Universitat Oberta de Catalunya

uoc.edu

Análisis filogenético de la óxido nítrico reductasa (Nor): establecimiento del perfil de la proteína

Ana Belén Valverde Guirao
Máster de Bioinformática y Bioestadística
Área 4. Subárea 6: Microbiología, biotecnología y biología molecular

Consultor/a
Paloma Pizarro Tobías

Profesor responsable de la asignatura
Antoni Pérez Navarro

06/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Análisis filogenético de la óxido nítrico reductasa (Nor): establecimiento del perfil de la proteína*
Nombre del autor:	*Ana Belén Valverde Guirao*
Nombre del consultor/a:	*Paloma Pizarro Tobías*
Nombre del PRA:	*Antoni Pérez Navarro*
Fecha de entrega (mm/aaaa):	06/2021
Titulación:	*Máster universitario de Bioinformática y Bioestadística*
Área del Trabajo Final:	*Microbiología, biotecnología y biología molecular*
Idioma del trabajo:	Castellano
Número de créditos:	15
Palabras clave	*nitric oxide reductase, phylogenetic analysis, protein profile*

Resumen del Trabajo (máximo 250 palabras):

La finalidad de este trabajo fue la realización de un análisis filogenético, así como la construcción del perfil de la proteína óxido nítrico reductasa(Nor). Esta enzima contribuye a la formación de un potente gas de efecto invernadero, el óxido nitroso (N_2O), que participa en la destrucción de la capa de ozono.

El punto de partida del estudio fue la creación de una base de datos de proteínas Nor bacterianas mediante PSI-BLAST. La metodología llevada a cabo en este trabajo se basó en el uso de distintas herramientas bioinformáticas como MEGA-X, ModelTest o TREE-PUZZLE para el análisis filogenético, así como HMMER para la construcción del perfil. Además, se utilizó R para la creación de pequeños programas que ayudaron a continuar con el estudio, además de otros programas.

Los resultados obtenidos mostraron que el árbol resultante agrupaba las secuencias en dos clados distintos. En uno de ellos predominaban bacterias de la familia *Rhizobiaceae*, mientras que, en el otro, la mayoría eran del orden *Rhodobacterales.* Por otro lado, se vio que los perfiles obtenidos eran capaces de reconocer proteínas anotadas como Nor en una base de datos de proteínas, así como proteínas denotadas como *unclassified*.

Los resultados apuntaron a que los perfiles podrían usarse para seguir estudiando el papel que tiene la enzima Nor en aumento de los niveles de N_2O atmosférico. De este modo, se podrían llevar a cabo distintos abordajes para mitigar este grave problema a nivel medio ambiental.

Abstract (in English, 250 words or less):

This project aimed to carry out a phylogenetic analysis and the development of the profile of the nitric oxide reductase (Nor). This enzyme contributes to the production of nitrous oxide (N_2O), which is a powerful greenhouse gas, and it participates in Ozone Depletion.

The starting point of the study was the obtaining of a database of bacterial Nor proteins through PSI-BLAST. The methodology carried out was based on the use of different bioinformatics frameworks. They were used software such as MEGA-X, ModelTest or TREE-PUZZLE for the phylogenetic analysis, and HMMER for the development of the profile. In addition, R was used for the creation of small programs that helped to continue with the study, among other programs.

The results showed that the tree organized the sequences in two different clades. In one of them, bacteria of the *Rhizobiaceae* family predominated, while in the other, the majority were of the order Rhodobacterales. On the other hand, it was found that the sequence-based profiles were able to identify proteins annotated as Nor in a protein database and proteins denoted as unclassified, as well.

The results suggested that the profiles could be used in the future to study the role of Nor in increasing levels of atmospheric N_2O. In this way, different approaches could be carried out to mitigate this serious environmental problem.

Índice

Lista de figuras

Lista de tablas

1 Introducción

1.1 Contexto y justificación del Trabajo

La enzima de estudio en este trabajo es la óxido nítrico reductasa o Nor, que se encuentra catalizando una reacción cuyo producto es el óxido nitroso (N_2O). Este gas resulta ser uno de los principales gases de efecto invernadero, además de contribuir al agotamiento de la capa de ozono.

Algunos autores piensan que debería ser posible reducir las emisiones de N2O mediante el estudio a nivel enzimático y microbiológico de la desnitrificación, de forma que apareciesen protocolos de manipulación del suelo a nivel físico y químico para modificar la fisiología de los microorganismos desnitrificantes y, por tanto, disminuir las emisiones de óxido nitroso [1].

El análisis filogenético y la posterior determinación del perfil de la proteína pueden ser un punto de partida para una profunda investigación contra el cambio climático y la destrucción de la capa de ozono, ya que el uso de perfiles proteicos es un abordaje útil en la caracterización de enzimas implicadas en ciclos biogenéticos.

1.2 Objetivos del Trabajo

Objetivos generales

- Estudiar las relaciones filogénicas entre las secuencias proteicas anotadas como Nor (nitric oxide reductase).
- Establecer el perfil de la enzima óxido nítrico reductasa.

Objetivos específicos

- Crear una base de datos propia de secuencias proteicas a partir del resultado del PSI-BLAST y realizar el alineamiento múltiple de las mismas.
- Construir el árbol filogenético y realizar el *bootstrapping*.
- Construir tanto perfiles proteicos como clados muestre el árbol a partir de los alineamientos múltiples de las secuencias contenidas.
- Validar el/los perfil/es generados.

1.3 Enfoque y método seguido

Como punto de partida se acudió a la bibliografía para conocer qué microorganismo bacteriano desnitrificante se encontraba mejor caracterizado, ya que la secuencia de la proteína Nor de dicho organismo se usaría como molde para la búsqueda de homólogos lejanos en PSI-BLAST (Position Specific Iterative BLAST). Se decidió usar este algoritmo porque es rápido, eficiente y permite buscar en grandes bases de datos, además de ser útil para recuperar homólogos lejanos. La finalidad de este punto era conseguir una base de datos propia y depurada de secuencias aminoacídicas homólogas de óxido nítrico reductasas bacterianas. Para la realización de la búsqueda en PSI-BLAST, se llevaron a cabo tantas iteraciones como fueron posibles para recuperar la mayor cantidad de secuencias homólogas lejanas.

Una vez construida la base datos propia, resultaba necesario validar que dicha base de datos era adecuada para continuar con el análisis. Para la validación, se decidió volver a realizar una búsqueda en PSI-BLAST, pero usando otra secuencia como molde. En este caso, la secuencia de Nor pertenecía a otra conocida bacteria desnitrificante. Una vez obtenido este segundo archivo multi-FASTA, se compararon ambos mediante un pequeño programa creado en R. Tras encontrar que ambas búsquedas presentaban una gran cantidad de secuencias comunes, era seguro continuar usando la base de datos propia.

Una vez validada la base de datos, se procedió a realizar el alineamiento múltiple de las secuencias contenidas en esta. Dicho alineamiento se llevó a cabo en MEGA X mediante MUSCLE. Se escogió el software MUSCLE porque es el que mejor resultados proporciona cuando se trata de secuencias proteicas. Además, presenta una mayor precisión y rapidez que otros softwares como ClustalW2 o T-Coffee.

El siguiente paso consistió en determinar el modelo evolutivo que mejor explicaba los datos. En este caso se hizo uso de más de una herramienta, ya que se tenía un fácil acceso a ellas y era una forma de confirmar el modelo con mayor seguridad. Para la determinación del modelo, empleando el alineamiento múltiple como input, se usó MEGA X, el servidor web de IQ-TREE y el programa ModelTest-NG. Además de hallar el modelo evolutivo también se estudió la señal filogenética presente entre las secuencias de la base de datos propia mediante el software TREE PUZZLE. Este último programa precisa que el alineamiento se encuentre en formato Phylip. Para realizar la conversión de FASTA a Phylip se usó la herramienta en línea llamada EMBOSS seqret.

Tras confirmar que las secuencias presentaban una alta señal filogenética, se procedió a la construcción del árbol. Este punto fue crítico en el proceso, ya que en un principio de quería usar R, pero finalmente se empleó MEGA X, dado su sencillo e intuitivo manejo. Además, dado que en toda la bibliografía revisada

sobre filogenia se recomendaba el uso de un método de comprobación, se aplicó el método de *Bootstrap*.

Finalmente, como última etapa de la parte experimental del trabajo, se obtuvieron los perfiles HMM (Hidden Markov Model) de cada uno de los dos clados del árbol. En primer lugar, se extrajeron las etiquetas de cada clado mediante MEGA X y seguidamente, usando un programa creado en R y a partir del archivo multi-FASTA de todas las secuencias de la base datos, se obtuvieron dos nuevos archivos FASTA que contenían las secuencias aminoacídicas de cada uno de los clados. A continuación, se alinearon las secuencias de ambos archivos mediante MUSCLE en MEGA X y se obtuvieron las secuencias consenso de cada clado mediante EMBOSS Cons. Seguidamente, se procedió a la obtención de los perfiles. Como primer paso, se usó *hmmbuild* para construir los perfiles a partir de los alineamientos múltiples de las secuencias de cada clado. Por último, se usó *hmmsearch* para validad cada uno de los perfiles, ya que estos se enfrentaron a la base datos de proteínas UniRef100, que fue previamente descargada localmente, para estudiar cuántas secuencias eran capaces de reconocer dichos perfiles.

1.4 Planificación del Trabajo

Tareas

1) Buscar una proteína molde adecuada en la bibliografía para usarla como *query* del PSI-BLAST. Realizar el PSI-BLAST. (8 días)

2) Realizar el alineamiento múltiple en MUSCLE de las secuencias obtenidas en el output del PSI-BLAST. (8 días)

3) Determinación del modelo evolutivo (8 días)

4) Estudiar la señal filogenética (8 días)

5) Construir el árbol filogenético y realizar el *bootstrapping*. (6 días)

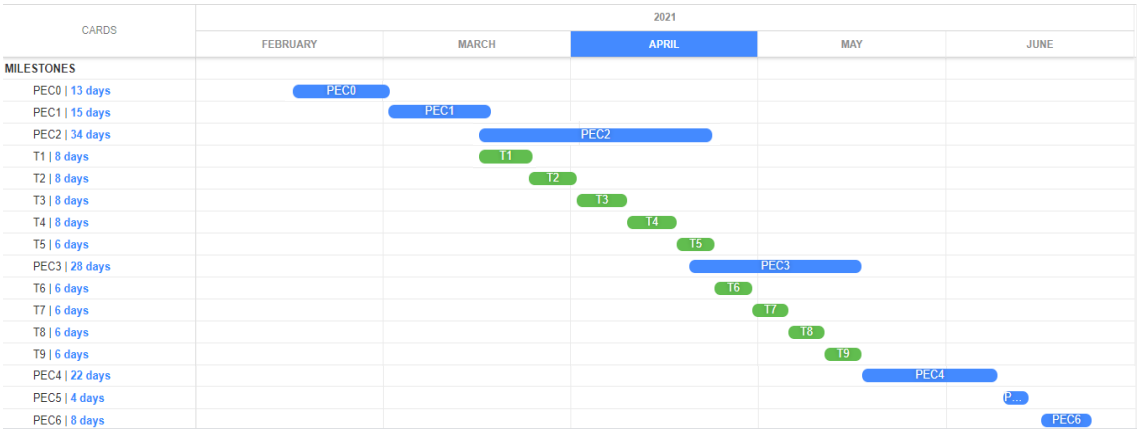
6) Construir el/los perfil/es proteico/s. (6 días)

7) Validar el/los perfil/es. (6 días)

8) Enfrentar el/los perfil/es a un conjunto de proteínas que contienen proteínas no caracterizadas y analizar el resultado. (6 días)

9) Validar el perfil obtenido en genomas completos y bibliotecas metagenómicas. Analizar el resultado. (6 días)

Calendario



Hitos

A continuación, se muestra un cuadro con los hitos en relación con las tareas y temporalización de estás para alcanzarlos:

Hito	Tareas	Periodo de tiempo	
Crear una base de datos propia de secuencias proteicas a partir del resultado del PSI-BLAST y realizar el alineamiento múltiple de las mismas	T1 y T2	17/03/2021 - 01/04/2021	PEC2
Determinación del modelo evolutivo y estudio de la señal filogenética	T3 y T4	02/04/2021 - 17/04/2021	
Construir el árbol filogenético y realizar el *bootstrapping*	T5	18/04/2021 - 23/04/2021	PEC3
Construir el perfil proteico a partir del alineamiento múltiple	T6	24/04/2021 - 29/04/2021	
Validar el perfil generado	T7, T8 y T9	30/04/2021 - 17/05/2021	

Tabla 1. Planificación temporal de cada tarea

1.5 Breve sumario de contribuciones y productos obtenidos

La propósito de este trabajo fue principalmente la obtención de dos productos que contribuyan a ampliar el conocimiento sobre al proceso de desnitrificación, así como su impacto sobre el medio ambiente, con la finalidad de intentar mitigarlo.

En primer lugar, se obtuvo un análisis filogenético a partir del alineamiento múltiple de secuencias de proteínas denotadas como enzimas Nor en diferentes especies bacterianas, usando como proteína molde la óxido nítrico reductasa de *Paracoccus denitrificans.*

Como segundo producto se pretendía obtener uno o varios perfiles de la proteína Nor, que contribuiría a la caracterización y mayor conocimiento de esta enzima de interés medioambiental.

1.6 Breve descripción de los otros capítulos de la memoria

En la memoria se encontrarán distintos capítulos como el de estado del arte, la metodología, resultados y discusión y, finalmente, conclusiones.

En el capítulo de estado del arte se expondrá la información que se ha encontrado en la bibliografía sobre la enzima de estudio, que sirve para contextualizar el resto de los capítulo del trabajo. Además, se incorpora por qué el trabajo es importante y aporta conocimiento de interés a la comunidad científica.

Un capítulo de relevancia será el de metodología, ya que se describirá detalladamente cada uno de los pasos seguidos, tanto para el análisis filogenético como para la obtención del perfil, así como el software empleado. En el caso del análisis filogenético se encontrarán apartados referentes a la construcción de la base de datos propia, la comprobación de la validez de dicha base de datos para el estudio, el alineamiento múltiple de las secuencias, la determinación del modelo evolutivo, el estudio de la señal filogenética y la construcción del árbol. En el caso de la construcción de los perfiles, se detallarán los pasos seguidos para seleccionar las secuencias de cada clado del árbol para poder alinearlas y obtener los distintos perfiles. Además, se nombrará el software empleado para la construcción y validación de los perfiles.

Por otro lado, en el capítulo de resultados y discusión se indicarán los resultados obtenidos tras aplicar cada uno de los procedimientos detallados en la metodología.

Finalmente, el capítulo de conclusiones se encontrará dividido en distintos apartados que describirán las conclusiones obtenidas tras las realización del trabajo, que incluirá una reflexión crítica en relación con los objetivos planteados, líneas de futuro y el seguimiento de la planificación.

2 Estado del arte

La desnitrificación es un proceso metabólico llevado a cabo por diferentes microorganismo en ausencia de oxígeno[2]. Se considera un proceso anóxico, ya que se trata de microrganismos aerobios facultativos, centrándose la mayoría de los estudios sobre desnitrificación en bacterias Gram-negativas[3]. Dicho proceso consiste en la reducción progresiva de diferentes moléculas nitrogenadas como NO_3, NO_2, NO, N_2O y siendo el producto final el N_2[2]. Estas reacciones son catalizadas por distintas reductasas como nitrato reductasas(Nap o Nar), nitrito reductasas (NirK / NirS), óxido nítrico reductasas (cNor, qNor o CuANor) y la óxido nitroso reductasa(Nos) codificadas por los genes nap / nar, nirK / nirS, nor y nos, respectivamente[3]. En la figura 1 se muestran algunas de estas enzimas, así como su ubicación celular y dependencia de cofactores metálicos. La vía que se muestra es representativa de la que se encuentra en muchas bacterias Gram negativas, como *Pa. denitrificans*.

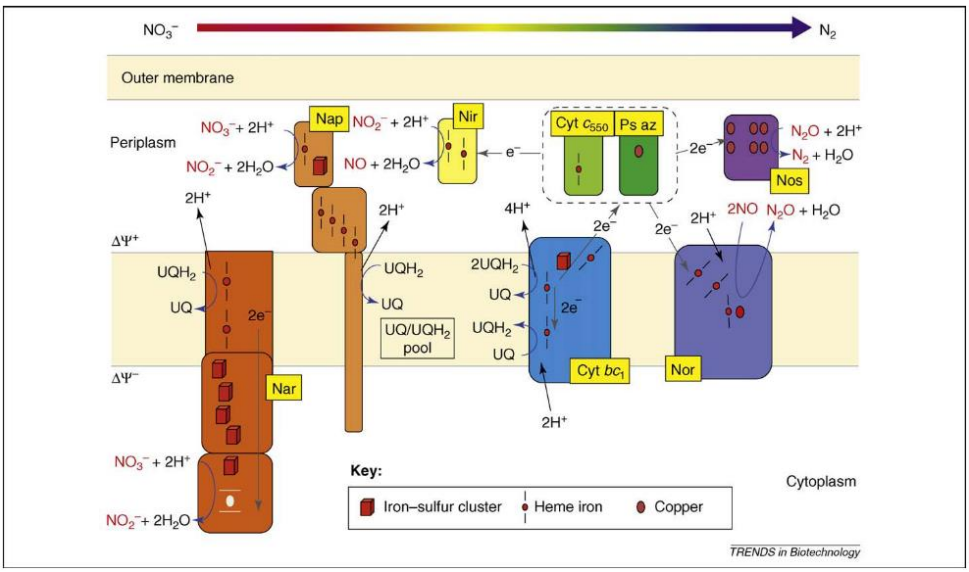


Figura 1. Representación de las diferentes enzimas que participan en el proceso de desnitrificación en muchas bacterias Gram negativas, como Pa. denitrificans. Abreviaturas: Nar, nitrato reductasa unida a membrana; Nap, nitrato reductasa periplásmica; Nir, nitrito reductasa; Nor, óxido nítrico reductasa; Nos, sistema de óxido nitroso reductasa; Ps az, pseudoazurina; Cyt bc1, citocromo bc1; Cyt c550, citocromo c550. A medida que el diagrama fluye de izquierda a derecha, el aceptor de electrones se reduce progresivamente con NO3 reduciéndose a gas N2. [2]

Algunos de los impulsores de la vía de la desnitrificación son la disponibilidad de oxígeno, la presencia de compuestos nitrogenados como nitrato, nitrito u óxido nítrico, y finalmente, donantes de electrones adecuados, como compuestos orgánicos del carbono[3].

El óxido nítrico(NO) es una molécula implicada en señalización celular y en la defensa del huésped, ya que debido a su toxicidad se han desarrollado mecanismos que lo transforman rápidamente a óxidos de nitrógeno[4]. Además, el NO es un precursor del óxido nitroso (N_2O), un importante gas de efecto invernadero. Esta característica hace esencial el estudio de diferentes mecanismos genéticos y moleculares con la finalidad de mitigar diferentes problemáticas medio ambientales actuales. Se han conocido factores de trascripción (NorR, NnrR, NsrR y DNR) implicados en la desnitrificación, en respuesta al NO[3].

Aunque el dióxido de carbono(CO_2) es el gas efecto invernadero más conocido, el N_2O resulta tener una potencia más de 300 veces mayor, y, además, también contribuye a la disminución del ozono estratosférico[5]. Se piensa que aproximadamente el 62% de la emisiones de N_2O a nivel mundial, provienen de suelos naturales y agrícolas debido a la desnitrificación bacteriana y a las oxidación del amoniaco. La intensificación de la agricultura conllevará en el futuro un mayor impacto medio ambiental si no se encuentran formas de reducir las emisiones biológicas de N_2O[1]. Se necesitan modelos para estimar las emisiones de óxido nitroso a diferentes escalas espacio-temporales para evaluar las opciones de mitigación[6].

El paso de NO a N_2O representa una reacción inusual, ya que se da la formación del enlace NN[7]. Dicha reacción es catalizada por la enzima óxido nítrico reductasa (Nor) durante el proceso de desnitrificación:

$$2NO + 2e^- + 2H^+ \rightarrow 2N_2O + H_2O \text{ [8]}$$

La reacción tiene lugar en la cara externa de la membrana citoplasmática, tal y como se puede observar en la Figura 1, y la mayoría de ellas se han identificado en Proteobacterias[3]. Nor se encuentra clasificada dentro de la familia de las hemo-cobre oxidasas (HCO)[9] y representa un tipo de respiración anaerobia [8]. En esta familia de proteínas se incluyen las citocromo c oxidasas, y la diferencia principal entre estas y Nor es que la última presenta hierro en lugar de cobre en su centro binuclear[4]. La forma más estudiada de Nor es cNOR o NorBC, que se encuentra principalmente en bacterias desnitrificantes y consiste en un dímero unido a la membrana que oxida el citocromo c [10]. Las subunidades de cNOR o NorBC se encuentran codificadas por los genes norB y norC [11]. Se confirmó la presencia de 12 hélices alfa transmembrana en la subunidad NorB, mientras que NorC se encontraba unido a la membrana mediante un único segmento transmembranal [12]. Por otro lado, aparece otra forma en bacterias patógenas no nitrificantes, bacterias desnitrificantes y arqueas, que acepta electrones del quinol, denominada qNor, presentado en este caso una única

subunidad[8]. Además, las enzimas qNor presentan una extensión N-terminal que está ausente en NorB del complejo cNor. Finalmente, se describe un grupo inusual de qNor denominado CuANor[3].

Tal y como se ha comentado anteriormente, las subunidades del complejo NorBC se encuentran codificadas por los genes norB y norC. Dichos genes habitualmente se cotranscriben con genes como norD, norE, norF y norQ. Se dabe que norQ y norD siempre están vinculados a norBC, sin embargo, los genes norE y norF pueden estas distantes o ausentes en algunos genomas[13]. En la figura 2 se pueden observar distintos patrones de genes Nor en bacterias desnitrificantes.

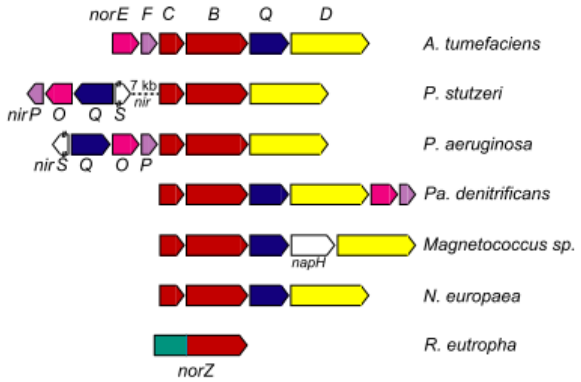


Figura 2. Patrones de genes Nor de bacterias desnitrificantes. Los genes homólogos se muestran en colores idénticos. Los marcos de lectura codificantes se escalan al tamaño relativo; los recuadros de flechas indican direcciones de transcripción[13].

Por otro lado, se ha visto que los hongos también presentan un sistema de desnitrificación que consiste en la reducción del nitrito a N_2O durante la respiración anaerobia mediante la actividad conjunta de la enzima nitrito reductasa (NirK) con cobre y una reductasa de óxido nítrico del citocromo p450 (P450nor) [1]. La reacción catalizada por P450nor es:

$$2NO + NAD(P)H + H^+ \rightarrow N_2O + H_2O + NAD(P) + P\text{-}450nor \text{ [11]}$$

El producto final de la desnitrificación fúngica es el óxido nitroso (N_2O), ya que estos organismos no presentan la enzima óxido nitroso reductasa (Nos). Además, P450nor no presenta anclaje a la membrana (presente en todas las enzimas eucariotas del citocromo p450) y no tiene actividad de mono-oxigenación[11].

La enzima que se encarga del paso de N_2O a N_2 es la óxido nitroso reductasa (Nos), que es un sumidero de eliminación del óxido nitroso[5]. No obstante, resulta importante estudiar el papel de Nor, ya que se ha visto una ausencia significativa de la regulación por N_2O[1]. Es decir, muchos microorganismos

desnitrificantes no responden con la producción de Nos frente al incremento de N_2O, de forma que el producto final es N_2O en lugar de N_2. De hecho, se dice que la desnitrificación es un proceso modular, ya que un organismo no siempre presenta el conjunto completo de enzimas desnitrificantes cuyo producto final es el N_2[5]. Esto resulta un importante problema a nivel medio ambiental porque conlleva un aumento de los niveles de N_2O en el planeta. Por este motivo es esencial modular el proceso de producción de producción de N_2O por Nor. Además, se piensa que la regulación de la emisión de N_2O tiene una base genética, ya que se estableció una relación causal entre la composición de la población desnitrificante y las posibles emisiones de N_2O[5].

3 Metodología

3.1 Construcción de la base de datos propia

Como punto de partida del trabajo, se usó la herramienta PSI-BLAST (Position-Specific Iterated Blast)[14] para la creación de una base de datos propia de secuencias de proteínas, ya que dicho algoritmo realiza una búsqueda tipo BLAST, pero mucho más sensible. Se decidió usar esta herramienta porque permite buscar proteínas homólogas lejanas a partir de una secuencia de consulta, de esta manera se puede obtener una base datos de secuencias homólogas repurada.

PSI-BLAST comienza buscando en una base datos a partir de una secuencia molde y después, construye un alineamiento múltiple de las secuencias obtenidas. Este primer paso de búsqueda es idéntico al seguido por BLASTp. A partir del alineamiento múltiple se obtiene una matriz de puntuación PSSM (position-specific scoring matrix), que se usa para buscar, nuevamente, secuencias homólogas en la base datos. El proceso es iterativo hasta que se decide finalizar la búsqueda o hasta convergencia.

En el caso concreto de este trabajo, se usó como secuencia de consulta la secuencia proteica de la óxido nítrico reductasa de *Paracoccus denitrificans* (Acc. Number: GEK68812.1), ya que es la bacteria desnitrificante mejor caracterizada[15].

La búsqueda se realizó en la base datos “non- redundant protein sequences(nr)” en organismos bacterianos. Además, se seleccionó un máximo de 1000 secuencias y un umbral restrictivo de 0,001. Tras ocho iteraciones, se obtuvieron un total de 230 secuencias anotadas como “nitric oxide reductase”, que fueron las seleccionadas para la construcción de la base de datos.

3.2 Comprobación de la validez de la base de datos

Con la finalidad de comprobar si la base de datos presentaba secuencias relevantes para el estudio, se decidió realizar una búsqueda similar en PSI-BLAST usando como *query* la secuencia de la óxido nítrico reductasa de *Bradyrhizobium japonicum* (Acc. Number: KMJ97954.1), ya que se describe en la bibliografía como un organismo Rizobios desnitrificante asociado a leguminosas[3]. Se mantuvieron los mismos parámetros de búsqueda que para *Pa. denitrificans* para que el resultado de la comparación fuera lo más preciso posible.

3.3 Alineamiento múltiple de las secuencias

Una vez confirmada la validez de la base datos propia se procedió al alineamiento múltiple de las secuencias proteicas. Para ello se usó el programa MEGA X[16], mediante el software MUSCLE[17]. Se decidió usar MEGA X porque se trata de un programa de uso cotidiano por muchos expertos en el ámbito científico, dado que permite un gran variedad de análisis, está disponible para diferentes sistemas operativos y es eficiente a nivel computacional. Por otro lado, se escogió MUSCLE para el análisis, ya que resulta ser eficiente computacionalmente, rápido y preciso.

El input para realizar el alineamiento fueron las 230 secuencias de la base de datos obtenida a partir de PSI-BLAST. Además, se tuvo que realizar una edición manual de algunas secuencias, dado que al cargarlas en MEGA X se incorporaron espacios al final de la secuencia, lo que llevaba a la obtención de un incorrecto alineamiento.

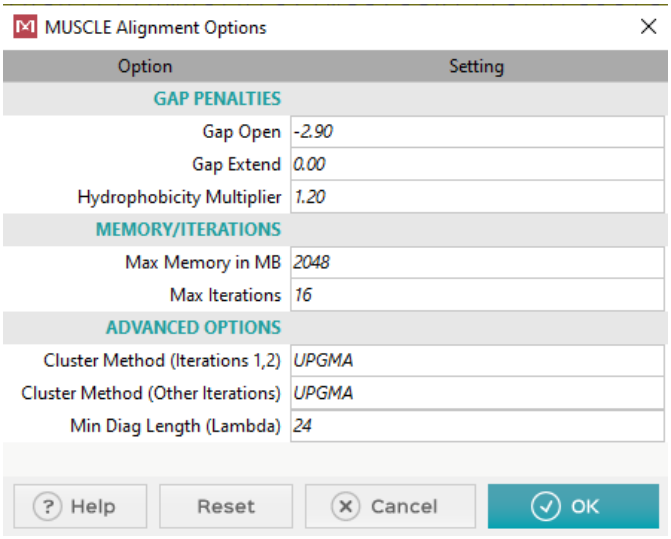


Figura 3. Parámetros seleccionados para el MSA

3.4 Determinación del modelo evolutivo

El siguiente paso del estudio consistió en determinar a qué modelo evolutivo se ajustaban mejor los datos obtenidos en el alineamiento, con la finalidad de obtener una estimación filogenética mucho más precisa. Dicha determinación se realizó con diferentes softwares, para comparar los resultados. En primer lugar, se usó MEGA X, ya que ofrece esta función y se puede realizar de forma sencilla debido a su interfaz gráfica. Además, se determinó el modelo evolutivo con ModelTest-NG[18], dado que aparece descrito en la bibliografía como una implementación de dos programas ampliamente conocidos y usados (jModelTest y ProtTest), siendo más eficiente y con nuevas características. Finalmente, se usó ModelFinder[19], el servidor web de IQ-TREE[20], descrito como un método rápido de selección de modelos.

3.5 Estudio de la señal filogenética

Para comprobar que los organismos de la base datos tenían relación y el alineamiento de sus secuencias no se había producido por azar, se realizó el estudio de la señal filogenética. Para llevar a cabo dicho estudio se usó el software TREE-PUZZLE[21]. Se eligió este programa dado el fácil manejo que presenta y la variedad de estudios que incluye. En concreto se usó la herramienta “Likelihood mapping”, que resulta ser una forma sencilla de visualizar el contenido filogenético de un alineamiento de secuencias [22]. Cabe destacar que fue necesario realizar una conversión del alineamiento múltiple en formato FASTA a formato Phylip para que este pudiera ser usado como input de TREE-PUZZLE. Para realizar dicha conversión se usó la herramienta *EMBOSS seqret* del EMBL-EBI.

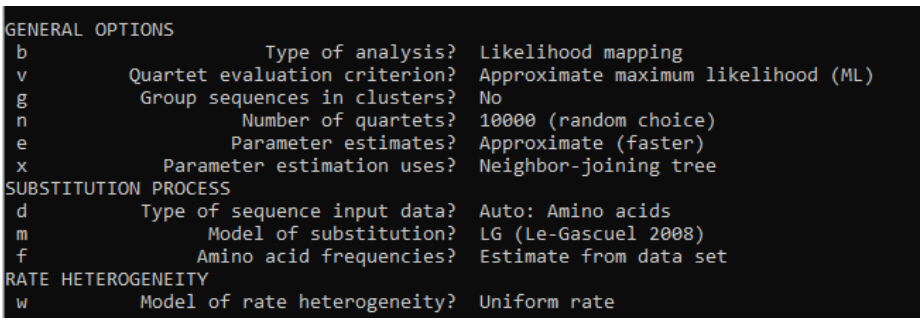


Figura 4. Opciones seleccionadas para el estudio de la seña filogenética con TREE-PUZZLE en el terminal

3.6 Construcción e inferencia del árbol filogenético

Para la construcción del árbol filogenético a partir del alineamiento múltiple se usó el programa MEGA X, dado el sencillo e intuitivo manejo de este. Por otro lado, se decidió emplear el método probabilístico de máxima verosimilitud para la construcción del árbol, ya que es el método más frecuentemente usado en filogenia molecular[23]. Además, dado que en este análisis se usaron secuencias homólogas distantes, resulto de interés el uso de este método porque requiere que la que evolución de diferentes sitios y linajes sea estadísticamente independiente.

Para la inferencia de la historia evolutiva mediante el método de máxima verosimilitud se indicó que el modelo de sustitución a aplicar debía ser el modelo LG[24], ya que resultó ser el modelo evolutivo al que mejor se ajustaban los datos en análisis previos. Por otro lado, los árboles iniciales para la búsqueda heurística se obtuvieron automáticamente aplicando los algoritmos Neighbour-Join y BioNJ a una matriz de distancias. Además, se usó una distribución gamma discreta.

Finalmente, para la evaluación de la filogenia se aplicó el método de *Bootstrap*[25]. En concreto, se indicó un total de 1000 replicaciones de *Bootstrap*, ya que, a mayor número de réplicas, mejor es la estimación del intervalo de confianza.

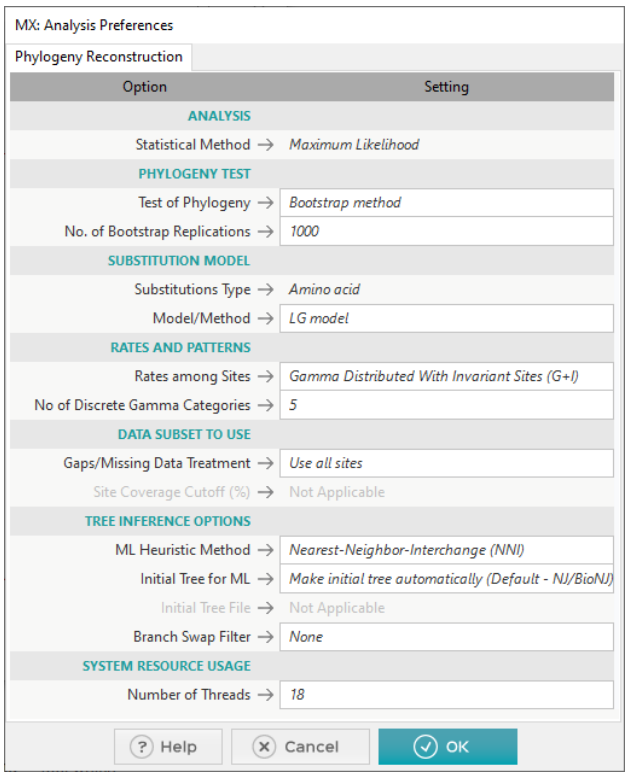


Figura 5. Opciones seleccionadas para la construcción de un árbol filogenético empleando el método probabilístico de máxima verosimilitud

Para estudiar qué organismos estaban presentes en cada uno de los clados del árbol, se empleó la herramienta en línea *Taxonomy browser* del NCBI, en concreto *Taxonomy name/id Status Report Page*, donde se introduce una lista de organismos y se obtiene un documento que contiene los IDs taxonómicos de cada uno de ellos. Este último documento se importó en R y, mediante el uso de paquetes como *myTAI* y *plyr* se creó una función para obtener la taxonomía completa de cada uno de los organismos de los diferentes clados(Anexo III).

3.7 Creación de perfiles proteicos y búsqueda de coincidencias en bases de datos

Como paso previo a la obtención de los perfiles, se realizaron los alineamientos múltiples de las secuencias de cada uno de los dos grupos identificados en el árbol. Para realizar dichos alineamientos, fue necesario obtener las secuencias de cada clado en formato FASTA. Para ello, se creó un script de R(Anexo VI) donde a partir de los *labels* de cada clado recuperados de MEGA X, se usó el fichero multi-FASTA de las 230 secuencias tras la búsqueda en PSI BLAST para encontrar las coincidencias y obtener dos ficheros FASTA distintos con las secuencias aminoacídicas de cada clado. A continuación, se alinearon las secuencias con MEGA X mediante MUSCLE. Además, se obtuvieron las secuencias consenso de cada clado del árbol mediante la herramienta EMBOSS Cons[26], usando como input los alineamientos múltiples de cada uno de los dos clados.

Se usó HMMER para la construcción de los perfiles HMM (Hidden Markov Models) [27]. En primer lugar, se empleó *hmmbuild,* usando los alineamientos múltiples de las secuencias de cada clado como input, para obtener los perfiles HMM. Una vez obtenidos los dos perfiles, se procedió a usar *hmmsearch* para buscar coincidencias entre ambos perfiles y la base de datos Uniref100(descargada en mayo de 2021). Cabe destacar que no fue necesario la calibración del perfil con *hmmcalibrate* porque dicha función de calibración ya se encuentra incorporado en *hmmbuild* de la última versión de HMMER.

Figura 6. Workflow llevado a cabo

4 Resultados y discusión

Tras las búsqueda en la bibliografía y la elección de la secuencia de la enzima Nor de *Paracoccus denitrificans* como molde para la búsqueda en PSI-BLAST, se obtuvo una base de datos refinada de 230 secuencias homólogas anotadas como *nitric oxide reductase*, las cuales aparecen en el Anexo II de la memoria.

Dado que era de vital importancia la validación de la base de datos construida para continuar con el análisis, se comparó dicha base de datos con el resultado de la búsqueda en PSI-BLAST usando como *query* una secuencia de Nor de un organismo distinto (*Bradyrhizobium japonicum*). En este segundo caso y tras siete iteraciones, se obtuvo un total de 247 secuencias anotadas como *nitric oxide reductase*. Para comparar los archivos multi-FASTA obtenidos en las ambas búsquedas se creó un pequeño programa en R (Anexo I) que comparaba ambos ficheros. Se obtuvo que 94,78% de las secuencias de la búsqueda con *Paracoccus denitrificans* también se encontraron para *Bradyrhizobium japonicum*. Esto indicó que la base de datos era representativa y se podía continuar con el estudio.

Una vez validada la base de datos propia, se continuó con el alineamiento múltiple de las secuencias que la componían. El alineamiento mostró regiones centrales conservadas entre las secuencias aminoacídicas de los distintos organismos homólogos, así como los extremos N y C-terminales muy variables, tal y como se puede observar en la Figura 7, donde se muestra el resultado del alineamiento múltiple de las diez primeras secuencias.

El siguiente paso consistió en determinar el modelo evolutivo que mejor explicaba los datos del alineamiento obtenidos. Tal y como aparece en el apartado 3.4 de la metodología, el modelo evolutivo se determinó mediante tres programas diferentes. En la tabla 2 aparecen los siete primeros modelos con menor valor BIC obtenidos a partir de MEGA X. El modelo que mejor explicaba los datos resultó ser LG+I+G4. Con los otros dos programas nombrados se obtuvo el mismo resultado.

Model	Parameters	BIC	AICc	lnL	(+I)	(+G)
LG+G+I	459	50762,050	46383,755	-22730,829	0,09	0,41
LG+G	458	50763,661	46394,896	-22737,409	n/a	0,37
LG+G+F	477	50862,337	46312,511	-22677,043	n/a	0,35
LG+G+I+F	478	50868,135	46308,780	-22674,168	0,07	0,38
WAG+G+I+F	478	51109,365	46550,011	-22794,784	0,10	0,44
WAG+G+F	477	51114,588	46564,763	-22803,169	n/a	0,39
JTT+G+F	477	51176,652	46626,826	-22834,201	n/a	0,38

Tabla 2. Resultado de las primeras columnas y los primeros siete modelos evolutivos con menor BIC a partir de MEGA X

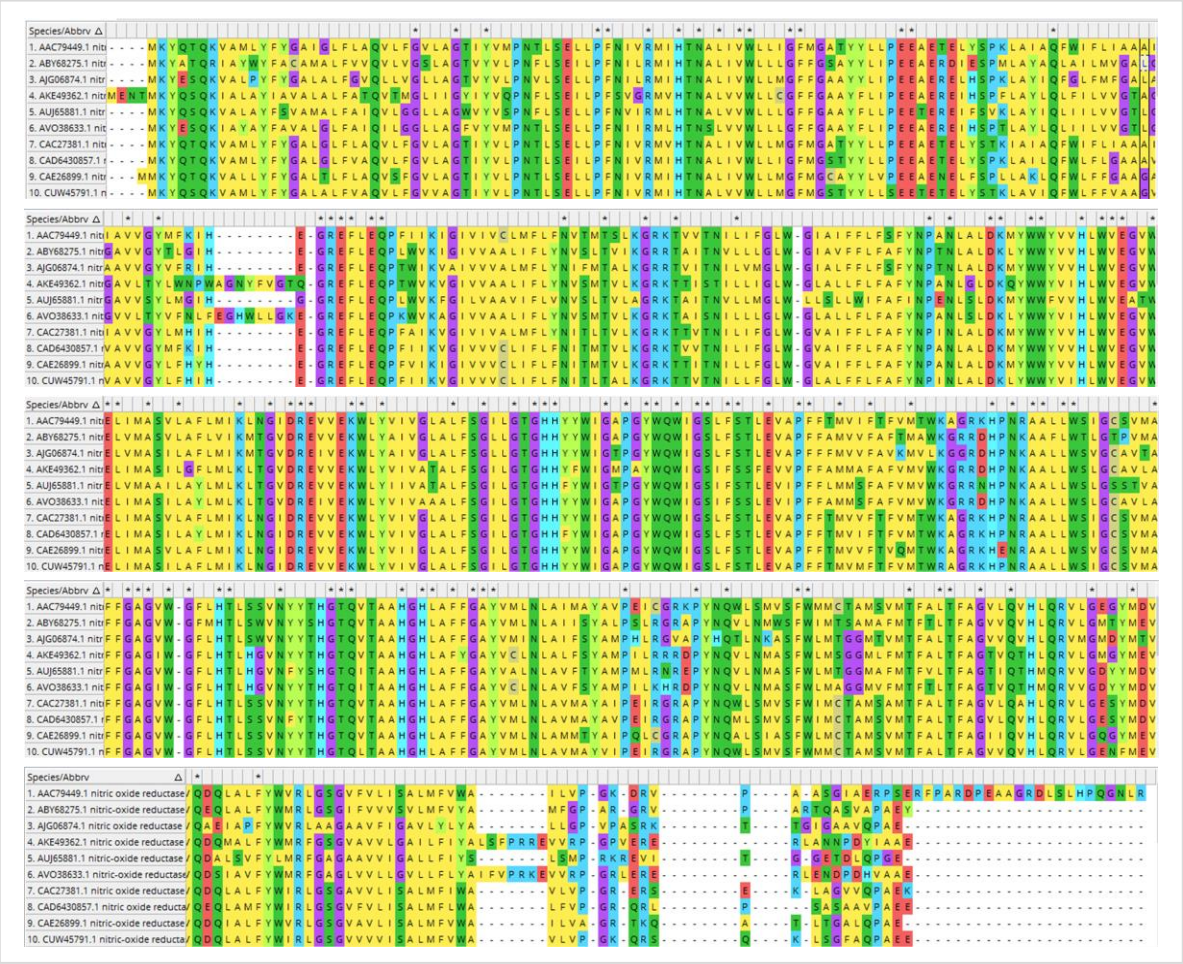


Figura 7. Alineamiento múltiple de las diez primeras secuencias obtenidas tres la búsqueda en PSI-BLAST. Dicho alineamiento se realizó en MEGA X con MUSCLE.

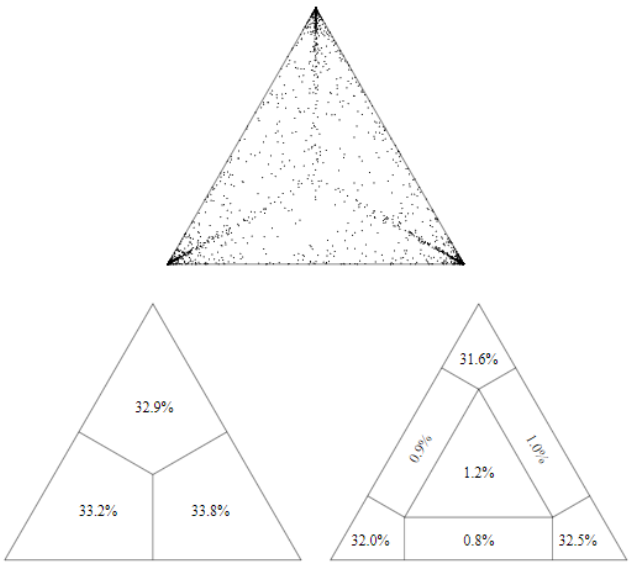


Figura 8. Resultado tras realizar el "Likelihood mapping" con TREE-PUZZLE para estudiar la señal filogenética

Por otro lado, resultaba interesante estudiar la señal filogenética y se obtuvo que las secuencias de la base datos presentaban una intensa señal filogenética. En la figura 8 aparece el resultado, y debido a la distribución de los puntos, se pudo afirmar los datos son adecuados para la reconstrucción filogenética.

Una vez obtenido y revisado el alineamiento múltiple de las 230 secuencias recuperadas de la búsqueda en PSI-BLAST, se procedió a la construcción del árbol filogenético (Figura 10), cuyos parámetros se puede revisar en el apartado 3.5 de la metodología. El árbol mostró dos ramas claramente diferenciadas, a las que se las llamó Grupo A(o clado 1) y Grupo B(o clado 2). El grupo A estaba constituido por un total de 162 secuencias correspondientes a la subunidad grande de la óxido nítrico reductasa. Entre los géneros bacterianos de este primer grupo, predominaron *Rhizobium*, *Ensifer*, *Sinorhizobium*, *Brucella* y *Ochrobactrum*. En el caso del grupo B, éste estaba conformado por 64 secuencias de una gran variedad de géneros bacterianos, siendo muy predominante la familia taxonómica *Rhodobacteraceae.*

Para una mejor interpretación de los resultados del árbol, se obtuvieron los identificadores taxonómicos de cada organismo y se determinó la clasificación taxonómica completa (Anexo III). Dicha clasificación completa se encuentra en las tablas de los anexos IV y V, que corresponden a los organismos del clado A y B, respectivamente.

Dado que no se disponía de la clasificación completa de todos los organismos del clado B, solo se pudo estudiar el conjunto completo a partir del taxón Clase. Se vio que prácticamente todas las bacterias del clado B pertenecían a la clase *Alphaproteobacteria.* Por otro lado, se pudo obtener que, al menos, el 70% de los organismos pertenecían al orden *Rhodobacterales.*

En el caso del clado A, se disponía de una clasificación taxonómica más completa y se pudo estudiar el conjunto completo de organismos en función del taxón Familia. En este caso, más del 99% de las bacterias pertenecían a la clase *Alphaproteobacteria* y la distribución de las distintas familias bacterianas presentes en el clado A aparecen representadas en la figura 9.

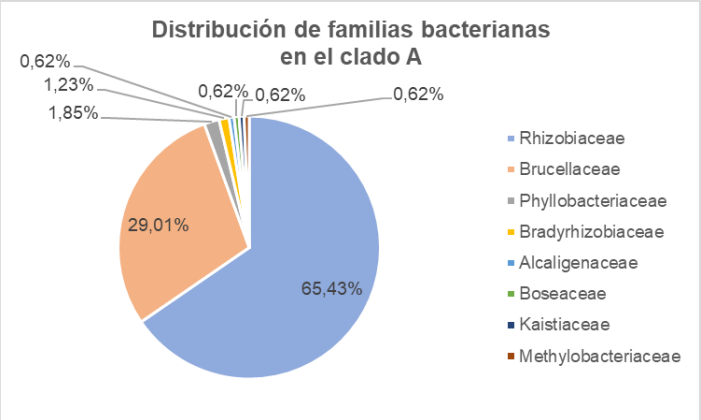


Figura 9. Proporción de familias bacterianas presentes en el clado A del árbol

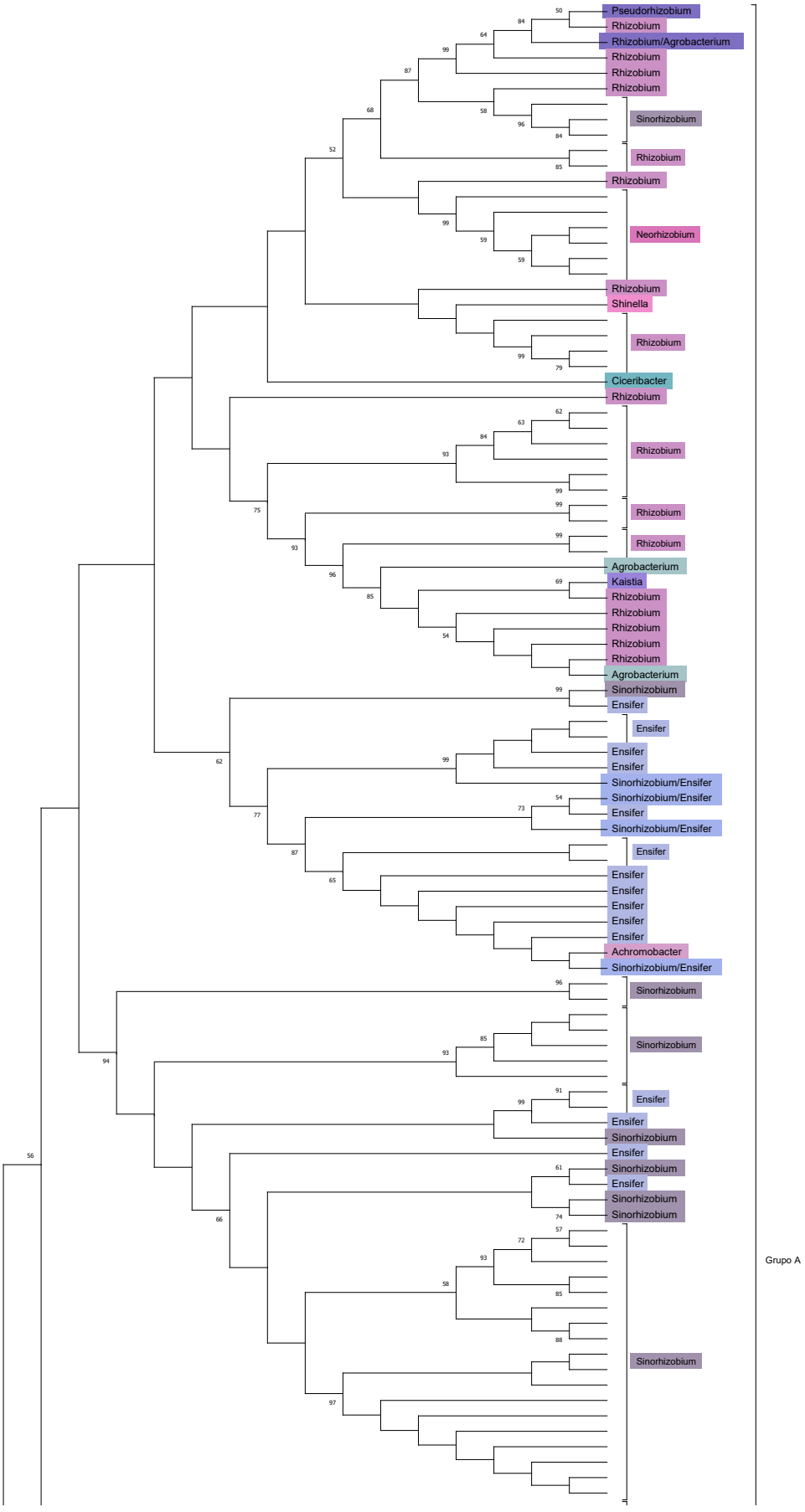


Figura 10. Árbol filogenético. (Continuación)

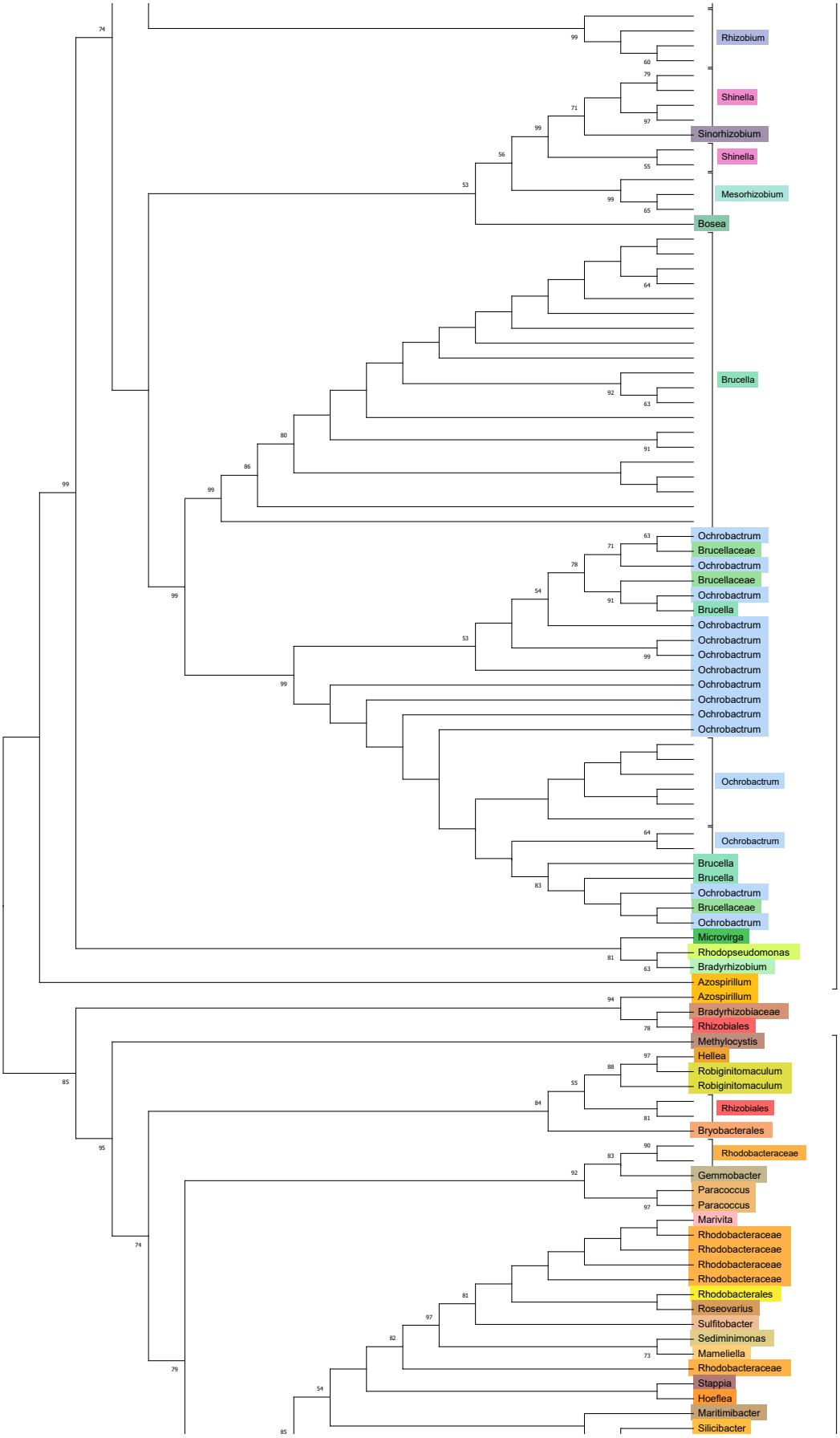


Figura 10. Árbol filogenético. (Continuación)

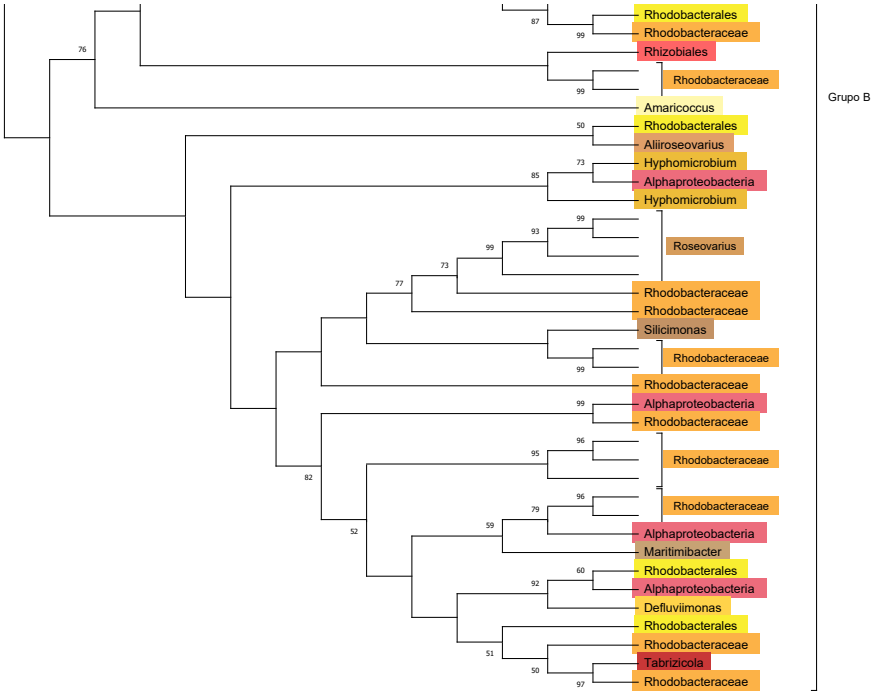


Figura 10. Árbol filogenético construido mediante el método de máxima verosimilitud. Se han indicado los valores de Bootstrap mayores al 50% y se ha coloreado en función del Genero para facilitar la interpretación.

Tas estudiar el resultado de la filogenia, se pasó al segundo gran apartado del trabajo, el de la construcción de perfiles HMM de la proteína Nor a partir de los alineamientos múltiples de las secuencias de cada clado.

El alineamiento múltiple de las 162 secuencias del grupo A presentó un total de 476 columnas alineadas. Dicho alineamiento se puede consultar en la Figura 11, donde se muestran las primeras cinco secuencias del clado A alineadas.

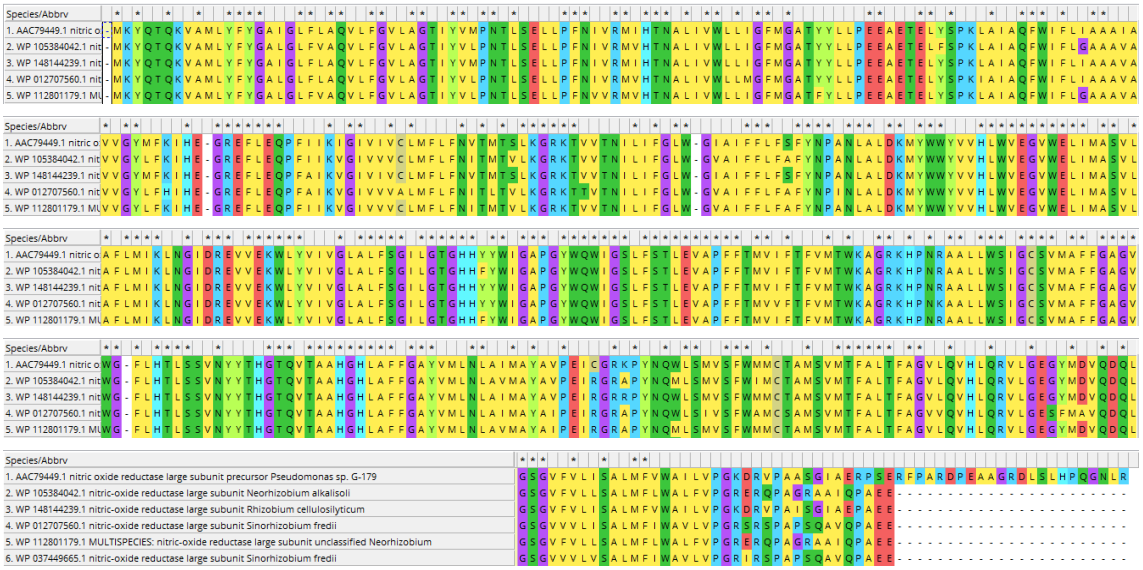


Figura 11. Alineamiento múltiple del primeras cinco secuencias del grupo A

En el caso del clado B, tras el alineamiento resultaron un total de 472 columnas alineadas de las 64 secuencias que lo componían. El alineamiento múltiple de las primeras cinco secuencias se muestra en la figura 12.

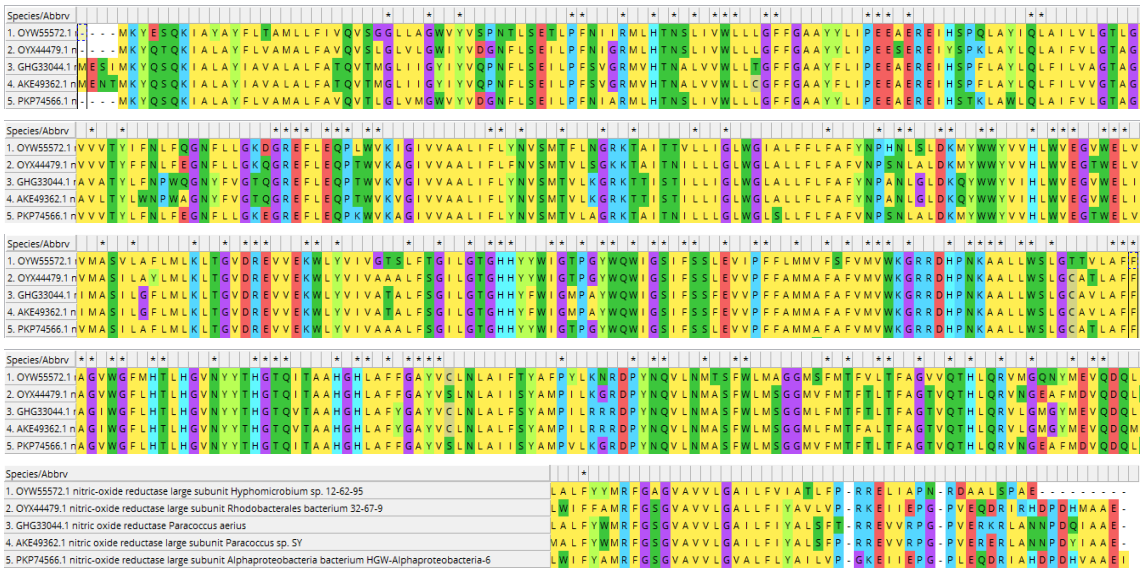


Figura 12. Alineamiento múltiple del primeras cinco secuencias del grupo B

El resultado que se obtuvo tras alineamiento de las secuencias en función del clado mostró fue que aparecían muchas más zonas coincidentes. Por otro lado, las secuencias consenso calculadas mediante EMBOSS Cons fueron las siguientes:

```
>Seq_consenso_clado_1
xMKYQTQKVAMLYFYGALGLFLAQVLFGVLAGTIYVLPNTLSELLPFNIVRMIHTNALIV
WLLMGFMGATYYLLPEEAETELYSPKLAIAQFWIFLVAAAVAVVGYLFHIHExGREFLEQ
PFxIKVGIVVVxLMFLFNITLTVLKGRKTTVTNILIFGLWxGVAIFFLFAFYNPINLALD
KMYWWYVVHLWVEGVWELIMASVLAFLMIKLNGIDREVVEKWLYVIVGLALFSGILGTGH
HYYWIGAPGYWQWIGSLFSTLEVAPFFTMVIFTFVMTWKAGRKHPNRAALLWSIGCSVMA
FFGAGVWGxFLHTLSSVNYYTHGTQVTAAHGHLAFFGAYVMLNLAVMAYAIPEIRGRAPY
NQWLSMVSFWMMCTAMSVMTFALTFAGVVQVHLQRVLGESFMDVQDQLALFYWIRLGSGV
VVLISALMFIWAVLVPGRERSxSLSxxVQPAEExxxxxxxxxxxxxxxxxxxxxxxx

>Seq_consenso_clado_2
xxxxMKYQSQKIAxAYFLVAMALFAIQVLGGLLAGWIYVSPNFLSEILPFNIIRMLHTNS
LIVWLLLGFFGAAYYLIPEESEREIHSpKLAYLQLIILVVGTLGVVVTYLFNLFEGNWLL
GxEGREFLEQPLWVKxGIVVAALIFLYNISMTVLAGRKTAITNILLLGLWGLALLFLFAF
YNPDNLSLDKMYWWYVVHLWVEGTWELVMASILAFLMLKLTGVDREVVEKWLYVIVATAL
FSGILGTGHHYYWIGTPGYWQWIGSIFSSLEVIPFFAMMSFAFVMVWKGRRDHPNKAALL
WSLGCATLAFFGAGVWGFLHTLHGVNYYTHGTQITAAHGHLAFYGAYVxLNLAIFTYAMP
ILRGRDPYNQVLNMASFWLMSGGMxFMTFVLTFAGTIQTHLQRVMGEYYMDVQDQLAIFY
xMRFGAGVAVVLGALLFIYAVLVPxRREVISPGxxxxxxxxxxxxxxxxxxx
```

Aminoácido	Código de una letra
Alanina	A
Arginina	R
Asparagina	N
Ácido aspártico	D
Cisteína	C
Glutamina	Q
Ácido glutámico	E
Glicina	G
Histidina	H
Isoleucina	I
Leucina	L
Lisina	K
Metionina	M
Fenilalanina	F
Prolina	P
Serina	S
Treonina	T
Triptófano	W
Tirosina	Y
Valina	V

Figura 13. Secuencias consenso de los clados A(1) y B(2)

A continuación, se procedió a la construcción de los perfiles HMM mediante *hmmbuild* . El perfil generado a partir del alineamiento de las secuencias del

clado A presentó 449 posiciones de consenso, definiendo así 27 columnas del alineamiento con espacios. El número de secuencia efectivo resultó ser 0,40.

```
uoc@progbio:~/PERFIL_TFM$ hmmbuild perfil_clado_1.hmm alineamiento_clado_1.fas
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.1b2 (February 2015); http://hmmer.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# input alignment file:             alineamiento_clado_1.fas
# output HMM file:                  perfil_clado_1.hmm
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

# idx name                 nseq  alen  mlen eff_nseq re/pos description
#---- -------------------- ----- ----- ----- -------- ------ -----------
1     alineamiento_clado_1   162   476   449     0.40  0.585

# CPU time: 0.82u 0.00s 00:00:00.82 Elapsed: 00:00:00.83
```

Figura 14. Perfil a partir del MSA de las secuencias del clado A

En el caso del clado B, el perfil construido a partir de *hmmbuild* presentó 455 posiciones de consenso. En este caso el número de secuencia efectivo fue 0,46.

```
uoc@progbio:~/PERFIL_TFM$ hmmbuild perfil_clado_2.hmm alineamiento_clado_2.fas
# hmmbuild :: profile HMM construction from multiple sequence alignments
# HMMER 3.1b2 (February 2015); http://hmmer.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
# input alignment file:             alineamiento_clado_2.fas
# output HMM file:                  perfil_clado_2.hmm
# - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

# idx name                 nseq  alen  mlen eff_nseq re/pos description
#---- -------------------- ----- ----- ----- -------- ------ -----------
1     alineamiento_clado_2    64   472   455     0.46  0.592

# CPU time: 0.83u 0.00s 00:00:00.83 Elapsed: 00:00:00.84
```

Figura 15. Perfil a partir del MSA de las secuencias del clado B

Finalmente, se procedió a estudiar si los perfiles construidos eran capaces de reconocer secuencias proteicas depositadas en la base de datos Uniref100. Para ello, se descargó la base de datos y se usó *hmmsearch* . Cada uno de los outputs se editaron seleccionando todos los hits hasta el umbral (*threshold*). Seguidamente, se guardaron en un fichero de texto plano para el análisis en R.

Se obtuvo que el perfil del grupo A mostraba un total de 27747 hits con un e-valor menor a 0,001, mientras que con el perfil del grupo B fueron 17588, también con un e-valor menor a 0,001. La gran parte de las proteínas reconocidas por los perfiles estaban denotadas como óxido nítrico reductasas. Además, los perfiles también identificaron proteínas anotadas como *unclassified* en la base de datos.

5 Conclusiones

5.1 Conclusiones

En conclusión, se obtuvo una base datos depurada de enzimas óxido nítrico reductasas (Nor) a partir de las cuales se construyó un árbol filogenético, que las agrupó en dos clados. Para cada uno de los clados se construyó un perfil HMM distinto. Se demostró que dichos perfiles eran útiles para la identificación de proteínas anotadas como Nor en la base datos de UniRef100. Finalmente, si se continua con el estudio, estos perfiles podrían ser usados para explorar el papel que juegan las óxido nítrico reductasas en el proceso de desnitrificación llevado a cabo por bacterias y el impacto de este sobre el medio ambiente. De esta forma, se podrían establecer diferentes vías de acción para reducir las emisiones de óxido nitroso.

5.2 Líneas de futuro

Una línea futura por explorar sería la de estudiar el resultado obtenido tras enfrentar los perfiles generados a genomas completos. Por otro lado, como paso final para completar este estudió sería interesante utilizar los perfiles para sintetizar la proteína y clonarla para comprobar si *in vitro* lleva a cabo la función conocida de la enzima Nor.

- Consideraciones finales

Como conclusión principal del trabajo obtengo que se debe seguir trabajando para que el gasto computacional que requieren muchos de los análisis llevados a cabo sean más rápidos y eficientes, ya que en muchos casos ha supuesto una limitación en la realización del trabajo.

Por otro lado, este trabajo me ha permitido mejorar mi capacidad autodidacta, dado que, debido a mi inexperiencia en la construcción de filogenias y perfiles proteicos, he tenido que invertir gran parte de mi tiempo en la búsqueda de información para ser capaz de avanzar con el análisis, cumplir con los objetivos a tiempo e interpretar los resultados obtenidos.

Tal y como se ha comentado anteriormente y en los informes de seguimiento, el factor del gasto computacional ha resultado ser un límite. Es por ello, por lo que no se ha logrado alcanzar la última tarea planteada en la planificación inicial. Aunque comencé con el análisis de comparación de los perfiles generados frente a bibliotecas de genomas completos como MG-RAST, decidí no incluirlo en el estudio por falta de tiempo para ejecutar e interpretar los resultados.

La realización del TFM me ha ayudado a entender que como bioinformático se debe tener la iniciativa de buscar herramientas bioinformáticas que se adapten a lo que quieres conseguir. Siento que este trabajo ha sido un punto de partida en mi carrera como bioinformática y estoy segura de que cuando adquiera más experiencia en el sector seré capaz de perfeccionar cada uno de los procesos llevados a cabo. Espero poder realizar más análisis filogenéticos en el futuro, ya que la experiencia ha sido personalmente satisfactoria.

5.3 Seguimiento de la planificación

En general, se ha conseguido seguir la planificación prevista, aunque es cierto que he tenido que invertir mucho más tiempo en la realización del que esperaba, teniendo en cuenta que estaba cursando al mismo tiempo otras tres asignaturas. En muchas ocasiones he priorizado la realización del TFM frente a entregas voluntarias de otras asignaturas para poder llevar un buen seguimiento de la planificación.

En cuanto a la metodología, he de decir que me ha supuesto un desafío personal, ya que al no haber realizado nunca un trabajo igual, me sentía desorientada en cuanto a qué programas estaban a mi disposición. De hecho, la metodología ha sido el punto de más variación del trabajo, ya que se han tenido que introducir cambios para obtener el trabajo final de forma exitosa. La principal razón de esos cambios ha sido el tiempo ajustado(unos meses) y mi poca experiencia en cuanto al uso de ciertos softwares. Sin embargo, el proceso de investigación sobre posibles programas para intentar aplicarlos a mi trabajo me ha supuesto un gran enriquecimiento como futura bioinformática.

6 Glosario

- **Nor:** *Nitric oxide reductase* u óxido nítrico reductasa. Enzima que cataliza la reducción de óxido nítrico a óxido nitroso.

- **PSI-BLAST:** Position-Specific Iterative (PSI)-BLAST. Método de búsqueda de perfiles de secuencias de proteínas que se basa en las alineaciones generadas por una ejecución del programa BLASTp

- **Secuencia consenso:** Secuencia ideal que representan los nucleótidos o aminoácidos que se encuentran con mayor frecuencia en cada posición de un fragmento de DNA o de una proteína, respectivamente.

- **Clado:** Cada una de las ramas del árbol filogenético propuesto para agrupar a los seres vivos. Se interpreta como un conjunto de especies emparentadas (con un antepasado común).

- **Perfil HMM:** Un modelo oculto de Márkov o HMM (Hidden Markov Model) es un modelo estadístico en el que se asume que el sistema a modelar es un proceso de Márkov de parámetros desconocidos.

- **Taxones:** Grupos en los que en biología se clasifican científicamente a los seres vivos, atendiendo a su semejanza y proximidad filogenética. Se estructuran en una jerarquía de inclusión, en la que un grupo abarca a otros menores y este, a su vez, subordinado a uno mayor.

- **Modelo evolutivo:** Descripciones matemáticas de la evolución de las secuencias y constituyen el engranaje que nos permite conectar los datos (alineamientos) con los métodos de reconstrucción filogenética.

- **Señal filogenética:** Tendencia de especies cercanas a expresar rasgos similares.

- **Matriz de puntuación PSSM:** Una matriz de puntuación de posiciones específicas (PSSM, position-specific scoring matrix) es una tabla que contiene información posicional de los aminoácidos o nucleótidos en un alineamiento múltiple de secuencias en el cual no hay huecos.

- **Método Bootstrap:** Método de remuestreo que se utiliza para aproximar la distribución en el muestreo de un estadístico. Se usa frecuentemente para aproximar el sesgo o la varianza de un análisis estadístico, así como para construir intervalos de confianza o realizar contrastes de hipótesis sobre parámetros de interés.

7 Bibliografía

[1] A. J. Thomson, G. Giannopoulos, J. Pretty, E. M. Baggs, and D. J. Richardson, “Biological sources and sinks of nitrous oxide and strategies to mitigate emissions,” *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 367, no. 1593, pp. 1157–1168, 2012.

[2] D. Richardson, H. Felgate, N. Watmough, A. Thomson, and E. Baggs, nitricMitigating release of the potent greenhouse gas N2O from the nitrogen cycle - could enzymic regulation hold the key?,” *Trends in Biotechnology*, vol. 27, no. 7. Trends Biotechnol, pp. 388–397, Jul-2009.

[3] M. J. Torres *et al.*, “Nitrous Oxide Metabolism in Nitrate-Reducing Bacteria: Physiology and Regulatory Mechanisms,” in *Advances in Microbial Physiology*, vol. 68, Academic Press, 2016, pp. 353–432.

[4] N. J. Watmough, S. J. Field, R. J. L. Hughes, and D. J. Richardson, “The bacterial respiratory nitric oxide reductase,” *Biochem. Soc. Trans.*, vol. 37,

no. 2, pp. 392–399, Apr. 2009.

[5] D. R. H. Graf, C. M. Jones, and S. Hallin, “Intergenomic comparisons highlight modularity of the denitrification pathway and underpin the importance of community structure for N2O emissions,” *PLoS One*, vol. 9, no. 12, Dec. 2014.

[6] K. Butterbach-Bahl, E. M. Baggs, M. Dannenmann, R. Kiese, and S. Zechmeister-Boltenstern, “Nitrous oxide emissions from soils: How well do we understand the processes and their controls?,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 368, no. 1621. Philos Trans R Soc Lond B Biol Sci, 05-Jul-2013.

[7] G. Braker and J. M. Tiedje, “Nitric oxide reductase (norB) genes from pure cultures and environmental samples,” *Appl. Environ. Microbiol.*, vol. 69, no. 6, pp. 3476–3483, Jun. 2003.

[8] Y. Shiro, “Structure and function of bacterial nitric oxide reductases: Nitric oxide reductase, anaerobic enzymes,” *Biochim. Biophys. Acta - Bioenerg.*, vol. 1817, no. 10, pp. 1907–1913, Oct. 2012.

[9] J. van der Oost, A. P. N. de Boer, J. W. L. de Gier, W. G. Zumft, A. H. Stouthamer, and R. J. M. van Spanning, “The heme-copper oxidase family consists of three distinct types of terminal oxidases and is related to nitric oxide reductase,” *FEMS Microbiol. Lett.*, vol. 121, no. 1, pp. 1–9, Aug. 1994.

[10] K. L. Casciotti and B. B. Ward, “Phylogenetic analysis of nitric oxide reductase gene homologues from aerobic ammonia-oxidizing bacteria,” *FEMS Microbiol. Ecol.*, vol. 52, no. 2, pp. 197–205, Apr. 2005.

[11] J. Hendriks, A. Oubrie, J. Castresana, A. Urbani, S. Gemeinhardt, and M. Saraste, “Nitric oxide reductases in bacteria,” *Biochim. Biophys. Acta - Bioenerg.*, vol. 1459, no. 2–3, pp. 266–273, Aug. 2000.

[12] T. Hino *et al.*, “Structural basis of biological N2O generation by bacterial nitric oxide reductase,” *Science (80-.).*, vol. 330, no. 6011, pp. 1666–1670, Dec. 2010.

[13] W. G. Zumft, “Nitric oxide reductases of prokaryotes with emphasis on the respiratory, heme-copper oxidase type,” *J. Inorg. Biochem.*, vol. 99, no. 1, pp. 194–215, 2005.

[14] M. Bhagwat and L. Aravind, “PSI-BLAST Tutorial,” in *Comparative Genomics*, vol. 1 and 2, N. Bergman, Ed. Humana Press, 2007, pp. 177–186.

[15] S. J. Field, F. H. Thorndycroft, A. D. Matorin, D. J. Richardson, and N. J. Watmough, “The Respiratory Nitric Oxide Reductase (NorBC) from Paracoccus denitrificans,” in *Methods in Enzymology*, vol. 437, Academic Press Inc., 2008, pp. 79–101.

[16] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, “MEGA X: Molecular evolutionary genetics analysis across computing platforms,” *Mol. Biol. Evol.*, vol. 35, no. 6, pp. 1547–1549, Jun. 2018.

[17] R. C. Edgar, “MUSCLE: Multiple sequence alignment with high accuracy and high throughput,” *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–1797, 2004.

[18] Di. Darriba, D. Posada, A. M. Kozlov, A. Stamatakis, B. Morel, and T. Flouri, “ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models,” *Mol. Biol. Evol.*, vol. 37, no. 1, pp. 291–294, Jan. 2020.

[19] S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. Von Haeseler, and L. S. Jermiin, “ModelFinder: Fast model selection for accurate phylogenetic estimates,” *Nat. Methods*, vol. 14, no. 6, pp. 587–589, May 2017.

[20] L. T. Nguyen, H. A. Schmidt, A. Von Haeseler, and B. Q. Minh, “IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies,” *Mol. Biol. Evol.*, vol. 32, no. 1, pp. 268–274, Jan. 2015.

[21] H. A. Schmidt, K. Strimmer, M. Vingron, and A. Von Haeseler, “TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing,” *Bioinformatics*, vol. 18, no. 3, pp. 502–504, 2002.

[22] K. Strimmer and A. Von Haeseler, “Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 94, no. 13, pp. 6815–6819, Jun. 1997.

[23] D. R. Brooks *et al.*, “Quantitative Phylogenetic Analysis in the 21 st Century Análisis Filogenéticos Cuantitativos en el siglo XXI,” *Rev. Mex. Biodivers.*, vol. 78, pp. 225–252, 2007.

[24] S. Q. Le and O. Gascuel, “An improved general amino acid replacement matrix,” *Mol. Biol. Evol.*, vol. 25, no. 7, pp. 1307–1320, Jul. 2008.

[25] B. Efron, E. Halloran, and S. Holmes, “Bootstrap confidence levels for phylogenetic trees,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 93, no. 23, pp. 13429–13434, Nov. 1996.

[26] F. Madeira *et al.*, “The EMBL-EBI search and sequence analysis tools APIs in 2019,” *Nucleic Acids Res.*, vol. 47, no. W1, pp. W636–W641, Jul. 2019.

[27] X. Meng and Y. Ji, “Modern Computational Techniques for the HMMER Sequence Analysis,” *ISRN Bioinforma.*, vol. 2013, pp. 1–13, Sep. 2013.

Anexos

Anexo I: Comparación de secuencias obtenidas tras la búsqueda en PSI-BLAST usando como secuencias problema las oxido nitrico reductasas de *Paracoccus denitrificans* y *Bradyrhizobium japonicum*

```
# Se leen los archivos multi-FASTA de ambos organismos
f1 <- readLines("C:/Users/Ana/Desktop/MASTER BIOINFORMATICA Y BIOESTADISTICA
UOC/TFM/PSI BLAST/29_3_2021/Paracoccus denitrificans/multiFASTA_output_PSIBLA
ST.txt")
f2 <- readLines("C:/Users/Ana/Desktop/MASTER BIOINFORMATICA Y BIOESTADISTICA
UOC/TFM/PSI BLAST/Bradyrhizobium japonicum/multiFasta_Bradyrhizobium.txt")

# Nos quedamos solo con la cabecera
f1 = grep(">", f1, value = TRUE)
f2 = grep(">", f2, value = TRUE)

# Se crea un bucle for que recorre f2(porque tiene más secuencias) y se ven
# almacenan en la variable "count" cuantas secuencias son coincidentes en f1
# y f2
```

```
count <- 0
for(i in f2){
  k <- 1
  j <- f1[k]
  while (i!=j && k<length(f1)) {
    k <- k+1
    j <- f1[k]
  }
  if (i==j){
    count <- count+1
  }
}

# De las 230 secuencias de f1, 218 también están presentes en f2, el 94.78%
count

## [1] 218

218*100/230

## [1] 94.78261
```

Anexo II: Secuencias obtenidas tras la búsqueda en PSI-BLAST

Acc number	Descripción	Organismo
OYW55572.1	nitric-oxide reductase large subunit	Hyphomicrobium sp. 12-62-95
OYX44479.1	nitric-oxide reductase large subunit	Rhodobacterales bacterium 32-67-9
GHG33044.1	nitric oxide reductase	Paracoccus aerius
AKE49362.1	nitric oxide reductase large subunit	Paracoccus sp. SY
PKP74566.1	nitric-oxide reductase large subunit	Alphaproteobacteria bacterium HGW-Alphaproteobacteria-6
OJY34953.1	nitric-oxide reductase large subunit	Rhodobacterales bacterium 65-51
KAB2882005.1	nitric-oxide reductase large subunit	Defluviimonas sp.
NEY91386.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium KMS-5
PKP85241.1	nitric-oxide reductase large subunit	Alphaproteobacteria bacterium HGW-Alphaproteobacteria-2
RVT81832.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium CCMM004
NJM84227.1	nitric-oxide reductase large subunit	Tabrizicola sp.
PKQ13637.1	nitric-oxide reductase large subunit	Alphaproteobacteria bacterium HGW-Alphaproteobacteria-1
KUO58476.1	nitric oxide reductase	Alphaproteobacteria bacterium BRH_c36
PIV78215.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium CG17_big_fil_post_rev_8_21_14_2_50_63_15
HCQ67325.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium
KAB2942771.1	nitric-oxide reductase large subunit	Hyphomicrobium sp.
KAF0172400.1	nitric oxide reductase subunit B	Rhodobacteraceae bacterium
AVO38633.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium SH-1
PLX37084.1	nitric-oxide reductase large subunit	Rhizobiales bacterium
HBZ45283.1	nitric-oxide reductase large subunit	Maritimibacter sp.

NNE80580.1	nitric-oxide reductase large subunit	Silicimonas sp.
KPP80931.1	nitric oxide reductase subunit B	Rhodobacteraceae bacterium HLUCCO07
KAF0113941.1	nitric oxide reductase subunit B	Rhodobacteraceae bacterium
HGG64587.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium
OIP85341.1	nitric-oxide reductase large subunit	Rhodobacterales bacterium CG2_30_65_12
HHI69849.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium
OUS38022.1	nitric-oxide reductase large subunit	Rhodobacterales bacterium 56_14_T64
HAW47897.1	nitric-oxide reductase large subunit	Roseovarius sp.
MBK44797.1	nitric-oxide reductase large subunit	Roseovarius sp.
MAU53187.1	nitric-oxide reductase large subunit	Roseovarius sp.
OZB20456.1	nitric-oxide reductase large subunit	Rhodobacterales bacterium 34-62-10
TNF17798.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium
PCJ07774.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium
PIY73018.1	nitric-oxide reductase large subunit	Rhodobacterales bacterium CG_4_10_14_0_8_um_filter_70_9
AAC79449.1	nitric oxide reductase large subunit precursor	Pseudomonas sp. G-179
PLX44269.1	nitric-oxide reductase large subunit	Rhizobiales bacterium
PIV73745.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium CG17_big_fil_post_rev_8_21_14_2_50_65_11
HHL43851.1	nitric-oxide reductase large subunit	Hellea balneolensis
EEW56939.1	nitric oxide reductase subunit B	Silicibacter sp. TrichCH4B
EDZ48515.1	nitric oxide reductase large suBunit, cytochrome b	Rhodobacterales bacterium Y4I
MTJ06000.1	nitric-oxide reductase large subunit	Sediminimonas qiaohouensis
MBT52745.1	nitric-oxide reductase large subunit	Mameliella sp.
MBD12547.1	nitric-oxide reductase large subunit	Roseovarius sp.
MAM59927.1	nitric-oxide reductase large subunit	Maritimibacter sp.
AUJ65881.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium
OED50652.1	nitric oxide reductase	Rhodobacteraceae bacterium (ex Bugula neritina AB1)
MAA98698.1	nitric-oxide reductase large subunit	Stappia sp.
HBU15540.1	nitric-oxide reductase large subunit	Gemmobacter sp.
PCJ89543.1	nitric-oxide reductase large subunit	Rhizobiales bacterium
NNE51529.1	nitric-oxide reductase large subunit	Sulfitobacter sp.
ABY68275.1	nitric-oxide reductase subunit B	Azospirillum baldaniorum
PWL33395.1	nitric-oxide reductase large subunit	Marivita sp. XM-24bin2
PWR03358.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium TG-679
KJS18673.1	nitric oxide reductase	Hoeflea sp. BRH_c9
OHC49259.1	nitric oxide reductase, partial	Rhodobacteraceae bacterium GWF1_65_7
PHS23544.1	nitric-oxide reductase large subunit	Robiginitomaculum sp.
PHR62718.1	nitric-oxide reductase large subunit	Robiginitomaculum sp.

WP_105384042.1	nitric-oxide reductase large subunit	Neorhizobium alkalisoli
RFP90230.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium 63075
WP_148144239.1	nitric-oxide reductase large subunit	Rhizobium cellulosilyticum
WP_012707560.1	nitric-oxide reductase large subunit	Sinorhizobium fredii
PWG16703.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium WDS4C29
OHC45483.1	nitric oxide reductase, partial	Rhodobacteraceae bacterium GWE1_64_9
WP_112801179.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Neorhizobium
WP_037449665.1	nitric-oxide reductase large subunit	Sinorhizobium fredii
PJN95169.1	nitric-oxide reductase large subunit	Amaricoccus sp. HAR-UPW-R2A-40
WP_040958526.1	nitric-oxide reductase large subunit	Sinorhizobium fredii
WP_077501438.1	MULTISPECIES: nitric-oxide reductase large subunit	Sinorhizobium/Ensifer group
WP_077546983.1	MULTISPECIES: nitric-oxide reductase large subunit	Rhizobium/Agrobacterium group
WP_105370200.1	nitric-oxide reductase large subunit	Neorhizobium huautlense
WP_120275922.1	nitric-oxide reductase large subunit	Rhizobium sp. AG855
WP_065794810.1	nitric-oxide reductase large subunit	Ensifer sp. LC163
WP_110792351.1	nitric-oxide reductase large subunit	Rhizobium wuzhouense
WP_060521134.1	MULTISPECIES: nitric-oxide reductase large subunit	Ensifer
WP_034790597.1	MULTISPECIES: nitric-oxide reductase large subunit	Ensifer
WP_060529039.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Ensifer
WP_014327923.1	nitric-oxide reductase large subunit	Sinorhizobium fredii
WP_053248643.1	nitric-oxide reductase large subunit	Ensifer adhaerens
WP_058329731.1	MULTISPECIES: nitric-oxide reductase large subunit	Sinorhizobium/Ensifer group
CAD6430857.1	nitric oxide reductase	Rhizobium sp. Q54
WP_052640605.1	nitric-oxide reductase large subunit	Pseudorhizobium banfieldiae
WP_028737944.1	MULTISPECIES: nitric-oxide reductase large subunit	Rhizobium
WP_152336167.1	nitric-oxide reductase large subunit	Rhizobium sp. TCK
RTL74767.1	nitric-oxide reductase large subunit	Bradyrhizobiaceae bacterium
WP_025430209.1	MULTISPECIES: nitric-oxide reductase large subunit	Ensifer
WP_062276239.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Rhizobium
WP_153439979.1	nitric-oxide reductase large subunit	Sinorhizobium terangae
CAC27381.1	nitric oxide reductase cytochrome b subunit	Achromobacter cycloclastes
WP_127890144.1	MULTISPECIES: nitric-oxide reductase large subunit	Sinorhizobium/Ensifer group
WP_037145454.1	nitric-oxide reductase large subunit	Rhizobium sp. YS-1r
WP_065781691.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Ensifer
WP_057251869.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Ensifer

WP_043616924.1	nitric-oxide reductase large subunit	Ensifer sp. ZNC0028
WP_077964986.1	nitric-oxide reductase large subunit	Ensifer adhaerens
WP_136506346.1	nitric-oxide reductase large subunit	Ensifer sp. MPMI2T
WP_113538314.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Ensifer
HBS50797.1	nitric-oxide reductase large subunit	Rhodobacteraceae bacterium
PZU21161.1	nitric-oxide reductase large subunit	Shinella sp.
WP_099059358.1	nitric-oxide reductase large subunit	Rhizobium sp. ACO-34A
WP_058322846.1	MULTISPECIES: nitric-oxide reductase large subunit	Sinorhizobium/Ensifer group
PJI38301.1	nitric-oxide reductase large subunit	Rhizobium sp.
WP_038590576.1	nitric-oxide reductase large subunit	Neorhizobium galegae
WP_104755335.1	nitric-oxide reductase large subunit	Ochrobactrum oryzae
MBA3041189.1	nitric-oxide reductase large subunit	Rhizobiaceae bacterium
WP_037431161.1	nitric-oxide reductase large subunit	Sinorhizobium fredii
WP_076397187.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Rhizobium
WP_140826294.1	nitric-oxide reductase large subunit	Rhizobium glycinendophyticum
WP_104666370.1	nitric-oxide reductase large subunit	Ensifer adhaerens
MBC7152261.1	nitric-oxide reductase large subunit	Rhizobium sp.
WP_105431262.1	nitric-oxide reductase large subunit	Neorhizobium sp. T6_25
WP_071201573.1	nitric-oxide reductase large subunit	Agrobacterium vitis
WP_061968291.1	nitric-oxide reductase large subunit	Bosea sp. Root670
WP_139678245.1	nitric-oxide reductase large subunit	Rhizobium smilacinae
WP_136559347.1	nitric-oxide reductase large subunit	Rhizobium sp. 7209-2
WP_069041047.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Agrobacterium
ODT23151.1	nitric oxide reductase	Kaistia sp. SCN 65-12
WP_105426816.1	nitric-oxide reductase large subunit	Neorhizobium tomejilense
WP_136600200.1	nitric-oxide reductase large subunit	Rhizobium ipomoeae
WP_141056775.1	nitric-oxide reductase large subunit	Brucella melitensis
WP_139975655.1	nitric-oxide reductase large subunit	Ochrobactrum sp. CGA5
WP_101442138.1	nitric-oxide reductase large subunit	Brucella melitensis
WP_105422855.1	nitric-oxide reductase large subunit	Neorhizobium sp. T25_27
WP_054149256.1	nitric-oxide reductase large subunit	Rhizobium sp. AAP116
WP_062579049.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Rhizobium
WP_018327474.1	nitric-oxide reductase large subunit	Rhizobium giardinii
WP_012093666.1	MULTISPECIES: nitric-oxide reductase large subunit	Ochrobactrum
WP_151643185.1	nitric-oxide reductase large subunit	Ochrobactrum tritici
WP_105440773.1	nitric-oxide reductase large subunit	Neorhizobium sp. T25_13
WP_004681527.1	MULTISPECIES: nitric-oxide reductase large subunit	Brucella
WP_018238764.1	nitric-oxide reductase large subunit	Ensifer sp. BR816
WP_010660606.1	MULTISPECIES: nitric-oxide reductase large subunit	Ochrobactrum
WP_094539469.1	nitric-oxide reductase large subunit	Ochrobactrum grignonense

WP_101433333.1	nitric-oxide reductase large subunit	Brucella melitensis
WP_002968298.1	nitric-oxide reductase large subunit	Brucella abortus
WP_097519557.1	nitric-oxide reductase large subunit	Sinorhizobium sp. BJ1
WP_113378561.1	nitric-oxide reductase large subunit	Rhizobium sp. SLBN-4
WP_140019728.1	nitric-oxide reductase large subunit	Ochrobactrum pecoris
WP_138145193.1	nitric-oxide reductase large subunit	Brucella sp. 10RB9215
WP_011970967.1	nitric-oxide reductase large subunit	Sinorhizobium medicae
WP_008936442.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Brucella
WP_064244366.1	nitric-oxide reductase large subunit	Ensifer glycinis
WP_127574700.1	nitric-oxide reductase large subunit	Sinorhizobium medicae
OYX75369.1	nitric-oxide reductase large subunit	Rhizobiales bacterium 32-66-11
WP_153492718.1	nitric-oxide reductase large subunit	Sinorhizobium medicae
WP_029928747.1	MULTISPECIES: nitric-oxide reductase large subunit	Ochrobactrum
WP_154169849.1	nitric-oxide reductase large subunit	Brucella sp. 10RB9213
WP_151091175.1	MULTISPECIES: nitric-oxide reductase large subunit	Ochrobactrum
WP_151661712.1	nitric-oxide reductase large subunit	Brucella intermedia
QFP60783.1	nitric-oxide reductase large subunit	Brucella melitensis
WP_071630711.1	nitric-oxide reductase large subunit	Ochrobactrum cytisi
WP_054160187.1	nitric-oxide reductase large subunit	Rhizobium sp. AAP43
WP_129333708.1	nitric-oxide reductase large subunit	Ciceribacter sp. F8825
WP_151655993.1	nitric-oxide reductase large subunit	Ochrobactrum sp. LMG 5442
WP_036350865.1	nitric-oxide reductase large subunit	Microvirga sp. BSC39
WP_002966337.1	MULTISPECIES: nitric-oxide reductase large subunit	Brucella
WP_151689276.1	nitric-oxide reductase large subunit	Ochrobactrum anthropi
WP_105541583.1	nitric-oxide reductase large subunit	Ochrobactrum sp. MYb71
WP_142593489.1	nitric-oxide reductase large subunit	Rhizobium endolithicum
WP_006466560.1	MULTISPECIES: nitric-oxide reductase large subunit	Brucellaceae
WP_037425244.1	nitric-oxide reductase large subunit	Sinorhizobium sp. CCBAU 05631
WP_101433160.1	nitric-oxide reductase large subunit	Brucella melitensis
KAB1120589.1	nitric-oxide reductase large subunit	Neorhizobium galegae
WP_112702321.1	nitric-oxide reductase large subunit	Brucella intermedia
WP_010967672.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
WP_124737065.1	nitric-oxide reductase large subunit	Brucella melitensis
WP_127641921.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
WP_014528651.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
HGG05328.1	nitric-oxide reductase large subunit	Aliiroseovarius sp.
WP_151576773.1	nitric-oxide reductase large subunit	Ochrobactrum anthropi
WP_028054805.1	MULTISPECIES: nitric-oxide reductase large subunit	Sinorhizobium
WP_006172098.1	MULTISPECIES: nitric-oxide reductase large subunit	Brucella
WP_141059706.1	nitric-oxide reductase large subunit	Brucella melitensis
WP_131606033.1	nitric-oxide reductase large subunit	Rhizobium leguminosarum

WP_105520061.1	MULTISPECIES: nitric-oxide reductase large subunit	Ochrobactrum
MBC7314691.1	nitric-oxide reductase large subunit	Rhizobium sp.
WP_007877571.1	MULTISPECIES: nitric-oxide reductase large subunit	Brucellaceae
WP_158528707.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
WP_014764525.1	nitric-oxide reductase large subunit	Sinorhizobium fredii
WP_127658321.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
WP_153494805.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
WP_064682771.1	MULTISPECIES: nitric-oxide reductase large subunit	Rhizobium
WP_100557614.1	MULTISPECIES: nitric-oxide reductase large subunit	Brucellaceae
HFB98026.1	nitric-oxide reductase large subunit	Bryobacterales bacterium
WP_064697068.1	nitric-oxide reductase large subunit	Rhizobium aegyptiacum
WP_154719273.1	nitric-oxide reductase large subunit	Rhizobium naphthalenivorans
WP_076772770.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Brucella
WP_127674949.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
WP_097545304.1	nitric-oxide reductase large subunit	Rhizobium anhuiense
WP_127659158.1	nitric-oxide reductase large subunit	Sinorhizobium medicae
WP_024895621.1	MULTISPECIES: nitric-oxide reductase large subunit	Ochrobactrum
WP_015242213.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
WP_094544236.1	nitric-oxide reductase large subunit	Ochrobactrum pseudogrignonense
WP_119255382.1	nitric-oxide reductase large subunit	Shinella zoogloeoides
WP_094508475.1	nitric-oxide reductase large subunit	Ochrobactrum thiophenivorans
WP_095918913.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
WP_117372379.1	nitric-oxide reductase large subunit	Shinella sp. WSJ-2
WP_094283800.1	nitric-oxide reductase large subunit	Brucella ceti
WP_095445109.1	nitric-oxide reductase large subunit	Ochrobactrum quorumnocens
WP_128094227.1	nitric-oxide reductase large subunit	Brucella pituitosa
WP_127712395.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
WP_127668036.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
CAE26899.1	nitric-oxide reductase subunit B	Rhodopseudomonas palustris CGA009
CUW45791.1	nitric-oxide reductase, large subunit	Brucella vulpis
WP_069062203.1	nitric-oxide reductase large subunit	Sinorhizobium sp. RAC02
WP_004688871.1	MULTISPECIES: nitric-oxide reductase large subunit	Brucella
WP_064331392.1	nitric-oxide reductase large subunit	Shinella sp. HZN7
WP_102763140.1	nitric-oxide reductase large subunit	Sinorhizobium sp. M4_45
WP_023081229.1	nitric-oxide reductase large subunit	Brucella canis
PPD10151.1	nitric-oxide reductase large subunit	Methylocystis sp.
TJW07042.1	nitric-oxide reductase large subunit	Mesorhizobium sp.
WP_006281056.1	nitric-oxide reductase large subunit	Brucella suis
WP_050743634.1	MULTISPECIES: nitric-oxide reductase large subunit	unclassified Shinella

WP_024271302.1	nitric-oxide reductase large subunit	Shinella sp. DD12
WP_132075437.1	nitric-oxide reductase large subunit	Sinorhizobium americanum
WP_056338997.1	nitric-oxide reductase large subunit	Rhizobium sp. Root482
WP_133033699.1	nitric-oxide reductase large subunit	Shinella granuli
WP_151613821.1	nitric-oxide reductase large subunit	Ensifer alkalisoli
HCH71545.1	nitric-oxide reductase large subunit	Ochrobactrum sp.
WP_069456899.1	nitric-oxide reductase large subunit	Ensifer alkalisoli
WP_037379897.1	nitric-oxide reductase large subunit	Sinorhizobium americanum
TIS57351.1	nitric-oxide reductase large subunit	Mesorhizobium sp.
TJW43853.1	nitric-oxide reductase large subunit	Mesorhizobium sp.
AJG06874.1	nitric oxide reductase large subunit	Azospirillum brasilense
WP_084364870.1	nitric-oxide reductase large subunit	Rhizobium sp. RU36D
ERM02179.1	nitric oxide reductase	Ochrobactrum intermedium 229E
WP_011085999.1	nitric oxide reductase subunit B	Bradyrhizobium diazoefficiens
WP_100673619.1	nitric-oxide reductase large subunit	Sinorhizobium meliloti
EAQ23442.1	Nitric oxide reductase large subunit, cytochrome b	Roseovarius sp. 217
WP_027999035.1	nitric-oxide reductase large subunit	Sinorhizobium arboris
WP_071019349.1	nitric-oxide reductase large subunit	Ensifer sp. LCM 4579

Tabla 3. Secuencias obtenidas tras la búsqueda en PSI-BLAST

Anexo III: Script de R para hallar la clasificación taxonómica de los organismos de cada clado de árbol

```
# Paquetes de R necesarios
library(myTAI)
library(plyr)

# Función para hallar la taxonomía
taxonomia <- function(names){

  df_list <- list()
  for (i in 1:length(names)){
    tax  <- taxonomy(organism = names[i], db = "ncbi",
                     output = "classification")

    df <- data.frame(lapply(tax$name, function(x) data.frame(x)))

    colnames(df) <- tax$rank
    df_list[[i]] <- df
    Sys.sleep(0.5)
  }
  df_list
}


##### Clasificación taxonómica del clado 1(A) #####

# Se carga el fichero que contiene el nombre y cada organismos y su ID taxonó
mico
tablaID_1 <- read.table("taxID_clado1.txt", sep="|", header = T)
```

```
# Se aplica la función
lista_clado_1 <- taxonomia(tablaID_1$name)

# Se obtiene la clasificación taxonómica de cada organismo de la lista
taxo_clado_1 <- do.call(rbind.fill, lista_clado_1)

# Se calcula el número de organismos presentes en cada clase
table(taxo_clado_1[,4])

##
## Alphaproteobacteria  Betaproteobacteria
##                 161                   1

# Se calcula el número de organismos presentes en cada orden
table(taxo_clado_1[,5])

##
##  Burkholderiales Hyphomicrobiales
##                1              161

# Se calcula el número de organismos presentes en cada familia, así como el
# la proporción de cada familia en el clado
tabla_familia <- data.frame(table(taxo_clado_1[,6]))
colnames(tabla_familia) <- c("Familia", "Frecuencia")
taxo_clado_1_length <- nrow(taxo_clado_1)
porcentaje <- c()
for (i in 1:nrow(tabla_familia)) {
  porcentaje <- c(porcentaje, (tabla_familia$Frecuencia[i] * 100) / taxo_clad
o_1_length)
}

tabla_familia$Porcentaje <- porcentaje

tabla_familia[order(-tabla_familia$Porcentaje),]

##               Familia Frecuencia Porcentaje
## 8        Rhizobiaceae        106  65.432099
## 4        Brucellaceae         47  29.012346
## 7  Phyllobacteriaceae          3   1.851852
## 3   Bradyrhizobiaceae          2   1.234568
## 1      Alcaligenaceae          1   0.617284
## 2           Boseaceae          1   0.617284
## 5         Kaistiaceae          1   0.617284
## 6 Methylobacteriaceae          1   0.617284

# Se exporta la clasificación completa a un fichero csv
write.csv(taxo_clado_1,"./complete_taxo_clado_1.csv", row.names = T)


##### Clasificación taxonómica del clado 2(B) #####

# Se carga el fichero que contiene el nombre y cada organismos y su ID taxonó
mico
tablaID_2 <- read.table("taxID_clado2.txt", sep="|", header = T)

# Se aplica la función
lista_clado_2 <- taxonomia(tablaID_2$name)

# Se obtiene la clasificación taxonómica de cada organismo de la lista
taxo_clado_2 <- do.call(rbind.fill, lista_clado_2)
```

```
# Se calcula el número de organismos presentes en cada clase
table(taxo_clado_2[,4])

##
##      Acidobacteriia Alphaproteobacteria
##                   1                  62

# Se exporta la clasificación completa a un fichero csv
write.csv(taxo_clado_2,"./complete_taxo_clado_2.csv", row.names = T)
```

Anexo IV: Tabla con la clasificación taxonómica de los organismos que conforman el clado A

superkingdom	phylum	class	order	family	genus	species
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Pseudorhizobium	Pseudorhizobium banfieldiae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. TCK
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	NA	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. Q54
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium endolithicum
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. YS-1r
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Neorhizobium	Neorhizobium alkalisoli
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Neorhizobium	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Neorhizobium	Neorhizobium huautlense
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium giardinii
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. Root482
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Neorhizobium	Neorhizobium galegae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Neorhizobium	Neorhizobium galegae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Neorhizobium	Neorhizobium sp. T6_25
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Neorhizobium	Neorhizobium sp. T25_13
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Neorhizobium	Neorhizobium tomejilense
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Neorhizobium	Neorhizobium sp. T25_27
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. RU36D
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Shinella	Shinella sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. SLBN-4
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium smilacinae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	NA	Pseudomonas sp. G-179
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium cellulosilyticum
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ciceribacter	Ciceribacter sp. F8825
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. ACO-34A
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	NA	Rhizobiaceae bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	NA

Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium naphthalenivorans
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium glycinendophyticum
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. AG855
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium wuzhouense
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Agrobacterium	Agrobacterium vitis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Kaistiaceae	Kaistia	Kaistia sp. SCN 65-12
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium ipomoeae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. AAP43
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. AAP116
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium sp. 7209-2
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Agrobacterium	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium terangae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer sp. MPMI2T
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer sp. LC163
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer adhaerens
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	NA	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	NA	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	NA	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer adhaerens
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer adhaerens
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer sp. ZNC0028
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Achromobacter	Achromobacter cycloclastes
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	NA	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium americanum
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium americanum
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium fredii
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium fredii
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium fredii
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium fredii
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium fredii
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer alkalisoli

Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer alkalisoli
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer sp. LCM 4579
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium arboris
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer sp. BR816
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium sp. BJ1
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Ensifer	Ensifer glycinis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium sp. CCBAU 05631
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium fredii
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium medicae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium medicae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium medicae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium medicae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium sp. M4_45
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium meliloti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium leguminosaru m
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium anhuiense
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Rhizobium	Rhizobium aegyptiacum
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Shinella	Shinella sp. HZN7
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Shinella	Shinella granuli
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Shinella	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Shinella	Shinella sp. DD12
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Sinorhizobium	Sinorhizobium sp. RAC02
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Shinella	Shinella zoogloeoides
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Rhizobiaceae	Shinella	Shinella sp. WSJ-2
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Phyllobacteria ceae	Mesorhizobium	Mesorhizobiu m sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Phyllobacteria ceae	Mesorhizobium	Mesorhizobiu m sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Phyllobacteria ceae	Mesorhizobium	Mesorhizobiu m sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Boseaceae	Bosea	Bosea sp. Root670

Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella melitensis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella melitensis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella ceti
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella melitensis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella melitensis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella melitensis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella melitensis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella canis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella suis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella melitensis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella abortus
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella sp. 10RB9215
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella sp. 10RB9213
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella vulpis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Ochrobactrum	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	NA	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella pseudogrignonensis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	NA	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Ochrobactrum	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella pituitosa
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella thiophenivorans
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Ochrobactrum	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	[Ochrobactrum] quorumnocens
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Ochrobactrum	Ochrobactrum sp. MYb71
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Ochrobactrum	Ochrobactrum sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella oryzae
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella cytisi
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Ochrobactrum	Ochrobactrum sp. CGA5
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Ochrobactrum	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella anthropi
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Ochrobactrum	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella tritici
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella anthropi
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella pecoris
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Ochrobactrum	NA

Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella grignonensis
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella intermedia
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella intermedia
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Ochrobactrum	Ochrobactrum sp. LMG 5442
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	NA	NA
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Brucellaceae	Brucella	Brucella intermedia
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Methylobacteriaceae	Microvirga	Microvirga sp. BSC39
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Bradyrhizobiaceae	Rhodopseudomonas	Rhodopseudomonas palustris
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Bradyrhizobiaceae	Bradyrhizobium	Bradyrhizobium diazoefficiens

Tabla 4. Clasificación taxonómica de los organismos que conforman el clado A

Anexo V: Tabla con la clasificación taxonómica de los organismos que conforman el clado B

Super kingdom	phylum	class	order	family	genus	species
Bacteria	Proteobacteria	Alphaproteobacteria	Maricaulales	Robiginitomaculaceae	Hellea	Hellea balneolensis
Bacteria	Proteobacteria	Alphaproteobacteria	Maricaulales	Robiginitomaculaceae	Robiginitomaculum	Robiginitomaculum sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Maricaulales	Robiginitomaculaceae	Robiginitomaculum	Robiginitomaculum sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	NA	NA	Hyphomicrobiales bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	NA	NA	Hyphomicrobiales bacterium
Bacteria	Acidobacteria	Acidobacteriia	Bryobacterales	NA	NA	Bryobacterales bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium GWF1_65_7
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium GWE1_64_9
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Gemmobacter	Gemmobacter sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Paracoccus	Paracoccus aerius
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Paracoccus	Paracoccus sp. SY
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Marivita	Marivita sp. XM-24bin2
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium WDS4C29
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium TG-679
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium 63075
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	NA	NA	Rhodobacterales bacterium 34-62-10
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Roseovarius	Roseovarius sp. 217
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Sulfitobacter	Sulfitobacter sp.

Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Sediminimonas	Sediminimonas qiaohouensis
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Mameliella	Mameliella sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Stappiaceae	Stappia	Stappia sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Phyllobacteriaceae	Hoeflea	Hoeflea sp. BRH_c9
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Maritimibacter	Maritimibacter sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Ruegeria	Silicibacter sp. TrichCH4B
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	NA	NA	Rhodobacterales bacterium Y4I
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	NA	NA	Hyphomicrobiales bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	NA	NA	Rhodobacterales bacterium 56_14_T64
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Amaricoccus	Amaricoccus sp. HAR-UPW-R2A-40
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	NA	NA	Rhodobacterales bacterium CG_4_10_14_0_8_um_filter_70_9
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Aliiroseovarius	Aliiroseovarius sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Hyphomicrobiaceae	Hyphomicrobium	Hyphomicrobium sp. 12-62-95
Bacteria	Proteobacteria	Alphaproteobacteria	NA	NA	NA	Alphaproteobacteria bacterium BRH_c36
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Hyphomicrobiaceae	Hyphomicrobium	Hyphomicrobium sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Roseovarius	Roseovarius sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Roseovarius	Roseovarius sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Roseovarius	Roseovarius sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Roseovarius	Roseovarius sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Pukyongiella	Pukyongiella litopenaei
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Silicimonas	Silicimonas sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium CCMM004
Bacteria	Proteobacteria	Alphaproteobacteria	NA	NA	NA	Alphaproteobacteria bacterium HGW-Alphaproteobacteria-1
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium HLUCCO07
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium

						CG17_big_fil_post_rev_8_21_14_2_50_63_15
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium CG17_big_fil_post_rev_8_21_14_2_50_65_11
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	NA	NA	Rhodobacterales bacterium CG2_30_65_12
Bacteria	Proteobacteria	Alphaproteobacteria	NA	NA	NA	Alphaproteobacteria bacterium HGW-Alphaproteobacteria-2
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Maritimibacter	Maritimibacter sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	NA	NA	Rhodobacterales bacterium 32-67-9
Bacteria	Proteobacteria	Alphaproteobacteria	NA	NA	NA	Alphaproteobacteria bacterium HGW-Alphaproteobacteria-6
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Defluviimonas	Defluviimonas sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	NA	NA	Rhodobacterales bacterium 65-51
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium KMS-5
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Tabrizicola	Tabrizicola sp.
Bacteria	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	NA	Rhodobacteraceae bacterium
Bacteria	Proteobacteria	Alphaproteobacteria	Hyphomicrobiales	Methylocystaceae	Methylocystis	Methylocystis sp.

Tabla 5. clasificación taxonómica de los organismos que conforman el clado B

Anexo VI: Script de R para obtener las secuencias que componen cada uno de los clados de árbol, con la finalidad de poder realizar un alineamiento múltiple de las mismas

```
setwd("C:/Users/Ana/Desktop/MASTER BIOINFORMATICA Y BIOESTADISTICA UOC/TFM/Perfil/Alineamiento_clados")

# Se carga el fichero multi-FASTA con todas las secuencias de la base de datos
library(seqinr)
multiFASTA <- read.fasta("./multiFASTA_output_PSIBLAST.fas", seqtype = "AA")

#############################################################################
# CLADO 1(A)

# Se cargan las etiquetas obtenidas a partir de MEGA X
```

```
labels_cado1 <- readLines("./labels_cado_1.txt")
labels_cado1 <- as.data.frame(labels_cado1)

labels_cado1_FASTA <- vector(mode = "list")
j <- 1

# Se usa un bucle for para seleccionar todas las secuencias pertenecientes al
# clado 1(A)
library(stringr)
for(i in multiFASTA){
  if(str_detect(labels_cado1, gsub("_", " ", seqinr::getName(i)))){
    cat(seqinr::getName(i),sep = "\n")
    labels_cado1_FASTA[[j]] <- i
    j <- j + 1
  }
}

write.fasta(labels_cado1_FASTA,
            seqinr::getAnnot(labels_cado1_FASTA),
            file.out = "labels_cado1_FASTA.fas",)

##############################################################################
###
# CLADO 2(B)

# Se cargan las etiquetas obtenidas a partir de MEGA X
labels_cado2 <- readLines("./labels_cado_2.txt")
labels_cado2 <- as.data.frame(labels_cado2)

labels_cado2_FASTA <- vector(mode = "list")
j <- 1

# Se usa un bucle for para seleccionar todas las secuencias pertenecientes al
# clado 2(B)
for(i in multiFASTA){
  if(str_detect(labels_cado2, gsub("_", " ", seqinr::getName(i)))){
    cat(seqinr::getName(i),sep = "\n")
    labels_cado2_FASTA[[j]] <- i
    j <- j + 1
  }
}


write.fasta(labels_cado2_FASTA,
            seqinr::getAnnot(labels_cado2_FASTA),
            file.out = "labels_cado2_FASTA.fas",)
```