
2014

"Prototipo A001"
Cerámica/engobes
35 x 25 x 25 cm
2008

"Prototipo AB"
Cerámica/engobes
32 x 20 x 20 cm
2009

"Prototipo H"
Cerámica/esmaltes y engobes
45 x 28 x 28 cm
2009

"Prototipo Z"
Cerámica/esmaltes y engobes
45 x 28 x 28 cm
2009

"Iris/radiales"
Cerámica /esmalte
12 x 60 x 60 cm
2014

"Tricomas"
Cerámica/engobes y esmalte
9.5 x 14 x 14 cm
2016

"Semilla voladora"
Cerámica/Raku
20 x 16.5 x 5 cm
2017

"Origen Mural"
Cerámica/esmaltes y engobes
165 x 48 x 6 cm
2010

"Aeonis"
Cerámica/ esmaltes y engobes
Dimensiones varias
2014

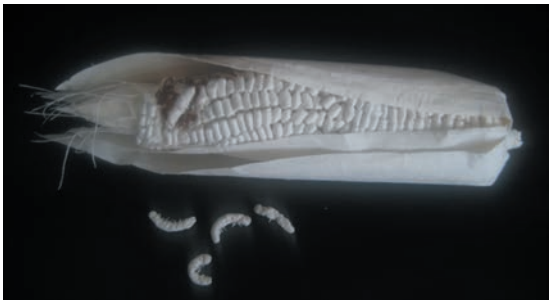
"Biocosmos"
Grafito sobre papel
40 x 50 cm
2013

"Diatomea I"
Cerámica/esmaltes y engobes
16.5 x 25 x 25 cm
2011

"Mandala
(Capullus)"
Raku
16.5 x 25 x 25 cm
2014

"Estructura"
Cerámica/esmalte
42 x 30 x 13 cm
2011

"Flor cósmica"
Raku
19 x 14.5 x 19 cm
2016



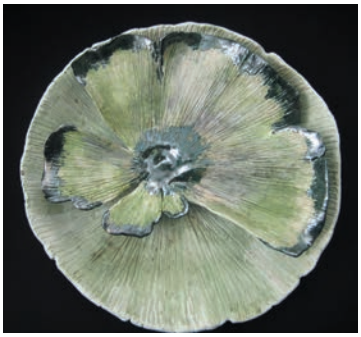
"Maíz MON 810"
Papel de algodón
30 x 13 x 7 cm
2014

"Pentapetalae"
Cerámica/
Esmalte
16 x 21.5 x 21.5 cm
2016

"Ordo"
Cerámica/engobes y esmalte
15 x 25 x 25 cm
2016

"Mural Lisis"
Cerámica/esmaltes y engobes
120 x 45 x 5 cm
2010

"Semilla voladora II"
Cerámica/Raku
27 x 10 x 10
2017



"Macrófita"
Cerámica/engobes y esmalte
4.5 x 28.5 x 30 cm
2016

"Semilla voladora I"
Cerámica/Raku
28 x 13 x 15 cm
2017

Normas editoriales para publicar en *TEQUIO*

TEQUIO. *Revista Interdisciplinaria de Investigación e Innovación* es una publicación cuatrimestral, editada y distribuida por la Universidad Autónoma "Benito Juárez" de Oaxaca.

Objetivo
Ser un espacio para difundir entre la comunidad universitaria y el público en general la investigación, las reflexiones teóricas y el conocimiento científico que se genera en diversas áreas del saber, en contextos regionales, nacionales e internacionales, desde una perspectiva interdisciplinaria de investigación e innovación.

Convocatoria de artículos
La convocatoria está dirigida a investigadores de las diferentes áreas del conocimiento, de la Universidad Autónoma Benito Juárez de Oaxaca y de la comunidad científica de México y el mundo.

TEQUIO recibe artículos originales e inéditos bajo convocatoria anual, por lo que los autores que contribuyan en ella deberán ajustarse a las siguientes normas:

1. La revista aceptará trabajos escritos en español o inglés, cuando sea la lengua nativa de los autores y/o tengan una lengua nativa diferente al español.
2. Los archivos deberán enviarse en formato Word 97-2013, en hoja tamaño carta, fuente Times New Roman de 12 puntos, con una extensión de 12 a 15 cuartillas (páginas), numeradas, al igual que los renglones. Los márgenes de la página deben ser de 2.5 cm para el superior e inferior, y 3 cm para los lados derecho e izquierdo, con un interlineado de 1.5.
3. En la redacción se respetarán las normas internacionales relativas a las abreviaturas, a los símbolos, a la nomenclatura anatómica, zoológica, botánica, química, a la transliteración terminológica, sistema de unidades, etcétera.
4. Todo trabajo deberá incluir las siguientes secciones, con las características especificadas.
4.1 En la primera página:
a. Título del trabajo en español e inglés. El título deberá ser tan corto como sea posible, siempre que contenga las palabras clave del trabajo, de manera que permita identificar la naturaleza y contenido de éste, aun cuando se publique en citas e índices bibliográficos. No se deben utilizar abreviaturas.
b. Nombre completo del o los autores, iniciando con el (los) nombre(s), apellido paterno apellido materno ejemplo: Andrés Hernández Scandy, Mariana Tafoya-Parra. El autor de correspondencia debe estar identificado con un asterisco e incluir su correo electrónico.
c. Institución a la que representan, sin abreviaturas y la dirección completa de la misma (en una nota a pie), especificando el país.
4.2 Resumen y abstract con un máximo de 250 palabras. A continuación de cada resumen se anotarán de tres a cinco palabras o frases cortas-clave (Key words), que ayuden a clasificar el artículo.
4.3 Notas a pie de página: a 10 puntos con las mismas características que el cuerpo del texto, deberán ser únicamente aclaratorias o explicativas, sólo servirán para ampliar o ilustrar lo dicho en el cuerpo del texto.
4.4 El trabajo puede incluir fotografías, gráficos, cuadros y mapas que ilustren el contenido, en el texto se debe mencionar dónde se insertarán las mismas y deberán enviarse por separado de manera electrónica y con sus respectivas fuentes de información.
4.5 Se recomienda presentar cada cuadro y figura en hojas separadas; los cuadros deberán estar numerados, tener título o leyenda explicativa, de manera que se comprendan por sí mismos sin necesidad de leer el texto.
a. Se entiende por cuadro al conjunto de nombres, cifras u otros datos presentados ordenadamente en columnas o renglones, de modo que se advierta la relación existente entre ellos. Deberán ser enviados en archivos individuales, en formato Word, con líneas horizontales y verticales, a fin de que pueda corregirse la ortografía o modificar su tamaño.
b. Las figuras (gráficas, dibujos, etcétera) deberán enviarse en los programas Excell para Windows, Corel Draw o Harvard Graphics, y presentarse en archivos individuales con el número progresivo correspondiente y pie de figura que la explique.
c. Las fotografías deberán ser enviadas en archivos individuales con alta resolución (300 pixeles por pulgada), en formatos gif; tiff, jpg. Se deben especificar los diámetros de aumento en las microfotografías que se incluyan.
4.6 Citas y referencias: al final del texto, las referencias deben separarse de acuerdo con el tipo de material que se consulta: bibliografía, hemerografía, referencias electrónicas, etcétera, en orden alfabético.

La forma de citar dentro del texto se apegará al formato APA 2016: entre paréntesis se anotará el primer apellido del autor o autores, separado con una coma del año de la publicación citada, luego una coma y la abreviatura "p.", y enseguida la página de donde fue tomada la cita: (Castañón, 2014, p. 25).

En caso de que sólo se mencione algún trabajo de otro autor o no se trate de una cita textual, se deberá anotar de esta forma: (Castañón, 2014) o bien dentro de la redacción: Como afirma Castañón (2014)...
Las referencias se consignarán de la siguiente forma:

Artículo impreso:
Apellido, A. A., Apellido, B. B. & Apellido, C. C. (Año). Título del artículo. *Título de la publicación,* volumen (número), pp-pp.
Libro:
Apellido, A. A. (Año). *Título*. Ciudad: Editorial.
Capítulo de libro:
Apellido, A. A. & Apellidos, A. A. (Año). Título del capítulo. En A. A. Apellido (Ed., Coord., etc.), Título del libro (pp-pp). Ciudad: Editorial.
Versión electrónica de libro impreso:
Apellido, A A (Año). *Título.* Recuperado de http://www.ejemplo.com
Simposios y conferencias:
Apellido, A., & Apellido, A. (mes, año). Título de la presentación. En A. Apellido del Presidente del Congreso (Presidencia), Título del simposio. Simposio dirigido por nombre de la Institución organizadora, lugar.
Tesis:
Apellido, A. & Apellido, A. (Año). *Título de la tesis* (Tesis de pregrado, maestría o doctoral). Nombre de la institución, lugar. Recuperado de www.ejemplo.com

5. La comisión editorial enviará los artículos que reciba a arbitraje con dos pares externos de reconocido prestigio nacional e internacional.
6. Si el artículo fue aceptado con correcciones y/o adaptaciones, éste deberá ser devuelto corregido a la revista en un plazo no mayor a 15 días naturales.
7. El dictamen final será inapelable. Los autores serán contactados vía correo electrónico.
8. ***TEQUIO*** solicitará una carta firmada por todos los coautores, en la que declaren estar de acuerdo con que su artículo sea publicado en la revista. En caso de ser coautores, indicarán en qué consistió su participación.
9. Los artículos contenidos en esta revista serán responsabilidad exclusivamente de los autores.
10. Cualquier circunstancia no contemplada en la presente convocatoria será resuelta por el Comité Editorial de ***TEQUIO.***

Tequio

Revista de Divulgación, Investigación e Innovación