Universidad Católica del Ecuador

Facultad de Ingeniería

Maestría en Biología Computacional

USO DE FILOGENÓMICA PARA LA IDENTIFICACIÓN DE PROTEÍNAS ASOCIADAS A LA RUTA BIOSINTÉTICA DE ALCALOIDES ANTICANCERÍGENOS EN PLANTAS

Proyecto de Titulación

Ángel Sebastián Rodríguez Pazmiño

Asesor

Dr. Alan Cervantes

Lectores

Dr. Daniel Chavez

Mtr. Laura González

Quito - Ecuador

Junio del 2023

ÍNDICE GENERAL

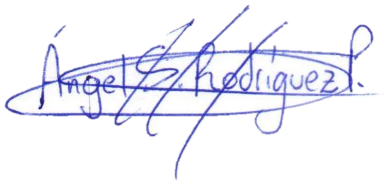
ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

DERECHOS DE AUTOR

Yo, **Ángel Sebastián Rodríguez Pazmiño**, con cédula de identidad **1804369518**, declaro que las ideas, juicios, valoraciones, interpretaciones, consultas bibliográficas, definiciones y conceptualizaciones expuestas en el presente trabajo; así como, los procedimientos y herramientas utilizadas en la investigación, son de absoluta responsabilidad del autor del proyecto de titulación. Así mismo, me acojo a los reglamentos internos de la Pontificia Universidad Católica del Ecuador.

Quito, julio 2023



Ángel Sebastián Rodríguez Pazmiño
C.I.: 1804369518

DEDICATORIA

Dedico este trabajo a las personas en situación de calle y a los vendedores ambulantes con los que me atravieso casi todos los días. Su esfuerzo diario por llevar un pan a la mesa de su hogar, a pesar de sus mil y un limitaciones, hace que valore lo que tengo, y aspire a llegar a tener una estabilidad que me permita ayudarles significativamente, aunque sea a unos pocos.

Por otro lado, debo reconocer a Miguel Ángel García, Ph.D. Él es un gran ser humano y excelente académico que fue mi maestro y tutor de tesis en el pregrado de Biología. Ahora, se repetirá esta historia. Él será mi supervisor de doctorado en estudios relacionados a la bacteria que produce la tuberculosis.

AGRADECIMIENTO

Quiero agradecer afectuosamente al profesor Sergio Alan Cervantes, Ph.D., por su apoyo en este trabajo. Aunque no fueron muchas las ocasiones en que platicamos, valoro cada minuto de su guía y de las discusiones que tuvimos. Fue muy enriquecedor el haber podido recibir algo de su vasta experiencia.

Espero que las circunstancias académicas nos permitan seguir colaborando.

RESUMEN

Vinblastina y vincristina son anticancerígenos de gran importancia para la industria farmacéutica. Actualmente su proceso de obtención es altamente complejo y costoso debido, por una parte, a la baja concentración en la planta *Catharanthus roseus*, siendo este el único organismo en el cual se han encontrado. Por ello, se han intentado diversas estrategias biotecnológicas para inducir una mayor producción de estos fármacos a un menor costo. En este trabajo, utilizando criterios filogenómicos y diferentes herramientas bioinformáticas, se explora la potencial presencia de los alcaloides vinblastina y vincristina en un grupo de 72 plantas, usando como secuencias de referencia cuatro proteínas asociadas a la ruta de biosíntesis de estos alcaloides en la planta *Catharanthus roseus*. Un enfoque de investigación como el propuesto en este proyecto, podría sugerir su presencia en una especie de planta distinta; impulsando, posteriormente, investigaciones de "wet lab" que confirmen o descarten su existencia.

Palabras claves: filogenómica, vinblastina, vincristina, *Catharanthus roseus*, plantas medicinales, herramientas bioinformáticas.

ABSTRACT

Vinblastine and vincristine are anticancer drugs of outstanding importance to the pharmaceutical industry. Currently, their production process is highly complex and expensive due, in part, to their low concentration in the *Catharanthus roseus* plant, which is the only organism in which they have been found. Therefore, various biotechnological strategies have been attempted to induce greater production of these drugs at a lower cost. In this work, using phylogenomic criteria and different bioinformatics tools, is explored the potential presence of the vinblastine and vincristine alkaloids in a group of 72 plants, using four reference protein sequences associated with the biosynthesis pathway of these alkaloids in the *Catharanthus roseus* plant. An investigative approach like the one proposed in this project could suggest their presence in a different plant species, subsequently prompting "wet lab" investigations to confirm or rule out their existence.

Keywords: phylogenomics, vinblastine, vincristine, *Catharanthus roseus*, medicinal plants, bioinformatics tools.

1. Introducción

Los alcaloides desempeñan un papel fundamental en la medicina humana y en la defensa natural de diversos organismos (Heinrich et al., 2021). En el caso de las plantas, estas sustancias químicas comprenden aproximadamente el 20% de los metabolitos secundarios conocidos (Kaur & Arora, 2015), cuyas funciones principales son proteger a las plantas de predadores y regular su crecimiento (Chik et al., 2013). Para la salud humana, los alcaloides pueden servir como anestésicos, cardioprotectores, y agentes anti-inflamatorios (Heinrich et al., 2021). En la actualidad, el interés por los compuestos naturales bioactivos ha aumentado debido a su potencial en el descubrimiento de fármacos y a un desarrollo muy proactivo sobre el estudio de los remedios tradicionales (etnofarmacología) (Atanasov et al., 2021). Hasta el 25 de octubre del 2020, 27 683 alcaloides habían sido registrados en el *Diccionario de Productos Naturales* (DNP, por sus siglas en inglés) con 990 nuevas inclusiones de alcaloides reportados como compuestos naturales entre el 2014 al 2020 (Heinrich et al., 2021).

Dentro de estos alcaloides de importancia, vinblastina y vincristina han sido exclusivamente identificados y producidos a partir de la planta *Catharanthus roseus* (*C. roseus*), encontrándose sólo en mínimas concentraciones (alrededor del 0.0005% de su peso seco), lo que vuelve su extracción compleja y costosa (Barrales-Cureño et al., 2019). El costo de extracción de 1 kg de vinblastina es de aproximadamente 1 millón de dólares americanos, mientras que para purificar la misma cantidad de vincristina hace falta un estimado de 3.5 millones de dólares americanos (Loyola-Vargas et al., 2004). Media tonelada de hojas secas son necesarias para la obtención de un solo gramo de vinblastina (Sottomayor et al., 2004) y para producir un kilogramo de vincristina se requieren 530 kg (*Digital Library Of The Commons*, s. f.). Además, su extracción se vuelve aún más complicada considerando que que *C. roseus* tiene 200 moléculas con propiedades químicas y físicas similares. Vinblastina y vincristina son muy valoradas en el mercado, ya que tienen bajos niveles de producción y son útiles para una variedad de condiciones médicas. Estos factores han fomentado el estudio de su biosíntesis y el desarrollo de técnicas alternativas de producción (Barrales-Cureño et al., 2019).

La biosíntesis de los alcaloides de *C. roseus* está sujeta a un riguroso control a nivel de células, tejidos y órganos, donde cuatro enzimas son claves: *tryptophan decarboxylase*

(TDC), *strictosidine synthase* (STR1), *desacetoxyvindoline 4-hydroxylase* (D4H) y *deacetylvindoline 4-O-acetyltransferase* (DAT) (Barrales-Cureño et al., 2019). Adicionalmente, esta biosíntesis depende en gran medida de las propias etapas de desarrollo de la planta y de factores ambientales (Barrales-Cureño et al., 2019). Varias investigaciones han tratado de comprender la regulación de ciertos genes que codifican para enzimas que participan en la síntesis de estos químicos y algunos de los mecanismos que regulan la expresión génica en cultivos de células en suspensión de *C. roseus* han sido identificados (van der Fits & Memelink, 2000).

La primera etapa en la biosíntesis de estos alcaloides es la formación de triptamina en una reacción catalizada por la enzima TDC (Goddijn et al., 1995). Otra etapa fundamental de la biosíntesis sucede cuando la triptamina se une al monoterpeno secologanina, compuesto final de la vía de biosíntesis de los iridoides, en una reacción catalizada por la enzima STR1 (Zhu et al., 2015). La condensación de triptamina con secologanina de glucósido iridoide bajo la catálisis de la sintasa de estrictosidina (STR) produce la formación de estrictosidina, el intermediario central en la biosíntesis de todos los tipos de alcaloides de indol (Zhu et al., 2015). Posteriormente, la estrictosidina se metaboliza a través de distintos pasos enzimáticos, considerando los catalizados por las enzimas D4H y DAT que conducen a la formación de vindolina y catarantina, los alcaloides monoterpénicos precursores de la vinblastina y la vincristina (De Luca et al., 1988).

La filogenómica es una de las herramientas que podrían utilizarse para estudiar las rutas biosintéticas de estos alcaloides en otros organismos. Esta es un área de estudio que había surgido, en primer lugar, en el contexto de la predicción de la función génica usando datos a nivel del genoma (Eisen, 1998), pero poco tiempo después también se utilizó para abarcar la inferencia filogenética en conjuntos de datos a esa escala (O'Brien & Stanyon, 1999). En investigaciones sobre plantas, la filogenómica ha sido una herramienta usada para esclarecer relaciones evolutivas entre, por ejemplo, angiospermas y gnetofitas (Ran et al., 2018) o para identificar proteínas de importancia en la evolución de plantas con semillas (Cibrián-Jaramillo et al., 2010). Además, el surgimiento de técnicas de secuenciación de alto rendimiento, ha ayudado al desarrollo del campo de la sistemática al habilitar el acceso a cantidades masivas de información de secuencias (*High Throughput*

Sequencing - an overview | ScienceDirect Topics, 2019), impulsando también investigaciones de carácter filogenómico.

Aunque para algunas cuestiones científicas puede ser suficiente usar un solo gen o proteína (por ejemplo, inventarios de biodiversidad, estudios de barcoding, etc.), para otras simplemente no es suficiente, cuando hay resultados incongruentes entre diferentes estudios y loci (Gee, 2003). Además, los análisis pueden tener una resolución limitada o estar mal respaldados debido a errores estocásticos cuando se basan en un pequeño número de secuencias y, en consecuencia, no en muchas características moleculares (Lozano-Fernandez, 2022). En cambio, los conjuntos de datos a una escala más grande son con frecuencia la fuente de árboles filogenéticos altamente respaldados y son menos propensos a errores de muestreo o estocásticos (Lozano-Fernandez, 2022).

La propuesta de Jonathan A. Eisen (1998) para usar la filogenómica en la predicción de la función génica se basa en la suposición de que la reconstrucción de la historia evolutiva de los genes debería ser capaz de predecir las funciones de los genes no caracterizados porque sus funciones se alteran como resultado de la evolución (Eisen, 1998). En síntesis, este es un proceso de varios pasos que puede involucrar la selección de homólogos, alineamiento múltiple de secuencias y construcción de un árbol filogenético; discriminando entre ortólogos y parálogos; y, finalmente, infiriendo la función de una proteína o gen en base a las secuencias identificadas por este proceso y las anotaciones generadas.

En el contexto del presente trabajo, se desea aplicar ciertos criterios metodológicos que pueden ser usados en filogenómica, para inferir la presencia de alcaloides anticancerígenos, de entre un grupo inicial de 72 plantas. Se tomó como referencia a *Catharanthus roseus*, una planta medicinal de la familia *Apocynaceae*, originaria de Madagascar y que ahora está presente en todas las regiones tropicales del mundo (Barrales-Cureño et al., 2019). De este modo, se intenta mostrar que estos análisis de las proteínas relacionadas con la biosíntesis de los alcaloides vincristina y vinblastina, en *C. roseus*, junto con sus secuencias homólogas, puede proporcionar información relevante sobre la evolución de estos compuestos y su posible distribución en otras plantas.

2. Determinación del problema

La producción de los alcaloides vincristina y vinblastina, fármacos anticancerígenos, es altamente compleja y costosa y, hasta la fecha, se conoce que existen solamente en la planta *Catharanthus roseus*. Estudios de relaciones evolutivas de estos alcaloides pueden sugerir su presencia en otras plantas medicinales como alternativa para su producción.

3. Objetivos

3.1. General

Analizar la existencia de la ruta de biosíntesis de los alcaloides vinblastina y vincristina en un grupo inicial de 72 plantas, a través de criterios filogenómicos, tomando como referencia las secuencias de cuatro proteínas importantes para la síntesis de estos alcaloides de la planta *Catharanthus roseus*.

3.2. Específicos

3.2.1. Aplicar un análisis de máxima verosimilitud para evaluar la calidad de las relaciones evolutivas.

3.2.2. Relacionar los resultados de este estudio con filogenias de rutas metabólicas de plantas.

3.2.3. Discutir las fortalezas y debilidades de la metodología planteada con relación a las herramientas bioinformáticas empleadas y el contexto biológico.

4. Revisión de Literatura

4.1. La filogenómica en el contexto de la genómica funcional

Aunque en la actualidad, la filogenómica tiene diversas aplicaciones que ayudan a entender de mejor manera las relaciones evolutivas, esta nació con la idea de predecir la función génica. En otras palabras, intenta abordar la cuestión: en un contexto evolutivo sobre qué función tiene determinada secuencia, siguiendo un flujo de trabajo que puede consistir en seleccionar secuencias homólogas, realizar alineamientos múltiples de secuencias, construir un árbol filogenético, sobreponer anotaciones en la topología del árbol y discriminar entre secuencias ortólogas y parálogas, para finalmente inferir la función de la proteína identificando los ortólogos y las anotaciones generadas (Eisen, 1998). Al menos hasta los años en los que Jonathan A. Eisen acuñara el término (1998) y en la década siguiente, la

inferencia filogenómica era aún muy poco utilizada. A la mayoría de las secuencias nuevas se les asignaba una posible función en base a similitudes de secuencia mediante búsquedas en una base de datos, lo cual conducía a numerosos errores de anotación. (Brown & Sjölander, 2006).

Anotar funciones génicas en base a similitudes de secuencias ignora que pudiera existir duplicaciones génicas, mezcla de dominios, errores en las anotaciones de bases de datos que no son detectados y eliminados, distancias evolutivas significativas en relación a especies divergentes, entre otros (Brown & Sjölander, 2006). Si el análisis filogenómico es aplicado correctamente, evita esos errores e, incluso, ayuda a detectarlos en las bases de datos (Sjölander, 2004). Una de las razones por las que la filogenómica no era ampliamente utilizada (al menos hasta ese entonces) es que su aplicación resulta mucho más complicada que una comparación por similitud de secuencia, a la vez que requiere de mayores conocimientos en evolución y bioinformática (Brown & Sjölander, 2006).

Se deben aplicar puntos de control de reproducibilidad para garantizar las mejores prácticas en filogenómica y bioinformática. Estos puntos de control son puntos en un flujo de trabajo donde se examina la integridad del flujo de trabajo, lo que permite validar los resultados en múltiples iteraciones para garantizar la coherencia de los resultados (Young & Gillung, 2020).

4.2. Filogenómica vs Filogenética

Jonathan Eisen dijo sobre la filogenómica que "*las predicciones funcionales pueden mejorar significativamente al enfocarse en cómo los genes llegaron a ser similares en su secuencia más que simplemente en la comparación de la similitud entre secuencias*" (Eisen, 1998). Ahora, en la actualidad, varias fuentes expanden este concepto al intento de inferir relaciones evolutivas a un nivel genómico (Lee, 2019; *Of Terms in Biology*, s. f.). No obstante, como señala Mike Lee (2019) esta es una visión muy simplificada del término porque, en la práctica, no se usan genomas enteros para realizar comparaciones evolutivas, ya que estos pueden ser imposibles de comparar debido a su enorme tamaño y hasta puede no tener sentido hacerlo dadas las grandes diferencias entre genomas. Es por esto que resultaría más apropiado decir que la filogenómica es el intento de inferir relaciones evolutivas en

algo más cercano al nivel del genoma de lo que nos lleva la filogenia de un gen individual (Lee, 2019).

La mayoría de las representaciones evolutivas que los biólogos están acostumbrados a ver, apunta Lee, son representaciones visuales de varias copias de un solo tipo de gen (como el gen 16S rRNA). En otras palabras, se intenta usar esas copias de genes como representantes del organismo, asumiendo relaciones evolutivas relevantes entre esos organismos que se están comparando. Con la filogenómica se realiza lo mismo, pero en lugar de comparar un solo gen, se comparan múltiples genes. Dicho esto, Lee acota que una definición más precisa podría ser: "*la filogenómica intenta inferir relaciones evolutivas entre secuencias compuestas de múltiples genes concatenados, mientras se asume que esas relaciones evolutivas inferidas nos dicen algo significativo con respecto a las relaciones evolutivas de sus genomas de origen*" (Lee, 2019).

4.3. Identificación de homólogos

Como punto de partida, de acuerdo a la propuesta metodológica de Eisen, debemos identificar los homólogos de los genes o proteínas de los cuales deseemos estudiar su evolución (Eisen, 1998). Esta es una tarea importante en investigaciones de genómica comparativa para entender la anotación del gen y la predicción de su función y una de las herramientas más conocidas para este fin es BLAST (Basic Local Alignment Search Tool) (Mesilaakso, 2019). La homología de secuencia es la homología biológica entre secuencias de ADN, ARN o proteínas, definida en términos de un mismo origen evolutivo. En otras palabras se refiere a la relación de descendencia común entre cualquier entidad biológica. Dichas entidades relacionadas por homología, de manera particular los genes, se denominan homólogos (Koonin, 2005).

4.4. Secuencias ortólogas y parálogas

Las secuencias ortólogas y parálogas son subcategorías de las secuencias homólogas (Koonin, 2005). Los ortólogos son genes que surgen a través de eventos de especiación (descendencia vertical), mientras que los parálogos son genes que se forman a través de duplicación (Fitch, 1970). La interrelación entre eventos de especiación y duplicación, con transferencia horizontal de genes, pérdida de genes y

rearreglos génicos, entrelazan ortólogos y parálogos en complicadas redes de interacción (Koonin, 2005).

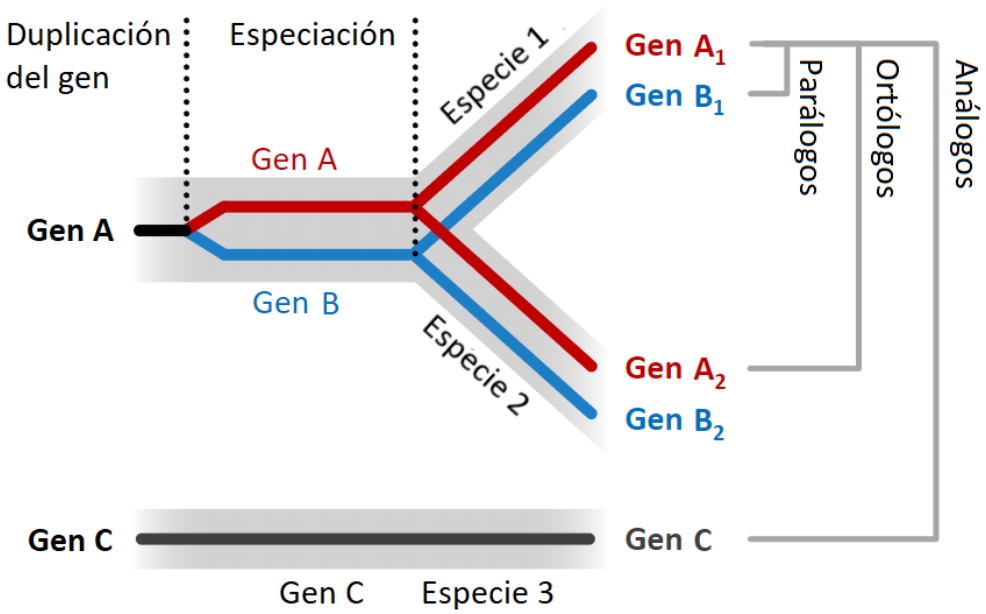


Figura 1. Ortólogos vs Parálogos

Un gen ancestral A se duplica para dar lugar a dos parálogos (genes A y B). Un episodio de especiación produce ortólogos en ambas especies hijas. Abajo: en una especie aparte, un gen que no está relacionado posee una función similar (gen C) pero tiene un linaje evolutivo separado y, por tanto, es análogo. Diagrama traducido del autor (Shafee, 2018), con licencia Creative Commons.

La estimación de las relaciones filogenéticas siempre deben realizarse con secuencias que están relacionadas por ortología, esto es, cuyo ancestro común divergió como resultado de la especiación (ortólogos), en lugar de un evento de duplicación (parálogos) (Fitch, 1970). Los genes que surgen de eventos de duplicación complican la inferencia de un árbol de especies basado en alineamientos de genes concatenados, porque el árbol de genes que describe las relaciones entre parálogos puede diferir del árbol de especies. La evaluación de la ortología se ha convertido en un problema central para los biólogos evolutivos y moleculares. En la inferencia filogenética, se asume con anterioridad que los loci

usados para inferir relaciones evolutivas son ortólogos, y la violación de esta suposición puede derivar en errores filogenéticos (Young & Gillung, 2020).

4.5. Las plantas medicinales

Desde los albores de la humanidad, la gente ha buscado remedios naturales para sus enfermedades y de esto existe amplia evidencia (Petrovska, 2012): documentos escritos, monumentos conservados, las plantas medicinales originales e, incluso, la sabiduría popular que se transmite a través de relatos orales. Los primeros usos de las plantas medicinales fueron puramente instintivos y en base a la experiencia porque no había suficiente información disponible en ese momento sobre las causas de las dolencias o las plantas específicas que podrían usarse como remedio (Dopico et al., 2008). A medida que la justificación del uso de plantas medicinales particulares para curar dolencias salió a la luz a lo largo del tiempo, el uso de plantas medicinales renunció gradualmente al marco empírico y, en cambio, se construyó sobre hechos explicativos y el método científico (Petrovska, 2012).

Las plantas producen muchos químicos que son biológicamente activos, no solamente para ellas mismas, sino también para otros organismos. Estos pueden actuar como herbicidas que inhiben el crecimiento de otras plantas competidoras, como el ácido salicílico producido por el sauce y otras producen sustancias que alejan a herbívoros o atraen a insectos. Los ingredientes activos de las plantas que pueden tener usos para los humanos son alcaloides, bitters, glucósidos cardiacos, glucósidos cianogénicos, flavonoides, minerales, fenoles, polisacáridos, proantocianinas, saponinas, taninas, vitaminas o aceites volátiles (*Medicinal Botany - Active Plant Ingredients*, s. f.).

4.6. Alcaloides

Existen numerosos compuestos naturales. Los alcaloides parecen ser particularmente únicos entre las muchas clases de compuestos orgánicos naturales, incluidos azúcares, lípidos, proteínas, aminoácidos, antocianinas, flavonoides y esteroides. Estos se originan a partir de aminoácidos y pueden ser producidos por plantas y algunos animales como metabolitos secundarios. Estas sustancias son esenciales para los seres vivos. En dosis muy pequeñas, los alcaloides exhiben potentes efectos biológicos tanto en organismos animales como humanos. Los

alcaloides se encuentran en alimentos y bebidas consumidos por humanos todos los días, así como en algunos medicamentos estimulantes (Kurek, 2019).

Alrededor del 20% de los metabolitos secundarios conocidos que se encuentran en las plantas son alcaloides (Kaur & Arora, 2015). Las funciones de los alcaloides en las plantas tienen que ver con protegerlas contra los depredadores y controlar su crecimiento (Chik et al., 2013). Los alcaloides son particularmente bien conocidos por sus usos terapéuticos como anestésicos, cardioprotectores y agentes antiinflamatorios. Entre los alcaloides bien conocidos que se utilizan en entornos clínicos se encuentran la nicotina, la efedrina, la estricnina y la quinina (Kurek, 2019).

4.7. Herramientas bioinformáticas utilizadas

4.7.1. *UniProt.* Es una base de datos de proteínas de acceso gratuito que también contiene información funcional. Mucha de su información se basa en proyectos de secuenciación genómica y de literatura de investigación. Esta iniciativa es mantenida por el consorcio UniProt que está conformado por varias organizaciones bioinformáticas Europeas y una fundación de Washington, DC, Estados Unidos («UniProt», 2015).

4.7.2. *Bash.* Es un shell de Unix y un lenguaje de comandos escrito por Brian Fox para el Proyecto GNU como un sustituto de software libre para el Bourne shell (*The A-Z of Programming Languages*, 2011). Lanzado por primera vez en 1989 (*Bash is in beta release! - gnu.announce | Google Groups*, 2013), se ha usado como shell de inicio de sesión predeterminado para la mayor parte de las distribuciones de Linux (Warren, 2019).

4.7.3. *MAFFT (versión 7)*. Es un software de alineación de secuencias múltiples para sistemas operativos similares a Unix. Proporciona una variedad de múltiples métodos de alineación, L-INS-i (precisa; para alineación de <~200 secuencias), FFT-NS-2 (rápida; para alineación de <~30 000 secuencias), etc (Katoh et al., 2019).

4.7.4. *Mesquite (versión 3.81)*. Es un programa modular y ampliable para biología evolutiva, proyectado para ayudar a los biólogos a organizar y analizar datos comparativos sobre organismos. Está enfocado

principalmente en el análisis filogenético, pero ciertos módulos se refieren a la genética de poblaciones, mientras que otros pueden ejecutar análisis multivariados no filogenéticos. Considerando que es modular, los análisis de los cuales dispone dependen de los módulos instalados (*Mesquite Project*, s. f.).

4.7.5. *CIPRES Science Gateway (versión 3.3)*. Es un recurso público para la inferencia de todo tipo de árboles filogenéticos. Está diseñado para facilitar a los investigadores acceso a recursos computacionales de NSF ACCESS por medio de una interfaz de navegador simple (Miller et al., 2010).

4.7.6. *FigTree (versión 1.4.4)*. Sirve como un visor gráfico de árboles filogenéticos y como un programa para generar figuras listas para una publicación. Particularmente está diseñado para mostrar árboles resumidos y anotados producidos por BEAST (*FigTree*, s. f.).

4.8. Modelos de sustitución evolutivos

Son modelos matemáticos y estadísticos que describen las tasas de cambio de mutaciones fijas entre secuencias y representan la base del análisis evolutivo de los datos genéticos a nivel molecular (Arenas, 2015). Los primeros modelos de sustitución fueron diseñados hace muchos años (Dayhoff, 1978), pero no fueron implementados sino hasta la creación de paquetes de software basados en métodos de máxima verosimilitud (Cummings, 2004). Diferentes investigadores han llegado a la conclusión de que el uso inadecuado de un modelo de sustitución puede conducir a sesgos en la construcción de inferencias filogenéticas (Lemmon & Moriarty, 2004). Por tanto, la elección del modelo de sustitución más apropiado es un paso esencial en el flujo de trabajo para la inferencia filogenética (Posada & Crandall, 1998).

Dentro del modelo de máxima verosimilitud utilizado en el presente proyecto, se empleó el programa RAxML que incluye, por defecto, el modelo de sustitución de aminoácidos PROTCAT. Este modelo trabaja con una matriz de aminoácidos con optimización de tasas de sustitución individuales por sitio y

clasificación de esas tasas individuales en el número de categorías de tasas especificadas (Stamatakis, 2008).

4.9. Principales formatos usados en el proyecto

FASTA. Es un formato de texto que se utiliza para representar secuencias de nucleótidos o aminoácidos, usando un código de letras. La línea antes de la secuencia, que se conoce como línea de definición FASTA, debe comenzar con el caracter ">", seguida de un SeqID (identificador de secuencia) único.

Figura 2. Formato FASTA

Ejemplo de un archivo en formato FASTA obtenido de una de las bases de datos de NCBI. Se puede apreciar, en el rectángulo en rojo, la línea de definición FASTA, que contiene el caracter ">" y el identificador de secuencia correspondiente, con el código de acceso, la proteína y especie. Bajo la línea de definición FASTA, está la secuencia de aminoácidos de ese archivo.

NEXUS. El formato de archivo NEXUS se usa cotidianamente en bioinformática. Guarda información sobre taxones, caracteres morfológicos y moleculares, distancias, códigos genéticos, suposiciones, conjuntos, árboles, etc. Muchos programas filogenéticos conocidos como PAUP, MrBayes, Mesquite, MacClade y SplitsTree usan este formato (Maddison et al., 1997).

```
#NEXUS

BEGIN DATA;
        dimensions ntax=546 nchar=1140;
        format missing=?
        symbols="ABCDEFGHIKLMNOPQRSTUVWXYZ"
        interleave datatype=DNA gap=- ;
matrix
Allegheny1                  ATGGCAAGCCTACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
Allegheny2                  ATGGCAAGCCTACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
Allegheny3                  ?????????????GAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
ArkansasA_1                 ???GCAAGCCTACGAAAAACCCACCCCCTACTAAAAA?TGCAAACAACGCACTGGTTGA
ArkansasA_2                 ????????????????????CCCACCCCCTACTAAAAATTGCAAACAACGCACTGGTTGA
ArkansasB_2                 ATGGCAAGCCTACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTGGTTGA
ArkansasB_3                 ????????????????????????????????????TGCAAACAACGCACTGGTTGA
ArkansasB_5                 ??GGCAAGCCTACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTGGTTGA
CumberlandA_1               ???????????????????AACCC?CCCCCTACTAAAAATTGC?AACAACGCACTTGT?GA
CumberlandA_2               ATGGCAAGCCTACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTT?TTGA
CumberlandA_3               ATGGCAAGCCTACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTG?
CumberlandA_4               ????????????????????????????CCCTACTAAAAATTGCAAACAACGCACTTGTTGA
CumberlandB_1               ????????????????????????????????CTACTAAAAATTGC?AAC?ACGCACTTGTTGA
CumberlandB_2               ????????????????????????????CCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
CumberlandB_3               ???GCAAGCCTACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
Elk_1                       ATGGCAAGCCTACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
Greenbriar_1                ?????????????GAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
Hudson_1                    ATGGCAAGC?TACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
Hudson_2                    ATGGCAAGC?TACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
Hudson_3                    ATGGCAAGC?TACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
Hudson_4                    ATGGCAAG?CTACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
Hudson_5                    ?????????????AAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
Genessee_1                  ATGGCAAGCCTACGAAAAACCCACCCCCTACTAAAAATTGCAAACAACGCACTTGTTGA
```

Figura 3. Formato NEXUS

Ejemplo de un archivo que emplea el formato NEXUS (*ALTER User's Guide*, s. f.).

PHYLIP. Este formato de archivo almacena una alineación de secuencia múltiple. Se definió y usó originalmente en el paquete PHYLIP de Joe Felsenstein, y desde entonces ha sido compatible con varias otras herramientas bioinformáticas (p. ej., RAxML) (*PHYLIP multiple sequence alignment format (skbio.io.phylip) — scikit-bio 0.2.3 documentation*, s. f.).

```
      5    42
Turkey    AAGCTNGGGC ATTTCAGGGT GAGCCCGGGC AATACAGGGT AT
Salmo gairAAGCCTTGGC AGTGCAGGGT GAGCCGTGGC CGGGCACGGT AT
H. SapiensACCGGTTGGC CGTTCAGGGT ACAGGTTGGC CGTTCAGGGT AA
Chimp     AAACCCTTGC CGTTACGCTT AAACCGAGGC CGGGACACTC AT
Gorilla   AAACCCTTGC CGGTACGCTT AAACCATTGC CGGTACGCTT AA
```

Figura 4. Formato PHYLIP

Ejemplo de un archivo que usa el formato PHYLIP (*sequence*, s. f.)

5. Metodología

Este estudio se basa en un *pipeline* de ocho etapas principales (ver figura 5). Las secuencias de referencia de *C. roseus* que se utilizaron, están relacionadas con la biosíntesis de los alcaloides vincristina y vinblastina siento estos: *tryptophan decarboxylase* (TDC), *strictosidine synthase* (STR1), *desacetoxyvindoline 4-hydroxylase* (D4H) y *deacetylvindoline 4-O-acetyltransferase* (DAT) (Barrales-Cureño et al., 2019).

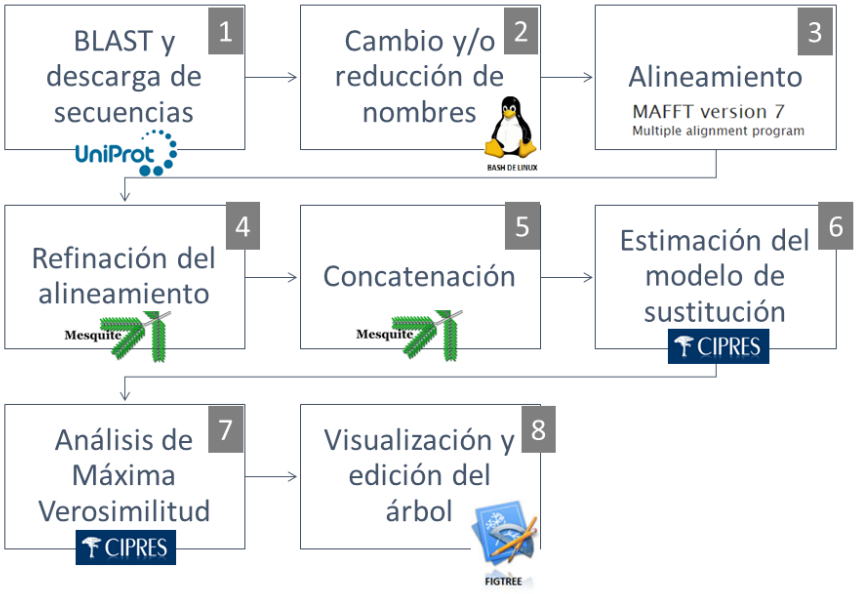


Figura 5. Pipeline del estudio

Pipeline usado en el proyecto para la inferencia filogenética de secuencias asociadas a la ruta de biosíntesis de los alcaloides vinblastina y vincristina en *C. roseus*.

La búsqueda de secuencias homólogas correspondientes a las cuatro proteínas de mi interés, se hizo con *Blast-p* en *UniProtKB* (una de las bases de datos de *UniProt*) y, posteriormente, se descargaron cuatro archivos de secuencias homólogas en formato FASTA. A continuación, mediante una línea de comandos en el *Bash* de *Linux*, procedí a reducir los nombres de las secuencias en los archivos FASTA, dejando los nombres de las especies y el código de su identificación taxonómica. En algunas secuencias fue necesario realizar un cambio manual puesto que aparecían nombres taxonómicos de familias en lugar de las especies o géneros correspondientes; esta información la cotejaba con los datos relacionados a la secuencia de la base de datos de *UniProt*.

A continuación, fueron cargados los archivos FASTA en el programa en línea de alineamiento múltiple MAFFTA (versión 7). Sin alterar ningún parámetro por defecto del alineamiento, procedí a correr el programa para después exportar cada archivo en formato NEXUS. En el programa *Mesquite* se cargaron los archivos NEXUS para refinar los alineamientos, nivelando los extremos 5' y 3', rellenando los datos faltantes con signos de interrogación y suprimiendo datos anómalos que se generaron en el alineamiento de MAFFTA. Posteriormente realicé la concatenación en el mismo programa *Mesquite* para generar una sola matriz con una lista inicial de 158 taxones que, tras varias pruebas en los análisis de máxima verosimilitud, fue reducida a una lista final de 30 taxones.

La matriz del concatenado de secuencias se exportó a un archivo PHYLIP, que luego fue subido a la plataforma CIPRES para realizar el análisis de máxima verosimilitud (ML). En dicho análisis se corrió, por defecto, el modelo de sustitución PROTCAT el cual trabaja con matrices de aminoácidos. Para correr el análisis de ML, en CIPRES seleccioné el archivo de entrada correspondiente, usando la herramienta *RAxML-HPC2 on XSEDE* (v. 8.2.12). Se obtuvieron varios archivos de salida, pero me enfoqué en revisar STDOUT, archivo que muestra los resultados generales del análisis de ML y, tras llegar a la prueba con mejores resultados, descargué el archivo de salida *RAxML_bipartitions.result* el cual contenía el árbol filogenético.

En el programa *FigTree* se cargó el archivo de salida *RAxML_bipartitions.result* para la visualización y edición del árbol filogenético. En esta última etapa del proceso destaqué los valores de *bootstrap* (apoyo de ramas) y realicé diferentes ediciones para mejorar la lectura e interpretación del árbol.

5.1. BLAST y descarga de secuencias

En la base de datos UniProtKB de *UniProt* (www.uniprot.org), se realizó la búsqueda de secuencias homólogas por cada una de las cuatro proteínas asociadas a la ruta biosintética de los alcaloides vinblastina y vincristina en *C. roseus* (TDC, STR1, D4H y DAT). Se siguió el siguiente proceso:

1. Introducimos el nombre de la secuencia de referencia y la especie.

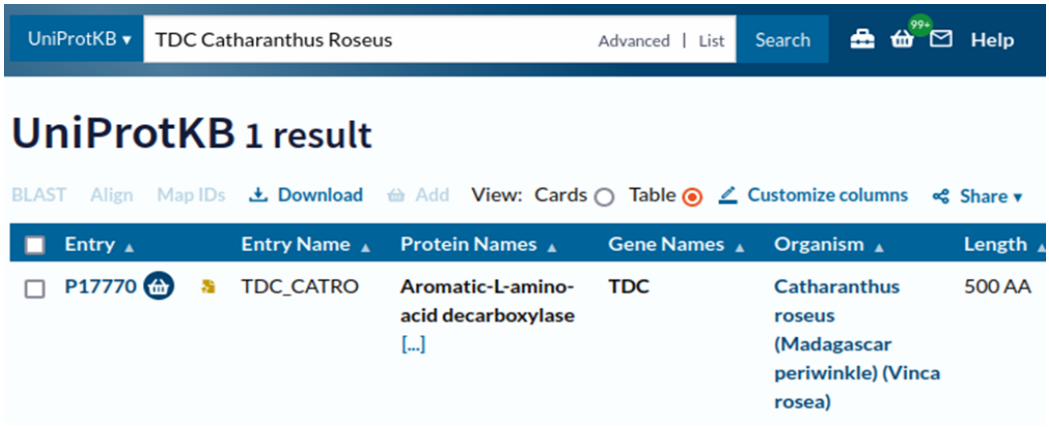


Figura 6. Base de datos UniProtKB

Vista principal de los resultados de búsqueda en UniProtKB.

2. Accedemos a la entrada (en la columna ENTRY) y luego ubicamos la herramienta BLAST en la sección *Sequence* opciones "Tools">"BLAST".

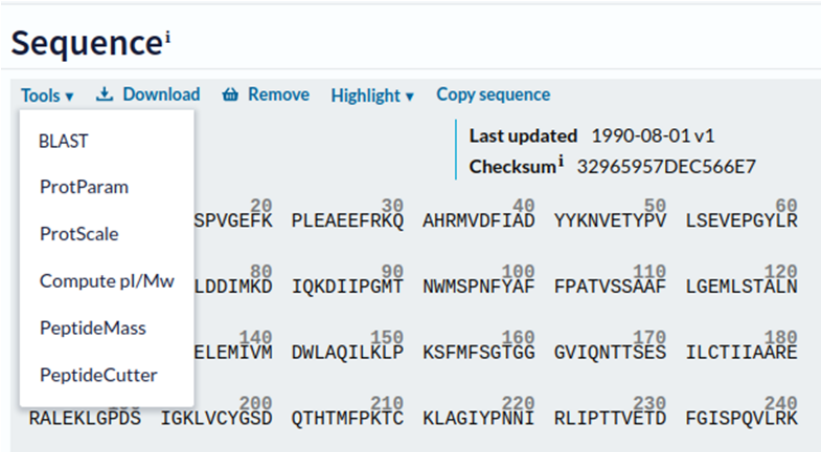


Figura 7. Secuencia de referencia en la base de datos

Información sobre la secuencia de referencia para realizar una búsqueda de BLAST en UniProtKB.

3. Refinamos los parámetros de la búsqueda de blast de la siguiente manera y, a continuación, damos clic en "Run BLAST".

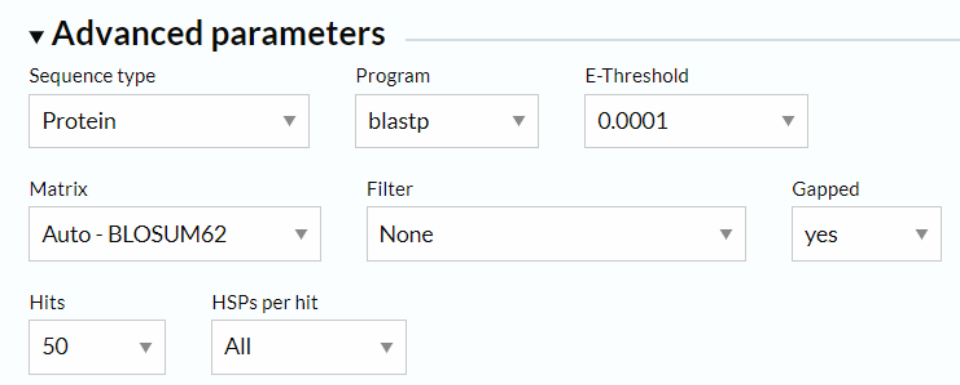


Figura 8. Parámetros avanzados del BLAST

Parámetros de la búsqueda de BLAST en la base de datos de UniProtKB.

4. Después de unos segundos, se habrá desplegado la lista de resultados como se muestra a continuación.

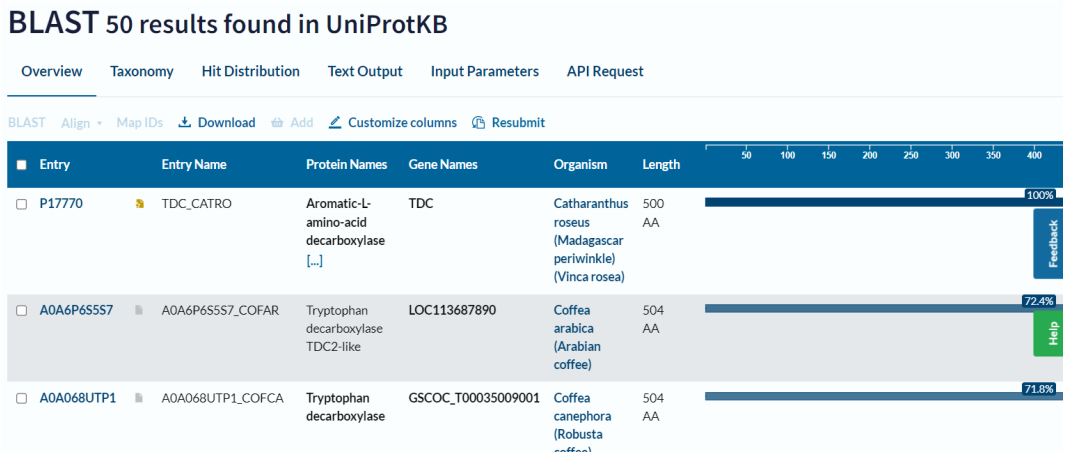


Figura 9. Resultados del BLAST

Lista de resultados de la búsqueda de BLAST.

5. Descarga de resultados en formato *FASTA*

Para descargar los resultados de la búsqueda de BLAST en *UniProt*, se deben seleccionar los resultados para mapear los IDs de las secuencias, haciendo luego clic en la opción “Map IDs”. Se desplegará una ventana con la lista de todas las secuencias dentro del trabajo “Retrieve/ID mapping”. Asignamos un nombre al trabajo como, por ejemplo, “TDC Blast” y damos clic en la opción “Map 50 IDs”.

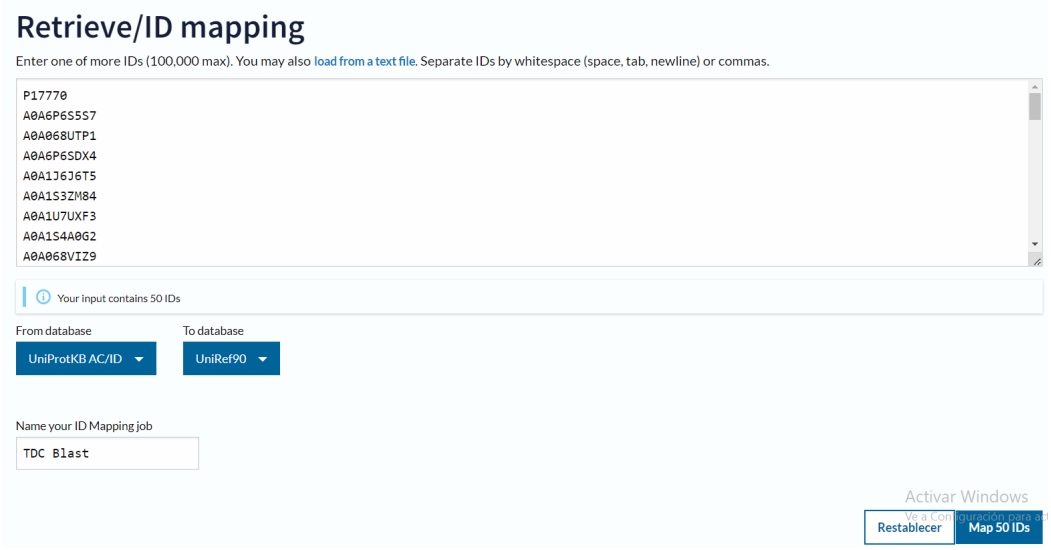


Figura 10. ID Mapping

Mapeos de ID para la obtención de secuencias FASTA a partir de los resultados de la búsqueda de BLAST.

6. Aparece la ventana *Tool results* donde se observan los trabajos realizados, como el ID mapping del paso anterior. Damos clic en el enlace del trabajo para observar los resultados del mapeo de los IDs de las secuencias "blasteadas".

Figura 11. Resultados ID Mapping

ID ya mapeados de los resultados de búsqueda de BLAST para la generación de secuencias FASTA.

7. Seleccioné todas las secuencias y di clic en “Download”. A continuación, apareció la siguiente ventana donde escogí la opción de formato “FASTA”. Se descargó el archivo de las secuencias homólogas encontradas. Repetí el proceso para el resto de secuencias.

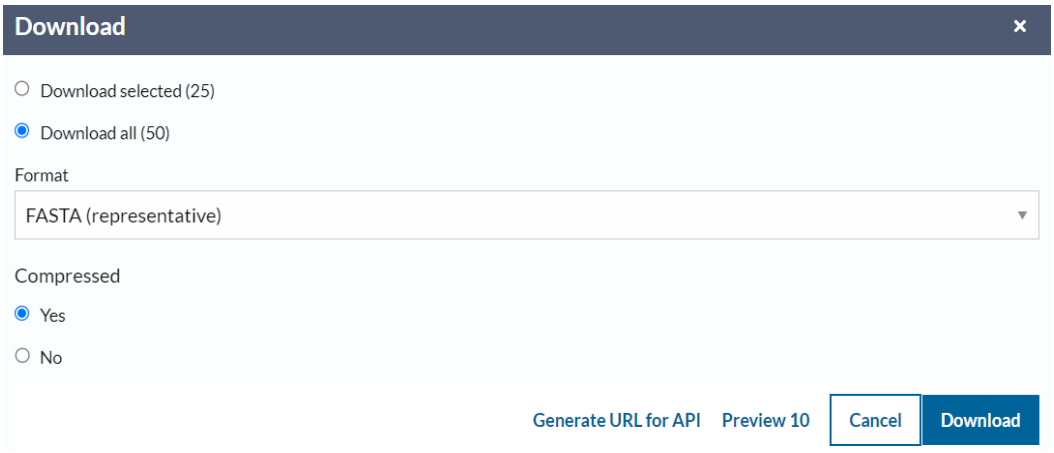


Figura 12. Descargando las secuencias en FASTA

Ventana de descarga de las secuencias en formato FASTA a partir de las búsquedas de BLAST.

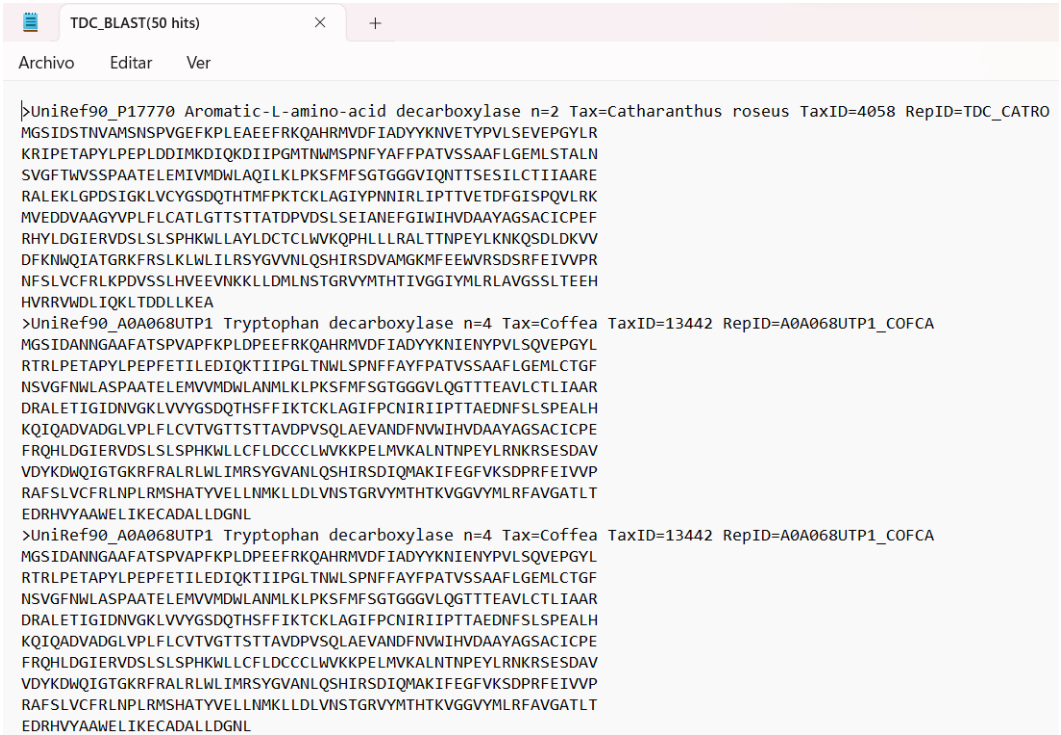


Figura 13. Uno de los archivos FASTA del proyecto

Archivo FASTA descargado de una de los resultados de búsqueda de BLAST para la proteína TDC.

5.2. Cambio y/o reducción de nombres

Se empleó una línea de comandos generada en el *Bash* de *Linux*, para reducir el nombre de las secuencias de los cuatro archivos FASTA, obteniendo solamente el nombre de la especie y su TaxID. Aquí se utilizó 'awk' para leer cada línea que empieza con ">" en el archivo FASTA. Después, se empleó el comando 'gsub()' para eliminar todo lo anterior al último carácter "|" en la línea (que representa el comienzo del nombre de la especie y el TaxID). Finalmente, se imprimió la línea resultante que contiene solo el nombre de la especie y el TaxID. El resultado se destina al archivo nombre_reducido.fasta. Cada línea en nombre_reducido.fasta contuvo el nombre de la especie y el TaxID correspondiente a cada secuencia en el mismo orden en que aparecen en el archivo FASTA original. La línea de comandos fue la siguiente:

```
~$ awk '/^>/{gsub(/.*\|/,"",$1); print $1}' archivo.fasta > nombre_reducido.fasta
```

Figura 14. Línea de comandos de Linux

Línea de comandos de Linux para la reducción de la línea de identificación FASTA de las secuencias.

Se realizó una revisión manual para determinar si los nombres de la especie coincidían con el mismo TaxID entre los cuatro archivos FASTA.

5.3. Alineamiento

Se cargaron los archivos *FASTA* en el programa en línea de alineamiento múltiple MAFFT (versión 7) (https://mafft.cbrc.jp/alignment/server/).

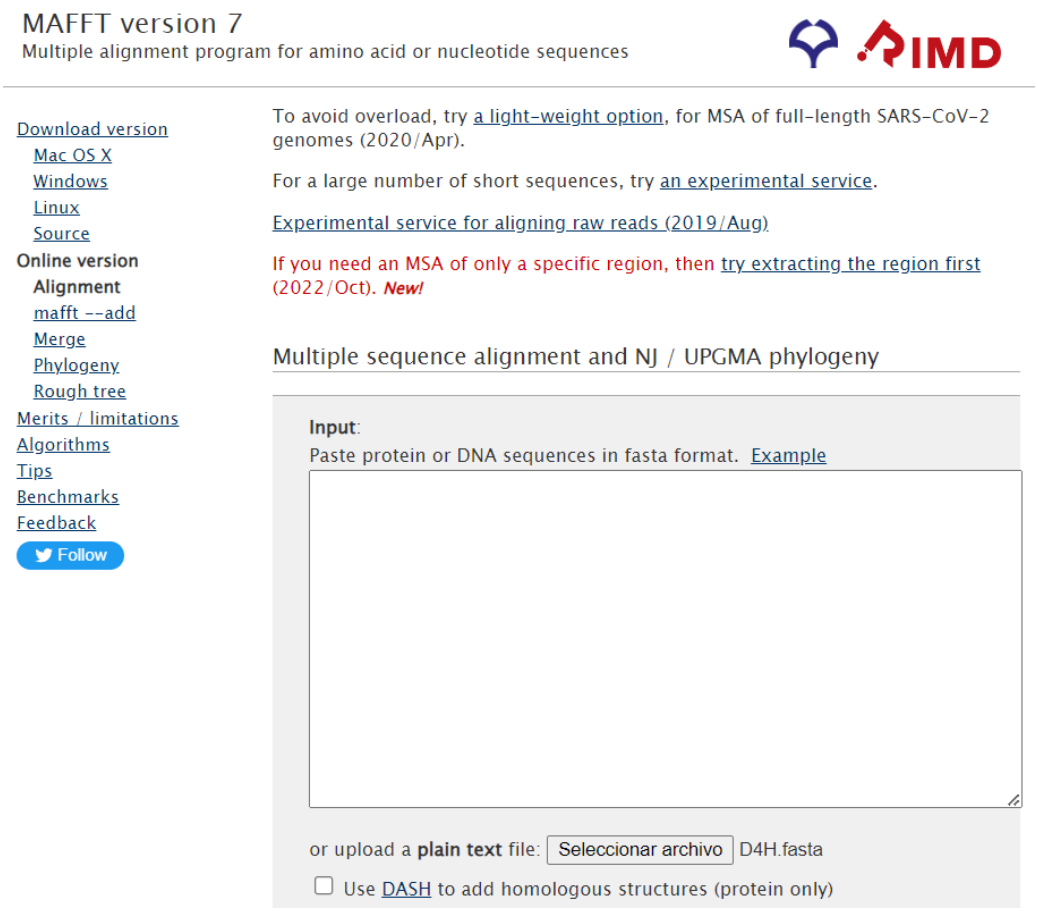


Figura 15. Programa MAFFT

Pantalla principal del programa MAFFT para la realización de alineamientos múltiples.

Sin alterar ningún parámetro por defecto del alineamiento, procedí a correr el programa, generando el siguiente resultado:

MAFFT–L–INS–i Result

```
CLUSTAL format alignment by MAFFT (v7.511)


Catharanth MPKSW----PIVISSHSFCFLPNSEQERKMKDLNFHAATLSEEESLRELKAFDETKAGVK
Nyssa_sine ML----------SPQEEKNMVITKTGE-------VQAITMPDYDRKSELKAFDESKAGVK
Coffea_Tax ---------------------------------------MTEYDRRSELEEFDNTKAGVK
Fagaceae3_ -------------------MIVTGPGE-------NQAELESKYDRNTELKAFDDSKAGVK
Camellia_s M-----------------VVTSNRGA-------IQTQIEPEYDWRSELKAFDDSKAGVK
Vitis_TaxI -------------------MVSSSSDE-------IQAGKASDYDRKSELKSFDDSKLGVK
Vitis2_Tax -------------------MVSSSSDE-------IQAGKASDYDRKSELKSFDDSKLGVK
Vitis3_Tax -------------------MVSSSSDE-------IQAGKASDYDRKSELKSFDDSKLGVK
Olea_europ ------------------------ME-------VKAKGISVYDRNSELRLFDDSKVGVK
Nyssa_sine -------------------MVATSVAE-------NTAVTLPDCDRKSELKAFDDSKAGVK
Quercus_Ta -------------------MEANYK---------NEAEVESVYDRQSEVKAFDDSRAGVK
Camellia_s MM----------------ANSTSTGK-------THAGSDLNYDRKTELKAFDDTKAGVK
Byttnerioi -------------------MVETKTGE-------VHADTNPDYDRQSEVKAFDDTKAGVK
Byttnerioi -------------------MVETKTGE-------VHADTNPDYDRQSEVKAFDDTKAGVK
Durio_zibe --------------------MGTKTGE-------VQGDSNPDYDSQSEVKAFDDTKAGVK
Jatropha_c -------------------MLATDSVT-------VEADSGSTNDRKSQLKAFDDTKSGVK
Olea_europ -------------------MFTTS----------RKFVTLEDDRIDELKAFDDTKAGVK
Manihot_es -------------------MDTPSG---------NFQKLECDYDRQAQLKAFDETRAGVK
Fagaceae_T -------------------MVVTSRDE-------V-----PDYDRASELKAFDDTKAGVK
Fagaceae2_ -------------------MVVTSRDE-------V-----PDYDRASELKAFDDTKAGVK
Castanea_m -------------------MVVTSRDE-------V-----PDYDRASELRAFDDTKAGVK
Juglandace M-----VRSPYEIESSRFQMVITSRDE-------VAATQKPGYDRASELKAFDSTKAGVK
Durio_zibe MG-TWSIRVELEVETGGKRQVRKEDRS-------PSNTSNPDYDRTSELKAFDDTKAGVK
```

Figura 16. Alineamiento MAFFT

Resultados del alineamiento en MAFFT.

Luego, se reformatearon los resultados a NEXUS.

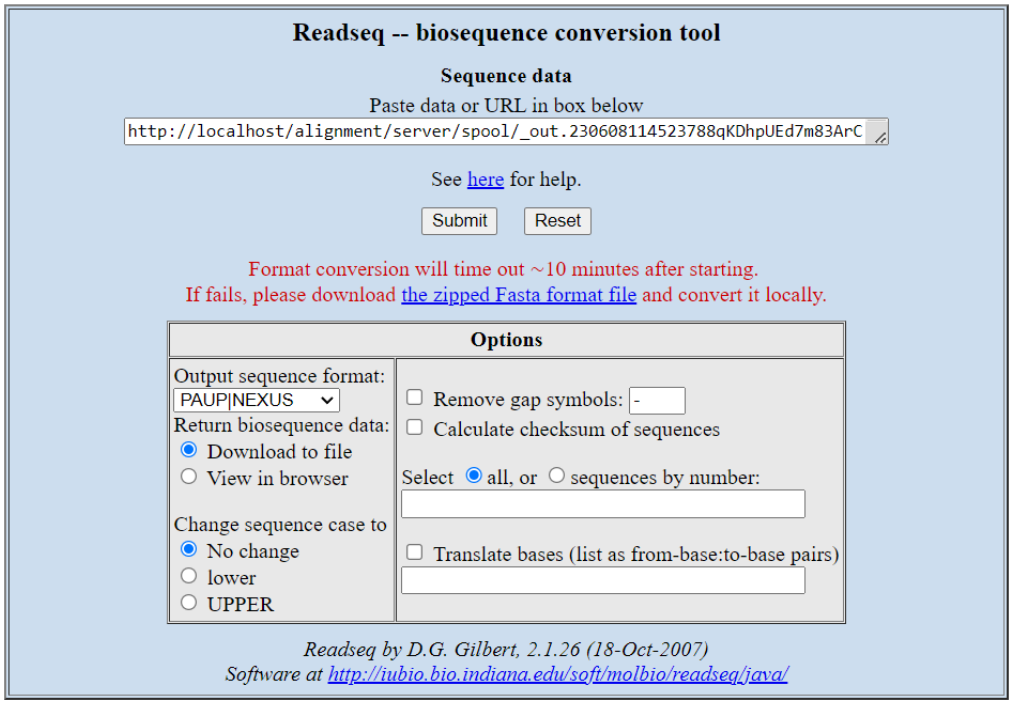


Figura 17. Descarga del alineamiento en NEXUS

Descarga de los resultados del alineamiento en formato NEXUS.

5.4. Refinación del alineamiento

Los archivos NEXUS se cargan en el programa MESQUITE.

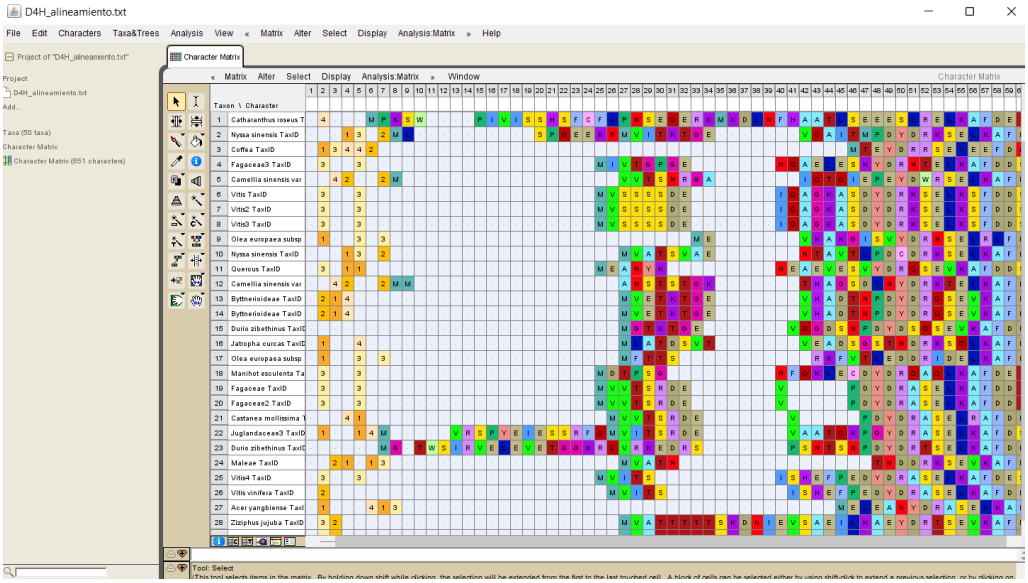


Figura 18. Visualización del archivo NEXUS

Archivo NEXUS de un alineamiento sin refinar, visualizado en el programa MESQUITE.

Aquí procedí a igualar los extremos de las secuencias en 5’ y 3’ cuando los archivos del alineado tenían muchos espacios vacíos (al menos en un 50% de las secuencias). Se eliminaron los residuos que presentaron números, reemplazando estos valores por gaps (“-”) y en los espacios vacíos que quedaron en los extremos, se colocó el símbolo de datos faltantes (“?”). Finalmente, también se eliminaron algunas columnas que no presentaban ningún residuo a lo largo del alineamiento. Esta etapa del refinamiento se repitió varias veces, luego de pasar por el análisis de máxima verosimilitud que, entre otras cosas, me ayudaba a identificar estos errores. Dejaron de aparecer errores en el análisis de máxima verosimilitud tras realizar siete refinamientos.

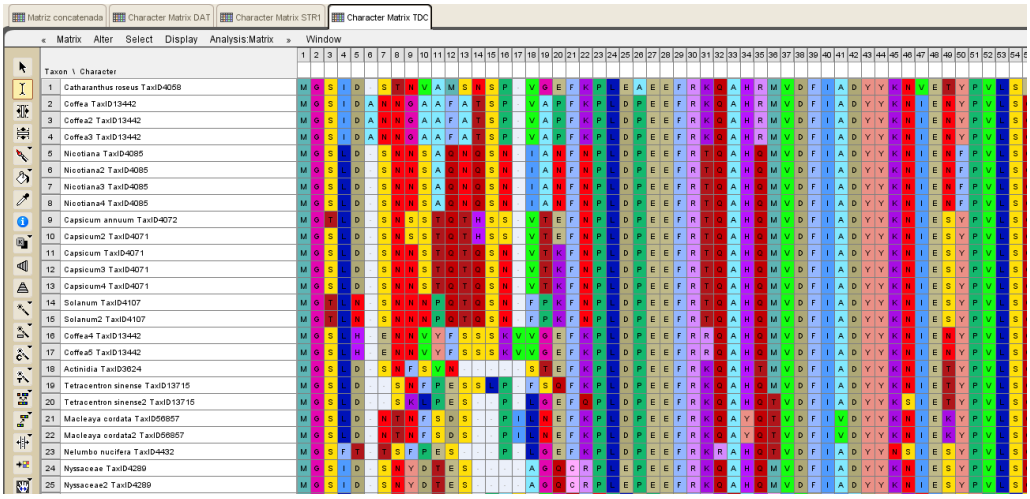


Figura 19. Archivo NEXUS ya editado

Archivo NEXUS de un alineamiento refinado, visualizado en el programa MESQUITE.

5.5. Concatenación

En el mismo programa que realizamos la refinación, procedemos a concatenar los cuatro archivos de alineaciones que ya refinamos. Para esto, hay que cargar todas las matrices de alineamientos en el área de trabajo y luego ingresar a *Matrix>Utilities>Concatenate Other Matrix*.

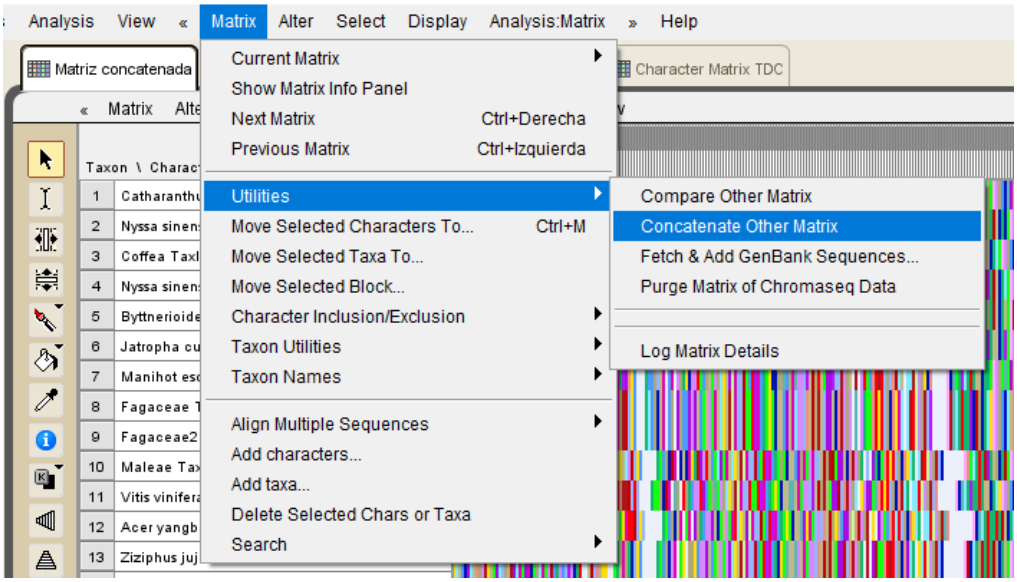


Figura 20. Concatenación

Archivo NEXUS de un alineamiento refinado, visualizado en el programa MESQUITE.

A continuación, se abrió una ventana donde seleccionamos las matrices a concatenar. Seleccioné a PHYLIP como el formato del archivo de salida de la concatenación.

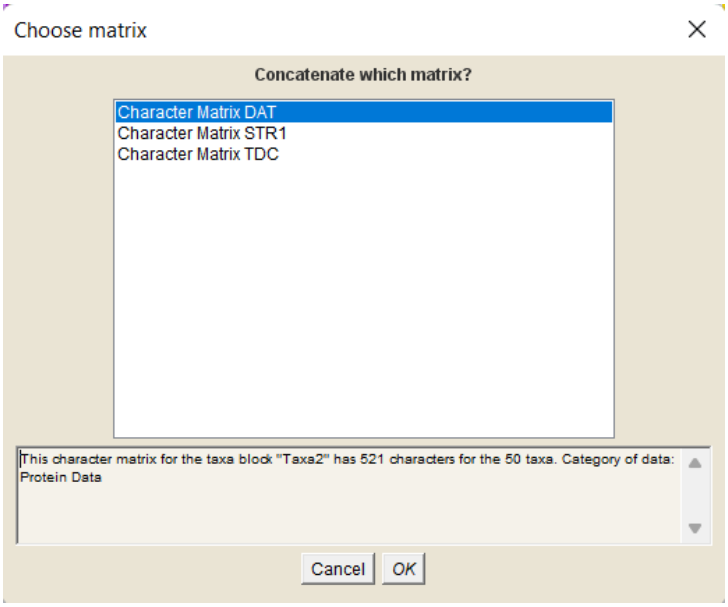


Figura 21. Ventana de concatenación

Ventana de concatenación en el programa MESQUITE.

5.6. Estimación del Modelo de Sustitución y Análisis de Máxima Verosimilitud en *CIPRES*

Tras haber creado una cuenta en CIPRES, un programa de análisis filogenéticos (www.phylo.org), creé una carpeta de trabajo llamada “Análisis filogenéticos”. De manera automática, se crearon las subcarpetas “Data” y “Tasks”. En “Data” cargué el archivo PHYLIP de la matriz concatenada.

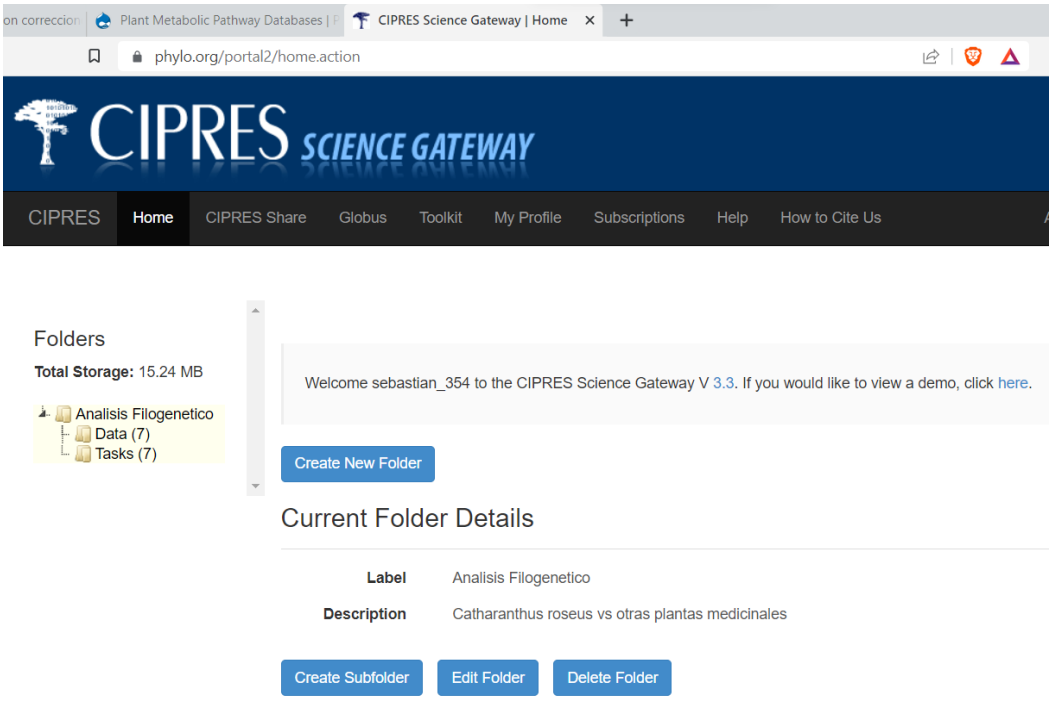


Figura 22. CIPRES

Vista principal del programa en línea CIPRES para realizar análisis filogenéticos.

Tras haber cargado el archivo PHYLIP, ingresé a "Tasks" donde creé un nuevo trabajo. Aquí debí establecer una descripción de la tarea, seleccionar el archivo de entrada para el análisis, la herramienta y la definición de los parámetros. La herramienta seleccionada para el análisis de Máxima Verosimilitud fue *RAxML-HPC2 on XSEDE* (v. 8.2.12), que está provista de un modelo de sustitución para matrices de aminoácidos llamado PROTCAT. Los parámetros que se modificaron fueron el tiempo máximo de análisis (30 min), grupo externo (*Catharanthus roseus*), el número de columnas de la matriz (1832) y las iteraciones de bootstrap (1000). El resto de parámetros de la herramienta se dejaron sin modificación.

RAxML-HPC2 on XSEDE - Parameters

ascertainment_corr_	lewis
bootstop_	1000
bootstrap_seed_	true
bootstrap_seed_val_	12345
convergence_criterion_	false
datatype_	protein
disable_ratehet_	false
disable_seqcheck_	false
intermediate_treefiles_	false
mesquite_output_	false
mlsearch_	true
mulcustom_aa_matrices_	false
nchar_	1832
no_bfgs_	false
number_cats_	25
outgroup_	Catharanthus_roseus_TaxID4058
outsuffix_	result
parsimony_seed_val_	12345
printbrlength_	false
prot_matrix_spec_	DAYHOFF
prot_sub_model_	PROTCAT
provide_parsimony_seed_	true
rearrangement_yes_	false
runtime_	0.25
specify_bootstraps_	true
use_apobootstopping_	false

Figura 23. Parámetros del análisis de máxima verosimilitud

Lista de parámetros de la herramienta *RAxML-HPC2 on XSEDE* (v. 8.2.12), que fue usada para el análisis de Máxima Verosimilitud, en la plataforma CIPRES.

Después de varios minutos de espera, se generaron, como resultado del análisis, varios archivos de salida. Me concentré en STDOUT y en el archivo del árbol RAxML_bipartitions.result. En STDOUT se presentaba un reporte de errores en el análisis y valores de calidad; sobre esto último, me fijé principalmente que el porcentaje de gaps y de caracteres indeterminados sea inferior a 50% y que el valor final de Máxima Verosimilitud vaya mejorando tras cada análisis (considerando que entre más cercano a 0 era mejor). En las primeras pruebas, el porcentaje siempre era superior a 50% por lo que tenía que ir modificando la matriz original eliminando los taxones que tenían solamente una secuencia homóloga, pasando, de este modo, de 158 a tan solo 30 taxones. Este último grupo tenía 2, 3 o 4 secuencias homólogas. Luego de haber tenido los valores de calidad esperados, descargué el archivo del árbol RAxML_bipartitions.result en formato TREE.

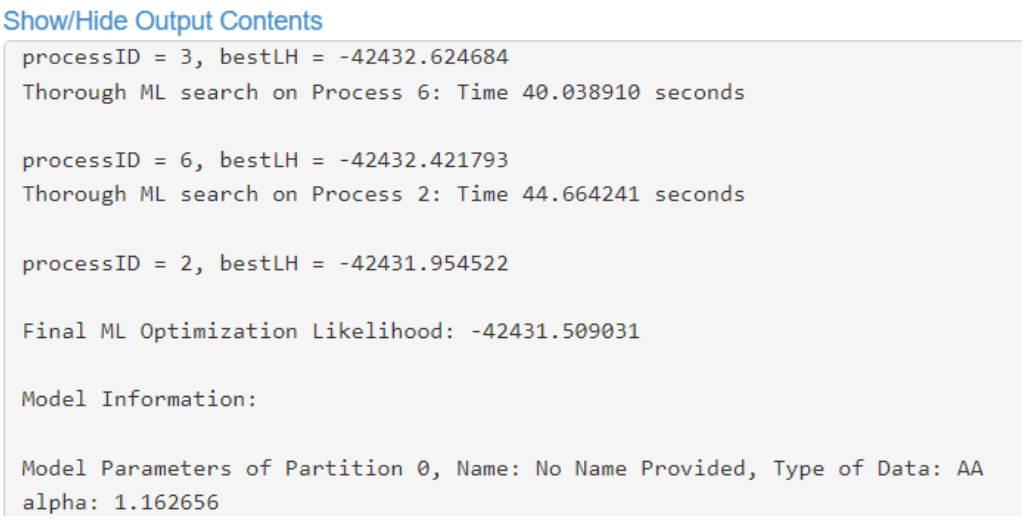


Figura 24. Reporte del análisis de máxima verosimilitud

Visualización del archivo STDOUT donde se aprecia una parte del reporte.

5.7. Visualización y edición del árbol

En el programa *FigTree* (v1.4.4) cargué el archivo del árbol RAxML_bipartitions.result y, con las diferentes herramientas que ofrece este programa, puse los valores de *bootstrap* (apoyo de ramas), aumenté el tamaño de la fuente, coloreé los clados, puse el árbol en versión circular, entre otras modificaciones de forma para mejorar sus visualización e interpretación.

6. Discusión y Análisis de resultados

6.1. Blast y descarga de secuencias homólogas

Tras haber realizado la búsqueda de secuencias homólogas en la base de datos de *UniProt*, se hallaron 158 secuencias entre las cuatro proteínas de referencia: *tryptophan decarboxylase* (TDC), *strictosidine synthase* (STR1), *desacetoxyvindoline 4-hydroxylase* (D4H) y *deacetylvindoline 4-O-acetyltransferase* (DAT). Sin embargo, muchas de estas secuencias corresponden a genes duplicados dentro de una misma especie de planta, totalizando así 72 especies (ver tabla 1 en la sección de anexos). De estas 72, sólo tres plantas tenían secuencias para las cuatro proteínas de búsqueda, cuatro plantas para tres proteínas, diez plantas para dos, y 55 plantas (el mayor número) tenían al menos una secuencia homóloga para una proteína.

Esta lista de plantas podía haber sido más extensa si se aumentaba el número de "hits" en la búsqueda de *blast-p*, pero esto implicaba obtener secuencias con porcentajes de identidad bajos. Si más secuencias hubieran sido incluidas en la filogenia, la calidad del árbol se habría reducido. Cada búsqueda de *blast-p* fue condicionada a 50 "hits" que arrojaron resultados con porcentajes de identidad superiores a 50%. Aún así, la lista de 72 plantas con sus respectivas secuencias homólogas (que sumaban 158 secuencias en total en la primera matriz de concatenados), fue reducida a 30 taxas en la matriz final (prueba #7), debido a ciertos criterios de calidad que serán discutidos más adelante.

Es importante mencionar que la presencia de más de una secuencia homóloga en varias plantas, se puede deber a la poliploidia del genoma. Es decir, existen múltiples copias de genes que podrían traducirse en varias proteínas homólogas, volviendo más complejos estos estudios.

6.2. Cambios y/o reducción de nombres en los archivos FASTA

Se descargaron cuatro archivos FASTA de las secuencias homólogas de cada una de las proteínas de referencia. Fue necesario reducir los nombres de la *línea de definición* FASTA, para mejorar la legibilidad del árbol filogenético que se realizaría posteriormente. Esto fue realizado con una línea de comandos de bash de Linux y con una revisión manual de ciertas secuencias que tenían el mismo TaxID pero diferente nombre. La matriz de secuencias concatenadas requiere que las secuencias homólogas tengan el mismo nombre para las cuatro proteínas, por lo que este paso es fundamental.

6.3. Refinación del alineamiento

En el programa Mesquite se pudieron observar ciertas imperfecciones en los alineamientos que necesitaban ser refinados para obtener una mejor calidad del árbol. Aparecieron números del 1 al 4 en regiones internas de los alineamientos que fueron eliminados. Hubo seis espacios en blanco en ciertas columnas de la matriz que también fueron suprimidos. Y se cortaron los extremos en 5' y 3' si el alineamiento presentaba regiones irregulares en más del 50% de secuencias. Además, se rellenaron los espacios que quedaron en los extremos con un signo de interrogación. La importancia de este relleno fue para que, en la construcción del

árbol, no se asocien estos espacios como un estado de carácter. También aparecían secuencias repetidas que eran identificadas en los análisis de máxima verosimilitud (figura 24) y que luego fueron eliminados.

```
IMPORTANT WARNING: Alignment column 1382 contains only undetermined values wh
ich will be treated as missing data

IMPORTANT WARNING: Alignment column 1383 contains only undetermined values wh
ich will be treated as missing data

IMPORTANT WARNING: Sequences Vitis_TaxID3603 and Vitis2_TaxID3603 are exactly
identical

IMPORTANT WARNING: Sequences Vitis_TaxID3603 and Vitis3_TaxID3603 are exactly
identical

IMPORTANT WARNING: Sequences Byttnerioideae_TaxID214909 and Byttnerioideae2_T
axID214909 are exactly identical

IMPORTANT WARNING: Sequences Juglandaceae_TaxID16714 and Juglandaceae2_TaxID1
6714 are exactly identical
```

Figura 25. Advertencias del análisis de máxima verosimilitud

Mensajes de advertencia del análisis de máxima verosimilitud sobre valores indeterminados (en relación a los números que aparecían en el alineamiento) y de secuencias con nombres repetidos.

Estas modificaciones se realizaron en seis rondas, cuando los resultados principales del análisis de máxima verosimilitud indicaban errores o valores de baja calidad relativa. Por ejemplo, en la prueba #6 de refinamiento, se obtuvieron los siguientes resultados de proporción de gaps y caracteres indeterminados y valor final de máxima verosimilitud:

```
Proportion of gaps and completely undetermined characters in this alignment:
64.09%

Final ML Optimization Likelihood: -54369.871930
```

Figura 26. Resultados del análisis de máxima verosimilitud en una de las pruebas

Resultados de la proporción de gaps y caracteres indeterminados y el valor final de máxima verosimilitud en la prueba #6.

El objetivo para alcanzar valores de calidad era que, en cada prueba, el valor final de máxima verosimilitud vaya acercándose cada vez más al cero y que sea

mejor que las pruebas anteriores. Respecto al porcentaje de proporción de gaps y caracteres indeterminados, este tenía que ser inferior al 50% para generar un árbol más confiable. Entre cada una de estas pruebas, para mejorar estos valores, se iban suprimiendo las taxas que tenían solamente una secuencia homóloga de las cuatro proteínas de referencia porque estas aportaban una alta proporción de gaps y caracteres indeterminados. Finalmente, para la prueba #7, se eliminaron las taxas que tenían solo una secuencia, quedándome únicamente con aquellas que tenían más de dos secuencias homólogas. Esto permitió que se obtengan los siguientes resultados:

```
Proportion of gaps and completely undetermined characters in this alignment:
48.54%

Final ML Optimization Likelihood: -42431.509031
```

Figura 27. Resultados finales del análisis de máxima verosimilitud

Resultados de la proporción de gaps y caracteres indeterminados y el valor final de máxima verosimilitud en la prueba #7.

Para el valor final de máxima verosimilitud no hay una cifra estándar de calidad, sino que se sugiere alcanzar el mejor valor de entre todas las pruebas realizadas para un estudio particular.

6.4. Concatenación

Luego de haber refinado los cuatro alineamientos, estos se concatenaron en el mismo programa Mesquite, generando una matriz final (en la prueba #7) de 30 taxas y 1832 columnas. No es relevante el orden de la concatenación para la construcción del árbol filogenético porque independientemente del orden de las secuencias, se genera el mismo árbol.

6.5. Interpretación del árbol filogenético obtenido

Usando el programa *FigTree* se generó un árbol circular (figura 26) para visualizar mejor la cercanía de *C. roseus* con respecto a los clados y un árbol rectangular (figura 27) para observar más claramente los valores de bootstrap (apoyo de ramas). Lo primero que me llamó la atención de los árboles generados, fueron la presencia de tres clados bien diferenciados. Estos se colorearon de rojo, verde y azul.

Ninguno de estos tuvo una cercanía filogenética importante a *C. roseus*, ya que este se encuentra fuera de los tres clados. En el clado rojo aparecen las especies *Vitis vinifera* (uva), cinco taxones de secuencias homólogas de *Coffea*, *Acer yangbiense* (arce) y *Rhamnella rubrinervis*. Todas pertenecen a una familia distinta. En el clado verde están dos taxones homólogos de *Nyssa sinensis* (tupelo chino) y también dos de *Ziziphus jujuba* (jinjolero), *Maleae*, *Morella rubra* (arrayán chino), *Carpinus fangiana* (carpe cola de mono), *Jatropha curcas* (piñón de template) y *Byttnerioideae*. Mientras que en el clado azul, que es el más cercano a *C. roseus*, están dos taxones homólogos de *Capsicum*, cuatro de *Nicotiana*, dos de *Fagaceae*, *Solanum*, *Manihot esculenta*, *Cannabaceae* y *Gossypium* (figuras 27 y 28).

En referencia al bootstrap (o apoyo de ramas), es importante mencionar que este es un parámetro estadístico que determina la confiabilidad entre las ramas del árbol filogenético, a nivel de taxas o clados (*Probability Reduced Evolution of Spatially-discrete Species*, s. f.). Lleva una escala de 0 a 100, donde un valor superior a 70 podría indicar una relación estadísticamente fiable. En el árbol rectangular (figura 27) se pueden observar valores superiores a 70 en la mayoría de ramas dentro de los clados.

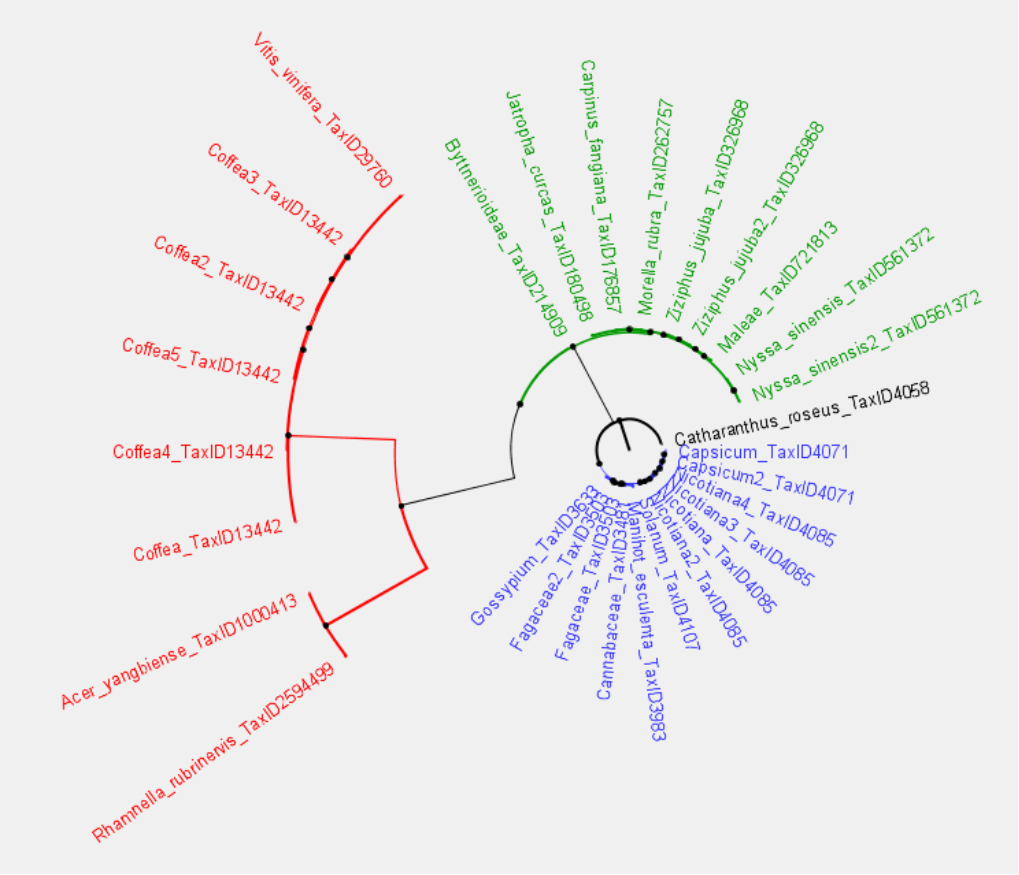


Figura 28. Árbol filogenético circular

Árbol filogenético en versión circular sin los valores de bootstrap. Se aprecia el grupo externo en negro (*Catharanthus roseus*) y, en azul, su clado más cercano.

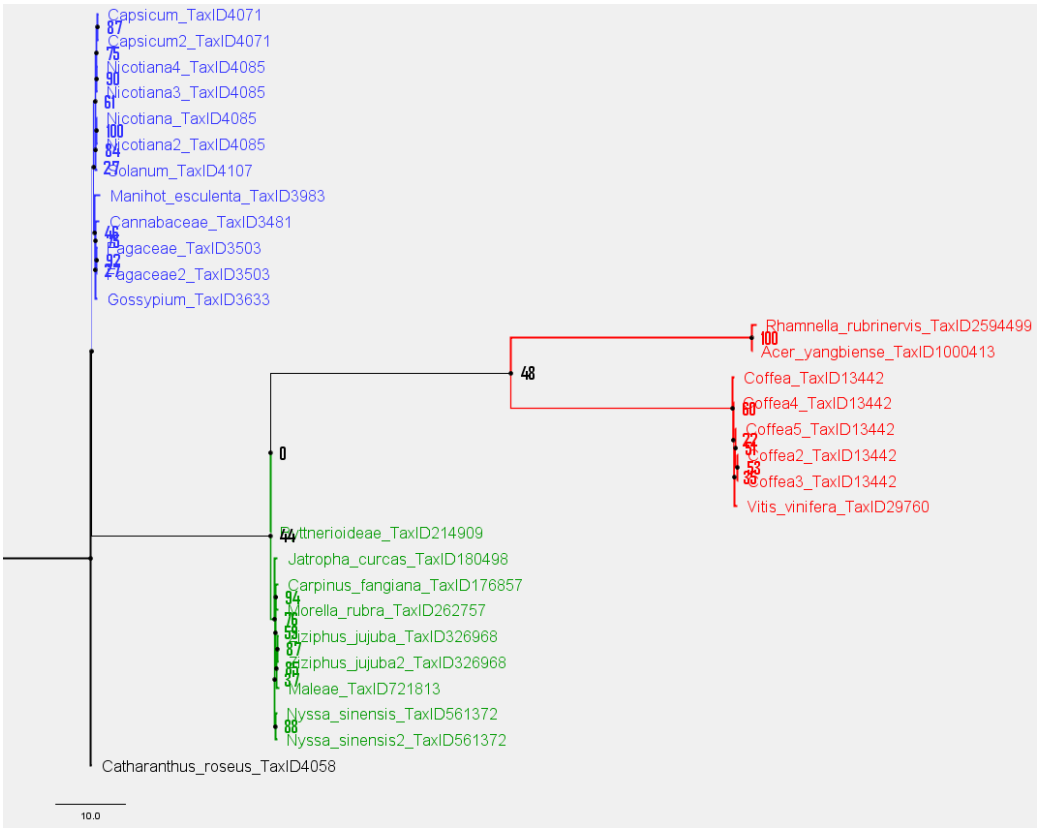


Figura 29. Árbol filogenético rectangular

Árbol filogenético en versión rectangular con los valores de bootstrap. Se aprecia el grupo externo en negro (*Catharanthus roseus*) y, en azul, su clado más cercano. La mayoría de los valores de los bootstrap, dentro de los clados, son superiores a 70.

A pesar de estos resultados, no se puede descartar el hecho de que existan una o varias especies cercanas evolutivamente a *C. roseus* en relación a las secuencias de la ruta biosintética tratada en este estudio. Por ello, es una limitante importante la falta de información genómica y proteómica de plantas medicinales. Apenas existe una base de datos exclusiva de plantas medicinales denominada *Medicinal Plant Genomic Resource* (http://mpgr.uga.edu/) que tiene solo 14 genomas y proteomas; este número resulta insignificante a comparación de bases de datos de otros organismos como bacterias o virus. Además, la gran mayoría de plantas identificadas en la base de datos de *UniProt* incluidas en este proyecto, son de interés para la industria de la alimentación, madera o textiles.

No obstante, como un soporte a los resultados de este estudio, es interesante destacar la filogenia de rutas metabólicas de la base de datos *Plant Metabolic Pathway* (plantcyc.org) (figura 28). Esta recopila información curada sobre diferentes rutas bioquímicas, enzimas, compuestos y genes de varias plantas. En la figura 28 se presenta una sección de la filogenia mostrada en esa base de datos construida en base a rutas bioquímicas de ese grupo de plantas. Se señalan con flechas algunas especies que coinciden con la investigación realizada en este proyecto. Lo más destacable en esta comparación es la coincidencia en la lejanía de *C. roseus* con especies del clado azul y rojo. No se encontró alguna especie del clado verde en la filogenia de esa base de datos.

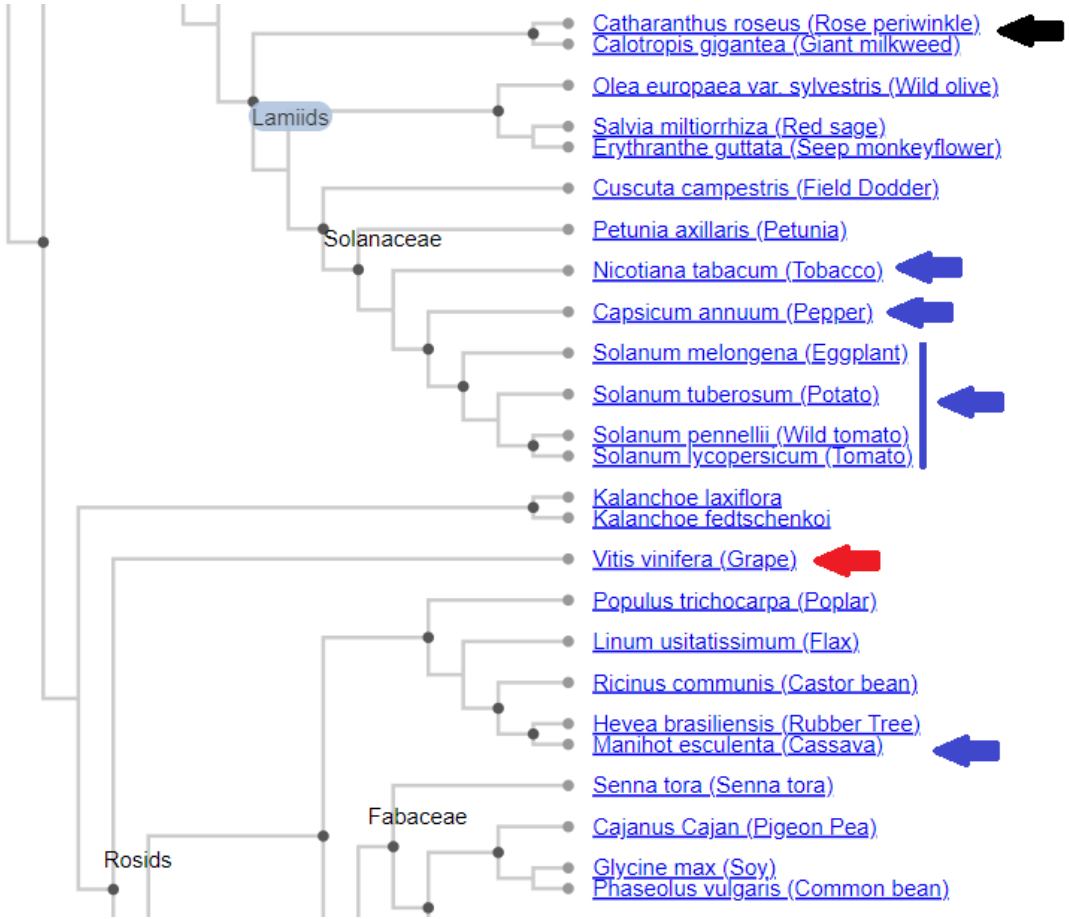


Figura 30. Filogenia de Plant Metabolic Pathway

Filogenia de la base de datos *Plant Metabolic Pathway*. La flecha negra señala a *C. roseus*, las flechas azules a especies del clado azul y, la flecha roja, al clado rojo, entre algunas especies consideradas en el presente trabajo.

6.6. Relación de este estudio con filogenómica

Para justificar el vínculo de la metodología propuesta en este proyecto con filogenómica, me basé en el concepto propuesto por Mike Lee (2019): "*la filogenómica intenta inferir relaciones evolutivas entre secuencias compuestas de múltiples genes concatenados, mientras se asume que esas relaciones evolutivas inferidas nos dicen algo significativo con respecto a las relaciones evolutivas de sus genomas de origen*" (Lee, 2019). Esta es una idea que puede extrapolarse a secuencias de ARN y proteínas. En este estudio, las secuencias de proteínas involucradas y que se concatenaron fueron: D4H, DAT, STR1 y TDC, de la planta *C. roseus*. Como bien apunta el mismo Lee, la mayoría de representaciones evolutivas que los biólogos están acostumbrados a ver, son representaciones de un solo tipo de secuencia (como el gen 16S rRNA). Sin embargo, si nos referimos a otras ideas de filogenómica, se destaca el volumen de datos utilizado. Patané et. al. (2018) señala que la filogenómica es el intento de reconstruir la historia evolutiva de organismos tomando en consideración genomas completos o largas fracciones de los genomas (Patané et al., 2018).

En cualquier caso, la metodología propuesta que involucra la concatenación de cuatro secuencias proteicas, puede aplicarse de manera similar a un mayor volumen de secuencias, a costas de emplear un mayor poder computacional. Aunque esto, a final de cuentas, dependerá de la pregunta de investigación. En el contexto de encontrar relaciones evolutivas de la ruta de biosíntesis de los alcaloides vinblastina y vincristina, podría no hacer falta considerar grandes porciones del genoma o proteoma de las plantas involucradas porque se trata de una ruta bioquímica en concreto, de entre miles que podría tener una planta. Por otro lado, para añadir más secuencias a este estudio, se tendría que conocer a mayor detalle la participación de otras enzimas o biomoléculas relacionadas a la biosíntesis de los alcaloides. Desafortunadamente, aún hay muchas enzimas involucradas en su biosíntesis que no se conocen (Barrales-Cureño et al., 2019).

De existir más secuencias de datos, tanto de la misma planta *C. roseus* sobre la ruta de biosíntesis de los alcaloides, como de plantas emparentadas, los resultados de un estudio como este podrían ser más precisos. Las plantas

encontradas en los resultados de la búsqueda de BLAST en *UniProt*, fueron, en su mayoría, de plantas relacionadas a la industria de alimentación. Es así que, para tener más oportunidades de encontrar mejores relaciones evolutivas, podría ser necesario contar con información del genoma y proteoma de otras especies de la familia *Apocynaceae*, donde se encuentra *C. roseus*. Como por ejemplo, *Vinca minor*, *Trachelospermum jasminoides*, *Mandevilla boliviensis*, entre muchos otros miembros de esta familia.

7. Conclusiones

7.1. No se logró determinar una relación filogenética de la ruta de biosíntesis de los alcaloides vinblastina y vincristina, en plantas diferentes a *Catharanthus roseus*. Sin embargo, no se podría descartar que sí exista un vínculo evolutivo, porque falta información genómica y proteómica de más especies que podrían estar emparentadas. De la lista de 72 plantas empleadas, solamente *Rauvolfia* corresponde a la misma familia de *C. roseus* (esta es, *Apocynaceae*). En la base de datos empleada en este estudio, no se encontró más información de secuencias en la familia *Apocynaceae*. Aunque, es importante puntualizar, que a pesar de contar con información de la misma familia, no hay garantía de que se podría hallar relaciones evolutivas, ya que esto depende de las secuencias homólogas que se estén investigando. En cualquier caso, habría sido positivo contar con más información.

7.2. El análisis de máxima verosimilitud empleado fue una herramienta eficiente para valorar la calidad del árbol filogenético generado. A partir de este análisis, se consideraron como principales parámetros de evaluación de calidad al porcentaje de gaps y caracteres indeterminados (< 50%), y a la cifra final de máxima verosimilitud que tenía que mejorar en cada análisis realizado (acercándose a cero). El valor de apoyo de ramas (bootstrap) fue otra cifra que se analizó, una vez construido el árbol. Se destaca que dentro de los tres clados formados hay apoyos de rama superiores a 70 puntos, en su mayoría, lo que otorga confiabilidad estadística a esas relaciones evolutivas. Pero el apoyo de ramas, entre los clados, es bajo 70 puntos lo que implica poca confiabilidad estadística de los tres clados formados.

7.3. El grupo inicial de secuencias en la matriz concatenada tenía 158 taxones, de 72 especies. En cambio, en el árbol filogenético final de este estudio, se obtuvieron 30 secuencias homólogas concatenadas, que se corresponden con 17 especies. Esto, sin considerar a *C. roseus* que fue el grupo externo del árbol. Esta reducción en el número de taxones se debió a que las primeras matrices tenían valores de calidad bajos (de acuerdo a lo expresado en 8.1.), hasta que, en la matriz final #7, se obtuvieron los valores más apropiados. De las 17 especies del árbol final, se pudieron encontrar seis en el árbol filogenético de la base de datos *Plant Metabolic Pathway* (incluyendo a *C. roseus*). Haciendo una comparación entre el árbol de mi estudio y el de esa base datos (respecto a las seis especies), coincide en que C. roseus no tiene una relación evolutiva cercana con las otras cinco especies (ver figura 28).

8. Recomendaciones

8.1. Para mejorar la fiabilidad del estudio, se pudo haber recolectado información en otras bases de datos bioinformáticas que incluyeran genomas y proteomas sobre plantas en general, y, de existir, con enfoque en plantas medicinales. *Phytozome* y *Medicinal Plant Genomics Resource* son dos opciones a considerar en estudios futuros.

8.2. Se pueden realizar muchas de las etapas del pipeline de este proyecto, dentro de un mismo software como *Geneious,* el cual es de pago y tiene una interfaz de usuario de uso simple e intuitiva. No obstante, las herramientas bioinformáticas usadas aquí son una opción gratuita y de software libre.

8.3. Estudios de este tipo se ven altamente limitados por la ausencia de información ómica, a comparación de otros organismos como bacterias o virus donde las bases de datos son más extensas y variadas. Se debe promover proyectos de secuenciación ómicos en plantas medicinales para impulsar los estudios filogenómicos y de fitofármacos. Cabe destacar que Ecuador es un país con una enorme diversidad de plantas medicinales, de las que suman 3118 especies, siendo el 75% nativas (*Bioknowledgy of the Ecuadorian Flora. Some medicinal plants and their uses.*, s. f.). Hasta lo investigado en este estudio, no se logró encontrar información genómica o proteómica de alguna de estas plantas.

8.4. En la materia de Metodología de la Investigación, recibida en la maestría de Biología Computacional de la PUCE, se pretendía iniciar con el diseño del proyecto de titulación, pero considero que fue un desacierto que tuviéramos esta materia en la primera parte del primer semestre, puesto que muchos de los estudiantes teníamos solamente nociones de lo que es Biología Computacional. Si nos daban esa materia en la segunda parte del primer semestre, probablemente habríamos estado más seguros de nuestro proyecto, con toda la inducción recibida durante la primera parte. De este modo, quizá podríamos haber optimizado los tiempos para desarrollar el proyecto. Dicho esto, quisiera recomendar que para las siguientes promociones se considere un cambio de este tipo, si es que aún no se lo ha hecho.

9. Referencias

ALTER User's Guide. (s. f.). Recuperado 11 de junio de 2023, de https://www.sing-group.org/ALTER/help/ALTER-UserGuide.htm

Arenas, M. (2015). Trends in substitution models of molecular evolution. *Frontiers in Genetics*, *6*. https://www.frontiersin.org/articles/10.3389/fgene.2015.00319

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., & Supuran, C. T. (2021). Natural products in drug discovery: Advances and opportunities. *Nature Reviews Drug Discovery*, *20*(3), Article 3. https://doi.org/10.1038/s41573-020-00114-z

Barrales-Cureño, H. J., Reyes, C. R., García, I. V., Valdez, L. G. L., Jesús, A. G. D., Ruíz, J. A. C., Herrera, L. M. S., Caballero, M. C. C., Magallón, J. A. S., Perez, J. E., Montoya, J. M., Barrales-Cureño, H. J., Reyes, C. R., García, I. V., Valdez, L. G. L., Jesús, A. G. D., Ruíz, J. A. C., Herrera, L. M. S., Caballero, M. C. C., … Montoya, J. M. (2019). Alkaloids of Pharmacological Importance in Catharanthus roseus. En *Alkaloids—Their Importance in Nature and Human Life*. IntechOpen. https://doi.org/10.5772/intechopen.82006

Bash is in beta release! - Gnu.announce | Google Groups. (2013, mayo 4). https://web.archive.org/web/20130504075535/http://groups.google.com/group/gnu.announce/msg/a509f48ffb298c35?hl=en

Bioknowledgy of the Ecuadorian Flora. Some medicinal plants and their uses. (s. f.). Recuperado 23 de junio de 2023, de https://libreriasiglo.com/biologia/128133-bioknowledgy-of-the-ecuadorian-flora-so

me-medicinal-plants-and-their-uses.html

Brown, D., & Sjölander, K. (2006). Functional Classification Using Phylogenomic Inference. *PLoS Computational Biology*, *2*(6), e77. https://doi.org/10.1371/journal.pcbi.0020077

Chik, S. C. C., Or, T. C. T., Luo, D., Yang, C. L. H., & Lau, A. S. Y. (2013). Pharmacological Effects of Active Compounds on Neurodegenerative Disease with Gastrodia and Uncaria Decoction, a Commonly Used Poststroke Decoction. *The Scientific World Journal*, *2013*, 896873. https://doi.org/10.1155/2013/896873

Cibrián-Jaramillo, A., De la Torre-Bárcena, J. E., Lee, E. K., Katari, M. S., Little, D. P., Stevenson, D. W., Martienssen, R., Coruzzi, G. M., & DeSalle, R. (2010). Using Phylogenomic Patterns and Gene Ontology to Identify Proteins of Importance in Plant Evolution. *Genome Biology and Evolution*, *2*, 225-239. https://doi.org/10.1093/gbe/evq012

Cummings, M. P. (2004). PAUP* [Phylogenetic Analysis Using Parsimony (and Other Methods)]. En *Dictionary of Bioinformatics and Computational Biology*. John Wiley & Sons, Ltd. https://doi.org/10.1002/0471650129.dob0522

Dayhoff, Margaret. (1978). *22 A Model of Evolutionary Change in Proteins*. https://www.semanticscholar.org/paper/22-A-Model-of-Evolutionary-Change-in-Proteins-O./ff3e2e966bf5ee6671dd757357a375b2463b072c

De Luca, V., Fernandez, J. A., Campbell, D., & Kurz, W. G. (1988). Developmental Regulation of Enzymes of Indole Alkaloid Biosynthesis in Catharanthus roseus. *Plant Physiology*, *86*(2), 447-450. https://doi.org/10.1104/pp.86.2.447

Digital Library Of The Commons. (s. f.). Recuperado 2 de abril de 2023, de https://dlc.dlib.indiana.edu/dlc/handle/10535/8426

Dopico, E., San Fabian, J. L., & Garcia-Vazquez, E. (2008). Traditional Medicine in Twenty-first Spain. *Human Ecology*, *36*(1), 125-129. https://doi.org/10.1007/s10745-007-9146-1

Eisen, J. A. (1998). Phylogenomics: Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis. *Genome Research*, *8*(3), 163-167. https://doi.org/10.1101/gr.8.3.163

FigTree. (s. f.). Recuperado 3 de junio de 2023, de http://tree.bio.ed.ac.uk/software/figtree/

Fitch, W. M. (1970). Distinguishing Homologous from Analogous Proteins. *Systematic Biology*, *19*(2), 99-113. https://doi.org/10.2307/2412448

Gee, H. (2003). Ending incongruence. *Nature*, *425*(6960), Article 6960. https://doi.org/10.1038/425782a

Goddijn, O. J., Pennings, E. J., van der Helm, P., Schilperoort, R. A., Verpoorte, R., & Hoge, J. H. (1995). Overexpression of a tryptophan decarboxylase cDNA in Catharanthus roseus crown gall calluses results in increased tryptamine levels but not in increased terpenoid indole alkaloid production. *Transgenic Research*, *4*(5), 315-323. https://doi.org/10.1007/BF01972528

Heinrich, M., Mah, J., & Amirkia, V. (2021). Alkaloids Used as Medicines: Structural Phytochemistry Meets Biodiversity—An Update and Forward Look. *Molecules*, *26*(7), 1836. https://doi.org/10.3390/molecules26071836

High Throughput Sequencing—An overview | ScienceDirect Topics. (2019). https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/high-throughput-sequencing

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics*, *20*(4), 1160-1166. https://doi.org/10.1093/bib/bbx108

Kaur, R., & Arora, S. (2015, junio 30). *ALKALOIDS-IMPORTANT THERAPEUTIC SECONDARY METABOLITES OF PLANT ORIGIN*. https://www.semanticscholar.org/paper/ALKALOIDS-IMPORTANT-THERAPEUTIC-SECONDARY-OF-PLANT-Kaur-Arora/a84acf030107b170e7787cecdc5ba9b4f6712c8b

Koonin, E. V. (2005). *Orthologs, Paralogs, and Evolutionary Genomics 1*. https://doi.org/10.1146/annurev.genet.39.073003.114725

Kurek, J. (2019). Introductory Chapter: Alkaloids - Their Importance in Nature and for Human Life. En *Alkaloids—Their Importance in Nature and Human Life*. IntechOpen. https://doi.org/10.5772/intechopen.85400

Lee, M. (2019). An Introduction to Phylogenomics. *Happy Belly Bioinformatics*. https://astrobiomike.github.io/genomics/phylogenomics

Lemmon, A. R., & Moriarty, E. C. (2004). The importance of proper model assumption in

bayesian phylogenetics. *Systematic Biology*, *53*(2), 265-277. https://doi.org/10.1080/10635150490423520

Loyola-Vargas, V. M., Sánchez-Iturbe, P., Canto-Canché, B., Gutiérrez-Pacheco, L. C., Galaz-Ávalos, R. M., & Moreno-Valenzuela, O. (2004). Biosíntesis de los alcaloides indólicos: Una revisión crítica. *Revista de la Sociedad Química de México*, *48*(1), 67-94.

Lozano-Fernandez, J. (2022). A Practical Guide to Design and Assess a Phylogenomic Study. *Genome Biology and Evolution*, *14*(9), evac129. https://doi.org/10.1093/gbe/evac129

Maddison, D. R., Swofford, D. L., & Maddison, W. P. (1997). Nexus: An Extensible File Format for Systematic Information. *Systematic Biology*, *46*(4), 590-621. https://doi.org/10.1093/sysbio/46.4.590

Medicinal Botany—Active Plant Ingredients. (s. f.). Recuperado 8 de abril de 2023, de https://www.fs.usda.gov/wildflowers/ethnobotany/medicinal/ingredients.shtml

Mesilaakso, L. (2019). *Bioinformatic approaches for detecting homologous genes in the genomes of non-model organisms: A case study of wing development genes in insect genomes*. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-398072

Mesquite Project. (s. f.). Recuperado 3 de junio de 2023, de https://www.mesquiteproject.org/

Miller, M. A., Pfeiffer, W., & Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. *2010 Gateway Computing Environments Workshop (GCE)*, 1-8. https://doi.org/10.1109/GCE.2010.5676129

O'Brien, S. J., & Stanyon, R. (1999). Ancestral primate viewed. *Nature*, *402*(6760), Article 6760. https://doi.org/10.1038/46450

Of Terms in Biology: Phylogeny, Phylogenetics, Phylogenomics. (s. f.). Small Things Considered. Recuperado 11 de abril de 2023, de https://schaechter.asmblog.org/schaechter/2020/09/of-terms-in-biology-.html

Patané, J. S. L., Martins, J., & Setubal, J. C. (2018). Phylogenomics. En J. C. Setubal, J. Stoye, & P. F. Stadler (Eds.), *Comparative Genomics: Methods and Protocols* (pp. 103-187). Springer. https://doi.org/10.1007/978-1-4939-7463-4_5

Petrovska, B. B. (2012). Historical review of medicinal plants' usage. *Pharmacognosy*

Reviews, *6*(11), 1-5. https://doi.org/10.4103/0973-7847.95849

PHYLIP multiple sequence alignment format (skbio.io.phylip)—Scikit-bio 0.2.3 documentation. (s. f.). Recuperado 11 de junio de 2023, de http://scikit-bio.org/docs/0.2.3/generated/skbio.io.phylip.html

Posada, D., & Crandall, K. A. (1998). MODELTEST: Testing the model of DNA substitution. *Bioinformatics (Oxford, England)*, *14*(9), 817-818. https://doi.org/10.1093/bioinformatics/14.9.817

Probability Reduced Evolution of Spatially-discrete Species. (s. f.). Recuperado 21 de junio de 2023, de https://www.biosym.uzh.ch/modules/models/BootStrapTree/Theory.html

Ran, J.-H., Shen, T.-T., Wang, M.-M., & Wang, X.-Q. (2018). Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1881), 20181012. https://doi.org/10.1098/rspb.2018.1012

Sequence. (s. f.). Recuperado 11 de junio de 2023, de https://evolution.genetics.washington.edu/phylip/doc/sequence.html

Shafee, T. (2018). *English: Top: An ancestral gene duplicates to produce two paralogs (histone H1.1 and 1.2). A speciation event produces orthologs in the two daughter species (human and chimpanzee). Bottom: in a separate species (E. coli), an gene has a similar function (histone-like nucleoid-structuring protein) but has a separate evolutionary origin and so is an analog.* Own work. https://commons.wikimedia.org/wiki/File:Ortholog_paralog_analog_examples.svg

Sjölander, K. (2004). Phylogenomic inference of protein molecular function: Advances and challenges. *Bioinformatics*, *20*(2), 170-179. https://doi.org/10.1093/bioinformatics/bth021

Sottomayor, M., Lopes Cardoso, I., Pereira, L. G., & Ros Barceló, A. (2004). Peroxidase and the biosynthesis of terpenoid indole alkaloids in the medicinal plant Catharanthus roseus (L.) G. Don. *Phytochemistry Reviews*, *3*(1), 159-171. https://doi.org/10.1023/B:PHYT.0000047807.66887.09

Stamatakis, A. (2008). The RAxML 7.0.4 manual. *Department of Computer Science. Ludwig-Maximilians-Universität München*.

The A-Z of Programming Languages: BASH/Bourne-Again Shell - a-z of programming languages - Computerworld. (2011, julio 6). https://web.archive.org/web/20110706103704/http://www.computerworld.com.au/article/222764/a-z_programming_languages_bash_bourne-again_shell/?pp=2&fp=16&fpid=1

UniProt: A hub for protein information. (2015). *Nucleic Acids Research*, *43*(Database issue), D204-D212. https://doi.org/10.1093/nar/gku989

van der Fits, L., & Memelink, J. (2000). ORCA3, a jasmonate-responsive transcriptional regulator of plant primary and secondary metabolism. *Science (New York, N.Y.)*, *289*(5477), 295-297. https://doi.org/10.1126/science.289.5477.295

Warren, T. (2019, junio 4). *Apple replaces bash with zsh as the default shell in macOS Catalina*. The Verge. https://www.theverge.com/2019/6/4/18651872/apple-macos-catalina-zsh-bash-shell-replacement-features

Young, A. D., & Gillung, J. P. (2020). Phylogenomics—Principles, opportunities and pitfalls of big-data phylogenetics. *Systematic Entomology*, *45*(2), 225-247. https://doi.org/10.1111/syen.12406

Zhu, J., Wang, M., Wen, W., & Yu, R. (2015). Biosynthesis and regulation of terpenoid indole alkaloids in Catharanthus roseus. *Pharmacognosy Reviews*, *9*(17), 24-28. https://doi.org/10.4103/0973-7847.156323

10. Anexos

Tabla 1. Presencia/ausencia de secuencias de proteínas homólogas en relación a *C. roseus*

	Planta	Taxon ID (UniProt)	Nombre común	Familia	D4H	DAT	STR 1	TDH	Observación
1	*Catharanthus roseus*	4058	Chavelita, alegría de la casa, cortejo.	Apocynaceae	Sí	Sí	Sí	Sí	Más de una secuencia homóloga en DAT
2	*Nyssa sinensis*	561372	Tupelo chino	Cornaceae	Sí	Sí	Sí	No	Más de una secuencia homóloga en D4H y STR1
3	*Olea europea subsp. europaea*	158383	Olivo	Oleaceae	Sí	No	No	No	Más de una secuencia homóloga en D4H
4	*Vitis*	3603	-	Vitaceae	Sí	No	No	No	Más de una secuencia homóloga en D4H
5	*Byttnerioideae*	214909	-	Malvaceae	Sí	No	No	Sí	Más de una secuencia homóloga en D4H
6	*Durio zibethinus*	66656	Durián	Malvaceae	Sí	No	No	No	Más de una secuencia homóloga en D4H
7	*Fagaceae*	3503	-	Fagaceae	Sí	Sí	No	Sí	Más de una secuencia homóloga en D4H y DAT
8	*Ziziphus jujuba*	326968	Jujube, jinjolero, azufaito...	Rhamnaceae	Sí	Sí	Sí	Sí	Más de una secuencia homóloga en D4H y STR1
9	*Coffea*	13442	Café	Rubiaceae	Sí	Sí	Sí	Sí	Más de una secuencia homóloga en D4H, STR1 y TDC
10	*Aquilegia coerulea*	218851	Aguileña	Ranunculaceae	Sí	No	No	No	Más de una secuencia homóloga en D4H
11	*Juglandaceae*	16714	-	Juglandaceae	Sí	No	No	No	Más de una secuencia homóloga en D4H
12	*Quercus*	3511	-	Fagáceas	Sí	No	No	No	
13	*Trema orientale*	63057	Árbol de carbón, olmo africano		Sí	No	No	No	
14	*Maleae*	721813	-	Rosáceas	Sí	Sí	No	No	Más de una secuencia homóloga en DAT
15	*Acer yangbiense*	1000413	Arce	Sapindaceae	Sí	No	Sí	No	
16	*Camellia sinensis var. sinensis*	542762	Planta del té	Theaceae	Sí	No	No	No	Más de una secuencia homóloga en D4H
17	*Cinnamomum micranthum f. kanehirae*	337451	-	Lauráceas	Sí	No	No	No	Más de una secuencia homóloga en D4H

	Planta	Taxon ID (UniProt)	Nombre común	Familia	D4H	DAT	STR1	TDH	Observación
18	*Rhamnella rubrinervis*	2594499	-	Rhamnaceae	Sí	No	No	No	
19	*Jatropha curcas*	180498	Piñón de tempate o jatropa	Euphorbiaceae	Sí	Sí	No	Sí	Más de una secuencia homóloga en D4H
20	*Malvaceae*	3629	-	Malvaceae	Sí	No	No	No	Más de una secuencia homóloga en D4H
21	*Thalictrum thalictroides*	46969	Ruda anémona	Ranunculaceae	Sí	No	No	No	
22	*Castanea mollissima*	60419	Castaño chino	Fagaceae	Sí	No	No	No	
23	*Cannabaceae*	3481	-	Cannabaceae	Sí	No	Sí	No	
24	*Nicotiana*	4085	-	Solanáceas	Sí	Sí	No	Sí	Más de una secuencia homóloga en D4H, DAT y TDC
25	*Gossypium*	3633	Plantas de algodón	Malvaceae	Sí	No	No	Sí	Más de una secuencia homóloga en TDC
26	*Manihot esculenta*	3983	Yuca	Euphorbiaceae	Sí	Sí	No	No	Más de una secuencia homóloga en DAT
27	*Phaseolus vulgaris*	3885	Fréjol	Fabaceae	Sí	No	No	No	
28	*Vitis vinifera*	29760	Uva	Vitaceae	Sí	No	Sí	No	Más de una secuencia homóloga en STR1
29	*Solanaceae*	4070	-	Solanaceae	No	Sí	No	No	Más de una secuencia homóloga en DAT
30	*Capsicum*	4071	Pimiento, ajíes...	Solanaceae	No	Sí	No	No	Más de una secuencia homóloga en DAT
31	*Nicotiana tabacum*	4097	Tabaco	Solanaceae	No	Sí	No	No	
32	*Coffea arabica*	13443	Café	Rubiaceae	No	Sí	No	No	Más de una secuencia homóloga en DAT
33	*Rosa chinensis*	74649	Rosa china	Rosaceae	No	Sí	No	No	Más de una secuencia homóloga en DAT
34	*Solanum subgen. Lycopersicon*	49274	Tomate	Solanaceae	No	Sí	No	No	Más de una secuencia homóloga en DAT
35	*Prunus*	3754	-	Rosaceae	No	Sí	No	No	Más de una secuencia homóloga en DAT
36	*Senna tora*	362788	Orozuz, guanima	Fabaceae	No	Sí	No	No	
37	*Populus*	3689	-	Salicaceae	No	Sí	No	No	Más de una secuencia homóloga en DAT

	Planta	Taxon ID (UniProt)	Nombre común	Familia	D4H	DAT	STR1	TDH	Observación
38	*Solanum*	4107	-	Solanaceae	No	Sí	No	Sí	Más de una secuencia homóloga en TDC
39	*Pyrus ussuriensis x Pyru communis*	2448454	Pera de Ussuri	Rosaceae	No	Sí	No	No	
40	*Rauvolfia*	4059	-	Apocynaceae	No	No	Sí	No	Más de una secuencia homóloga en STR1
41	*Morella rubra*	262757	Arrayán chino	Myricaceae	No	No	Sí	Sí	Más de una secuencia homóloga en STR1
42	*Cicer arietinum*	3827	Garbanzo	Fabaceae	No	No	Sí	No	
43	*Dorcoceras hygrometricum*	472368	-	Gesneriaceae	No	No	Sí	No	
44	*Phtheirospermum japonicum*	374723	-	Orobanchaceae	No	No	Sí	No	
45	*Populus deltoides*	3696	Álamo	Salicaceae	No	No	Sí	No	
46	*Carpinus fangiana*	176857	Carpe cola de mono	Betulaceae	No	No	Sí	Sí	
47	*Trifolium*	3898	-	Fabaceae	No	No	Sí	No	Más de una secuencia homóloga en STR1
48	*Brassicaceae*	3700	-	Brassicaceae	No	No	Sí	No	Más de una secuencia homóloga en STR1
49	*Lupinus*	3869	-	Fabaceae	No	No	Sí	No	Más de una secuencia homóloga en STR1
50	*Colocasia esculenta*	4460	Taro, pituca o malanga	Aráceas	No	No	Sí	No	
51	*Eutrema salsugineum*	72664	Berro de agua salada	Brassicaceae	No	No	Sí	No	
52	*Brassica*	3705	-	Brassicaceae	No	No	Sí	No	Más de una secuencia homóloga en STR1
53	*Camelineae*	980083	-	Brassicaceae	No	No	Sí	No	
54	*Handroanthus impetiginosus*	429701	Árbol trompeta rosa	Bignoniaceae	No	No	Sí	No	
55	*Coluteocarpeae*	1394505	-	Brassicaceae	No	No	Sí	No	
56	*Tetracentron sinense*	13715	-	Trochodendraceae	No	No	No	Sí	Más de una secuencia homóloga en TDC
57	*Capsicum annuum*	4072	Pimiento	Solanaceae	No	No	No	Sí	
58	*Capsicum*	4071	-	Solanaceae	No	No	No	Sí	Más de una secuencia homóloga en TDC

	Planta	Taxon ID (UniProt)	Nombre común	Familia	D4H	DAT	STR 1	TDH	Observación
59	*Macleaya cordata*	56857	Penacho de amapola	Papaveraceae	No	No	No	Sí	Más de una secuencia homóloga en TDC
60	*Nyssaceae*	4289	-	Nyssaceae	No	No	No	Sí	Más de una secuencia homóloga en TDC
61	*Cuscuta sect. Cleistogrammica*	1824901	-	Convolvulaceae	No	No	No	Sí	Más de una secuencia homóloga en TDC
62	*Actinidia*	3624	-	Actinidiaceae	No	No	No	Sí	
63	*Citrus*	2706	-	Rutaceae	No	No	No	Sí	Más de una secuencia homóloga en TDC
64	*Cynara cardunculus var. scolymus*	59895	Alcachofa	Asteraceae	No	No	No	Sí	
65	*Nelumbo nucifera*	4432	Lotus	Nelumbonaceae	No	No	No	Sí	
66	*Artemisia annua*	35608	Artemisa dulce	Asteraceae	No	No	No	Sí	
67	*Camptotheca acuminata*	16922	Árbol feliz	Cornaceae	No	No	No	Sí	
68	*Quercus lobata*	97700	Roble valle	Fagaceae	No	No	No	Sí	
69	*Rhamnella rubrinervis*	2594499	-	Rhamnaceae	No	No	No	Sí	
70	*Kalanchoe fedtschenkoi*	63787	Vieiras jaspeadas de lavanda	Crassulaceae	No	No	No	Sí	
71	*Corchorus olitorius*	93759	Yute nalta	Malvaceae	No	No	No	Sí	
72	*Eucalyptus grandis*	71139	Chicle de rosa	Myrtaceae	No	No	No	Sí	