

Análisis filogenético de la óxido nitroso reductasa (*nosZ*): establecimiento del perfil de la proteína

Sandra Segarra Jornet

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

Diciembre del 2021



Esta obra está sujeta a una licencia de
Reconocimiento-NoComercial-
SinObraDerivada [3.0 España de Creative Commons](https://creativecommons.org/licenses/by-nc-nd/3.0/es/)

FICHA DEL TRABAJO FINAL

Título del trabajo:	<i>Análisis filogenético de la óxido nitroso reductasa (nosZ): establecimiento del perfil de la proteína</i>
Nombre del autor:	<i>Sandra Segarra Jornet</i>
Nombre del consultor/a:	<i>Paloma Pizarro Tobías</i>
Nombre del PRA:	<i>Antoni Pérez Navarro</i>
Fecha de entrega (mm/aaaa):	12/2021
Titulación:	<i>Máster de Bioinformática y Bioestadística</i>
Área del Trabajo Final:	<i>Microbiología, Biotecnología y Biología Molecular</i>
Idioma del trabajo:	Castellano
Número de créditos:	15
Palabras clave	<i>nosZ, phylogenetic analysis, greenhouse gases</i>

Resumen del Trabajo

Uno de los objetivos principales del presente trabajo es la elaboración del análisis filogenético de la óxido nitroso reductasa, una enzima encargada de la reducción del N₂O a N₂ durante la desnitrificación biológica. El otro objetivo principal es el establecimiento del perfil de dicha proteína.

Para ello ha sido necesario crear una base de datos que incluya un conjunto de secuencias proteicas de distintos organismos microbiológicos, mediante la búsqueda de homólogos de la secuencia *query* en BLAST.

Posterior a la preparación de las secuencias, se realiza un alineamiento múltiple con el software MEGA-X, así como la determinación del modelo evolutivo y la señal filogenética, esta última usando TREE-PUZZLE.

Para el estudio de la clasificación taxonómica de los organismos poseedores del gen *nosZ* se prepara un árbol filogenético usando el método probabilístico de máxima verosimilitud con bootstrapping. De este análisis se obtiene la distinción de los organismos en dos Clados, según si poseen el gen "típico" o "atípico".

El último paso es la preparación del perfil proteico. Se ha preparado dos, uno para cada Clado, usando el servidor web Phyre2. Se han predicho las estructuras

tridimensionales de la proteína mediante el modelado por homología, ambas con un nivel de confianza del 100%.

Con los resultados obtenidos, se confirma que es posible ampliar y complementar los estudios ya hechos sobre la óxido nitroso reductasa usando bases de datos públicas, hecho que podría significar un paso importante en la potencial reducción de las emisiones de N₂O, un potente gas de efecto invernadero.

Abstract

One of the main objectives of this work is the phylogenetic analysis of nitrous oxide reductase, an enzyme responsible for the reduction of N₂O to N₂ during biological denitrification. The other main objective is the establishment of the profile of this protein.

To do this, it has been necessary to create a database that includes a set of protein sequences from different microbiological organisms, by searching for homologues of the *query* sequence in BLAST.

After the preparation of the sequences, a multiple alignment is carried out with the MEGA-X software, as well as the determination of the evolutionary model and the phylogenetic signal, the latter using TREE-PUZZLE.

For the study of the taxonomic classification of organisms possessing the *nosZ* gene, a phylogenetic tree is prepared using maximum likelihood probabilistic method with bootstrapping. From this analysis, the organisms are distinguished into two Clades, according to whether they possess the “typical” or “atypical” gene.

The last step is the preparation of the protein profile. Two have been prepared, one for each Clade, using the Phyre2 web server. The tree-dimensional structures of the protein have been predicted by homology modelling, both with a 100% confidence level.

With the results obtained, it is confirmed that is possible to expand and complement the studies already done on nitrous oxide reductase by using public databases, a fact that could mean an important step in the potential reduction of emissions of N₂O, a powerful greenhouse gas.

Índice

1	Introducción	1
1.1	Contexto y justificación del trabajo	1
1.2	Objetivos del Trabajo.....	1
1.2.1	Objetivos generales	1
1.2.2	Objetivos específicos	1
1.2	Enfoque y método a seguir	2
1.3	Planificación del trabajo	3
1.3.1	Tareas a realizar	3
1.3.2	Planificación temporal.....	3
1.4	Breve resumen de productos obtenidos	5
1.5	Breve descripción de los otros capítulos de la memoria	5
2	Estado del arte	6
3	Metodología	9
3.1	Creación y preparación de la base de datos.....	9
3.2	Alineamiento múltiple de las secuencias.....	10
3.3	Estudio del modelo evolutivo y la señal filogenética	10
3.4	Construcción del árbol filogenético	11
3.5	Preparación del perfil proteico	12
4	Resultados	14
4.1	La base de datos	14
4.2	El alineamiento múltiple.....	15
4.3	El modelo evolutivo y la señal filogenética	16
4.4	El árbol filogenético	17
4.5	El perfil de la proteína	20
4.5.1	Perfil para el Clado I.....	20
4.5.2	Perfil para el Clado II.....	23
5	Discusión	25
6	Conclusiones	26
7	Glosario.....	27
8	Bibliografía.....	29
9	Anexos	32

Lista de figuras

Figura 1. Workflow del estudio. En gris, las tareas; en azul, los softwares usados...	3
Figura 2. Diagrama de Gantt con la planificación temporal de las tareas y los hitos parciales. Cada color representa una PEC distinta.....	4
Figura 3. Ciclo del nitrógeno, en el que se describen las reducciones secuenciales que forman la desnitrificación. Las enzimas involucradas aparecen al lado de la reacción que catalizan (9)	6
Figura 4. Árbol filogenético creado por el método de máxima verosimilitud a partir de secuencias de aminoácidos de longitud completa del gen nosZ. Se indica la distribución del motivo del péptido señal detectado en cada uno de los clados. Los símbolos en las árboles especifican los grupos taxonómicos. (13)	7
Figura 5. Valores de los parámetros para el alineamiento múltiple de secuencias proteicas en MEGA-X.....	10
Figura 6. Valores de los parámetros para la determinación del modelo evolutivo en MEGA-X.	11
Figura 7. Parámetros seleccionados en TREE-PUZZLE para el análisis de la señal filogenética.....	11
Figura 8. Parámetros para la construcción del árbol filogenético	12
Figura 9. Representación de la frecuencia en el conteo de proteínas para todas las secuencias incluidas en la base de datos	15
Figura 10. Resultados del alineamiento múltiple para las 15 primeras secuencias. [Continúa en la siguiente página]	15
Figura 11. [Continuación]	16
Figura 12. Resultados del análisis "likelihood-mapping"	17
Figura 13. Árbol filogenético de las 222 secuencias proteicas. Se ha calculado usando máxima verosimilitud con Bootstrapping.	17
Figura 14. Recuento de las órdenes taxonómicas presentes en el Clado I del árbol filogenético.....	19
Figura 15. Plantillas con un 100% de confianza obtenidas durante el proceso de modelado por homología del perfil proteico para Aminobacter sp. (Clado I).	21
Figura 16. Resultados obtenidos de la modelización para predecir la estructura proteica de N ₂ OR. Este ha sido seleccionado como el mejor modelo.....	22
Figura 17. La figura A muestra la confianza del alineamiento, es decir, la fiabilidad de la alineación de la secuencia query. La figura B muestra la sensibilidad mutacional, que sirve para predecir si las mutaciones en la proteína pueden tener	

un efecto funcional o fenotípico. La figura C es la evaluación de la calidad ProQ2, un algoritmo de evaluación de la calidad del modelo.	22
Figura 18. Plantillas con un 100% de confianza obtenidas durante el proceso de modelado por homología del perfil proteico para Rhodobacteraceae bacterium (Clado II).	23
Figura 19. Resultados obtenidos de la modelización para predecir la estructura proteica de N ₂ OR. Este ha sido seleccionado como el mejor modelo.	24
Figura 20. Representación de la confianza del alineamiento (A), la sensibilidad mutacional (B) y la evaluación de la calidad (C).	24

Lista de tablas

Tabla 1. Los 10 modelos evolutivos calculados con el software MEGA.....	16
Tabla 2. Familias taxonómicas presentes en el Clado I del árbol filogenético construido en el apartado 4.3.	20

1 Introducción

1.1 Contexto y justificación del trabajo

El óxido nitroso (N_2O) es un gas de efecto invernadero, que a pesar de no ser uno de los principales gases emitidos, su potencial de calentamiento global es elevado. La mayor parte de las emisiones del óxido nitroso provienen de procesos de desnitrificación y nitrificación de bacterias y hongos en suelos. La intensificación de la agricultura y, por consiguiente, del uso de fertilizantes sintéticos a base de nitrógeno ha contribuido al aumento de sus emisiones.

La óxido nitrosa reductasa es una enzima que forma parte del sistema respiratorio de las bacterias, que es activa en condiciones anaeróbicas y con la presencia de nitrato o óxido nítrico. El gen que la codifica es el *nosZ*. Las especies poseedoras de este gen tienen la capacidad de convertir N_2O a N_2 , un gas inocuo y principal componente de la atmósfera.

Según las investigaciones más recientes, el estudio a nivel microbiológico y molecular de la desnitrificación y de la óxido nitroso reductasa podría significar la reducción de las emisiones de N_2O . Para ello, es necesario estudiar cuáles especies poseen el gen *nosZ* en su genoma, su clasificación taxonómica y el establecimiento del perfil de la proteína.

Con la caracterización de la enzima y la selección de especies poseedoras del gen, se facilitaría una futura investigación para su potencial uso en la disminución de las emisiones de gases de efecto invernadero.

1.2 Objetivos del Trabajo

Toda la línea de estudio que sigue el presente trabajo puede dividirse en dos objetivos principales, los cuales se pueden desglosar en cinco objetivos específicos

1.2.1 Objetivos generales

- Realización de un estudio filogenético de las secuencias del gen *nosZ* para un conjunto de especies previamente seleccionadas.
- Establecimiento del perfil de la proteína óxido nitroso reductasa.

1.2.2 Objetivos específicos

- Preparación de una base de datos de secuencias aminoácidas y proteicas del gen *nosZ* para un conjunto de especies seleccionadas en base a una secuencia molde.
- Realización de un alineamiento múltiple de todas las secuencias de la base de datos.
- Comprobación de la señal filogenética y determinación del modelo evolutivo más adecuado, para la correcta preparación del árbol filogenético

- Preparación de un árbol filogenético
- Establecer el perfil de la proteína
- Comprobación y validación del perfil, y comparación del perfil según los clados extraídos del árbol filogenético

1.2 Enfoque y método a seguir

El punto de partida del estudio es la búsqueda bibliográfica de un organismo cuya secuencia del gen *nosZ* esté bien secuenciada. Esta será usada como secuencia molde para preparar una base de datos.

Para la creación de la base de datos se hará uso de BLAST (Basic Local Alignment Search Tool) [1], un programa informático de alineamiento de secuencias capaz de comparar una query contra una gran cantidad de secuencias almacenadas en bases de datos, en concreto, se usará la de NCBI. Nuestra query será la secuencia molde, y como resultado se obtendrá un conjunto de secuencias homólogas.

El último paso para la creación y preparación de la base de datos es el filtrado, en el que se seleccionaran las secuencias y organismos de interés, además de realizar una limpieza de los datos para eliminar posibles errores tipográficos u organismos repetidos. Estos últimos pasos se harán usando el lenguaje de programación R [2].

Con las secuencias del gen *nosZ* agrupadas y preparadas para el estudio, se va a realizar un alineamiento múltiple usando el software MEGA [3].

Para asegurarse que el alineamiento no ha sido producido al azar es importante estudiar la señal filogenética. Se puede estudiar mediante el método “likelihood-mapping”, un método para la visualización del contenido filogenético de un alineamiento de secuencias. Esto va a hacerse con el software TREE-PUZZLE [4].

Previo a la construcción del árbol filogenético es necesario disponer del modelo evolutivo que mejor caracteriza el conjunto de secuencias de la base de datos. Para ello, se usa de nuevo MEGA. El criterio para la selección del mejor modelo se hará en base al criterio de información bayesiana (BIC score), que considera que el mejor modelo es el de menor puntuación.

La construcción del árbol filogenético se va a hacer con el alineamiento múltiple y el modelo evolutivo previamente preparados, y se aplicará un método de Bootstrap para asegurarse que los resultados son robustos. También se usará el software MEGA.

El último hito a conseguir es el establecimiento del perfil de la proteína. Se usará el programa Phyre2 [5], un servidor web para la predicción de estructuras tridimensionales de una secuencia proteica basándose en el modelado de

homologías. Se seleccionarán una o más secuencias proteicas según los resultados obtenidos en el análisis filogenético.

En el siguiente esquema (Figura 1) se muestra el workflow del estudio, con cada una de las tareas a realizar por orden cronológico y el software o servidor usado.

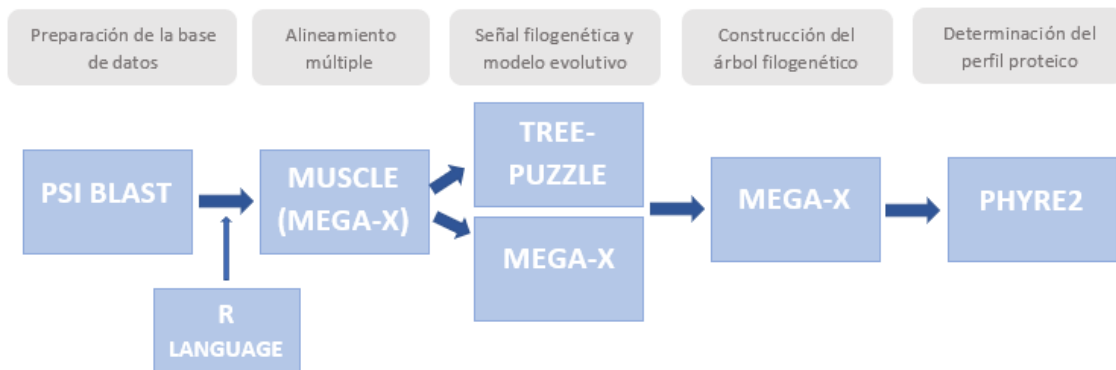


Figura 1. Workflow del estudio. En gris, las tareas; en azul, los softwares usados.

1.3 Planificación del trabajo

1.3.1 Tareas a realizar

Las tareas a realizar durante el estudio están estrechamente relacionadas con los objetivos específicos. Son las siguientes:

- Realizar una búsqueda bibliográfica para seleccionar una secuencia molde del gen *nosZ*
- Preparar una base de datos de secuencias de aminoácidos y de nucleótidos de especies seleccionadas a partir de la secuencia molde
- Filtrar y limpiar de la base de datos de secuencias previo al análisis
- Realizar un alineamiento múltiple de las secuencias
- Determinar el modelo evolutivo y la señal filogénica
- Construir un árbol filogenético
- Determinar el perfil proteico a partir del alineamiento de las secuencias
- Comprobar el perfil o perfiles creados, y realizar una comparación entre grupos según su filogenia

1.3.2 Planificación temporal

Los hitos que van marcando el desarrollo del proyecto se basan en las tareas nombradas en el apartado anterior. Se representa la temporalización de los hitos en la siguiente Figura 2.

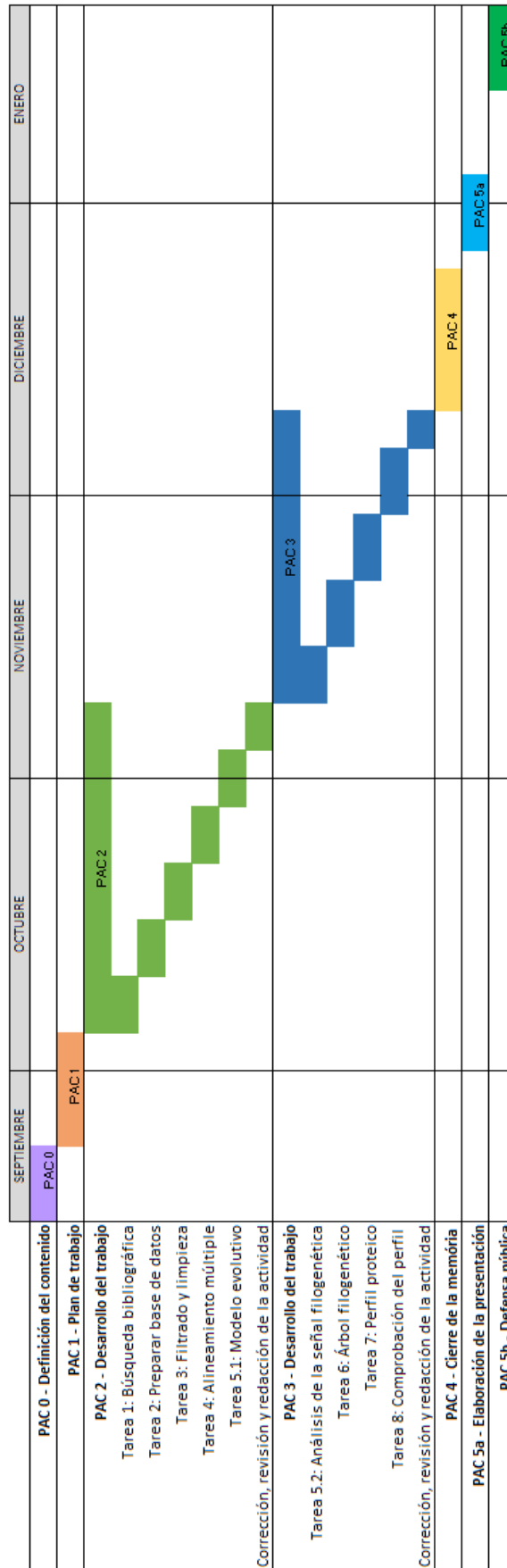


Figura 2. Diagrama de Gantt con la planificación temporal de las tareas y los hitos parciales. Cada color representa una PEC distinta.

1.4 Breve resumen de productos obtenidos

Los resultados esperados de este proyecto son una memoria y una presentación virtual.

Dentro de la memoria se incluirán los resultados obtenidos, a destacar:

- Una base de datos formada por un conjunto de organismos microbiológicos poseedores del gen *nosZ*.
- Un árbol filogenético, con la clasificación taxonómica de los organismos seleccionados.
- Un perfil proteico de la enzima óxido nítrico reductasa

1.5 Breve descripción de los otros capítulos de la memoria

El proyecto se estructurará en las siguientes partes:

- Introducción en la que se expondrá la justificación del trabajo, los objetivos y qué se espera obtener del estudio.
- Estado del arte con la información obtenida de la búsqueda bibliográfica, útil para poner sobre el papel toda la información necesaria previa al análisis.
- Metodología que incluirá las bases de datos públicas consultadas, todos los programas y softwares utilizados y los cálculos realizados.
- Resultados que reunirá todos los outputs obtenidos, tanto del alineamiento múltiple de secuencias, los árboles filogenéticos con su clasificación taxonómica, y el perfil de la proteína obtenida.
- Discusión y conclusiones en la que, a partir de los resultados obtenidos, se presentará una argumentación de los resultados obtenidos, así como una comparación con los estudios previos.
- Glosario con los conceptos y acrónimos que requieran de una descripción y explicación adicional.
- Bibliografía completa con todos los artículos científicos, capítulos de libros y páginas web consultados.
- Anexos con los scripts usados, tablas e información adicional.

2 Estado del arte

El óxido nitroso (N_2O) es el tercer gas de efecto invernadero más importante en cuanto a su potencial de calentamiento global, además de favorecer la destrucción de la capa de ozono. Las concentraciones de dicho gas han ido en aumento en (de 290ppb el 1940 a 330ppb el 2017), a un ritmo estrechamente ligado al aumento del nitrógeno reactivo (Nr). El Nr se puede originar de forma natural o antrópica, ya sea a través del proceso Haber-Bosch, la siembra de cultivos fijadores de nitrógeno o el uso de combustibles fósiles y biocombustibles [6].

Según los datos del Grupo Intergubernamental de Expertos sobre el Cambio Climático (IPCC) la principal fuente emisora de N_2O es la agricultura [7], ya que se genera este gas como subproducto de reacciones microbianas que usan el Nr como sustrato.

Las emisiones de N_2O de los suelos agrícolas provienen principalmente de los procesos biológicos nitrificantes y desnitrificantes, aunque se considera ésta última como la fuente principal [8].

La desnitrificación es un proceso fundamental en el ciclo del nitrógeno (ver Figura 3), en el cual la mayoría de procariontes y algunos hongos son capaces de usar los óxidos de nitrógeno como aceptores terminales de electrones.

El proceso completo incluye cuatro reducciones consecutivas, empezando por nitrato (NO_3^-), nitrito (NO_2^-), óxido nítrico (NO), óxido nitroso (N_2O), hasta dinitrógeno (N_2), que finalmente se libera a la atmósfera.

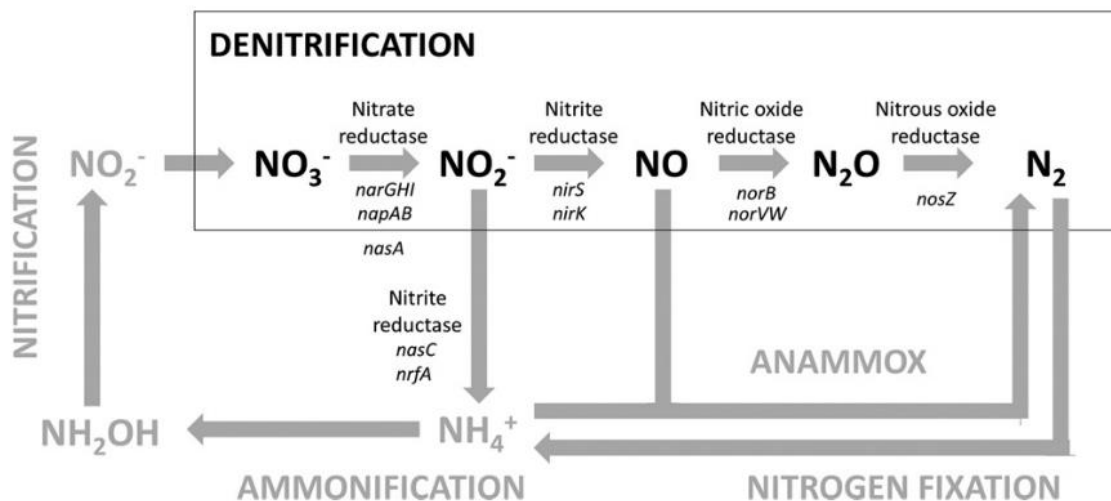


Figura 3. Ciclo del nitrógeno, en el que se describen las reducciones secuenciales que forman la desnitrificación. Las enzimas involucradas aparecen al lado de la reacción que catalizan [9]

Todas las reacciones en el proceso de desnitrificación son catalizadas por las metaloenzimas nitrato reductasa, nitrito reductasa, óxido nítrico reductasa y óxido nitroso reductasa.

En este trabajo nos centramos en la óxido nitroso reductasa (N₂OR), una enzima homodimérica encargada de catalizar la reducción de dos electrones de N₂O a N₂ y H₂O [10], codificada por el gen *nosZ*.

El gen *nosZ* ha sido objeto de muchos estudios para intentar dilucidar el control de las emisiones de N₂O de los suelos agrícolas. En el 2012 un estudio genómico identificó un gen *nosZ* “atípico”, el Clado II *nosZ*, a menudo más abundante en los suelos que el gen “típico” Clado I [11]. La principal diferencia entre los dos tipos de genes es que los microorganismos que tienen el Clado I tienen el potencial genético para ser desnitrificadores completos, mientras que solamente algunos de los microorganismos con el gen del Clado II poseen la capacidad para ello.

El Clado I *nosZ* está representado por los ya bien estudiados microorganismos desnitrificantes como *Bradyrhizobium spp.*, *Paraccocus denitrificans*, *Neisseria spp.* o *Pseudomonas spp.*, es decir, principalmente de los subgrupos alfa, beta y gamma del filo Proteobacteria. En cambio, los microorganismos del Clado II *nosZ* forman parte de múltiples filos bacterianos. Los recientes estudios de secuenciación metagenómica han demostrado el predominio de los microorganismos del Clado II en prácticamente todos los ecosistemas [12].

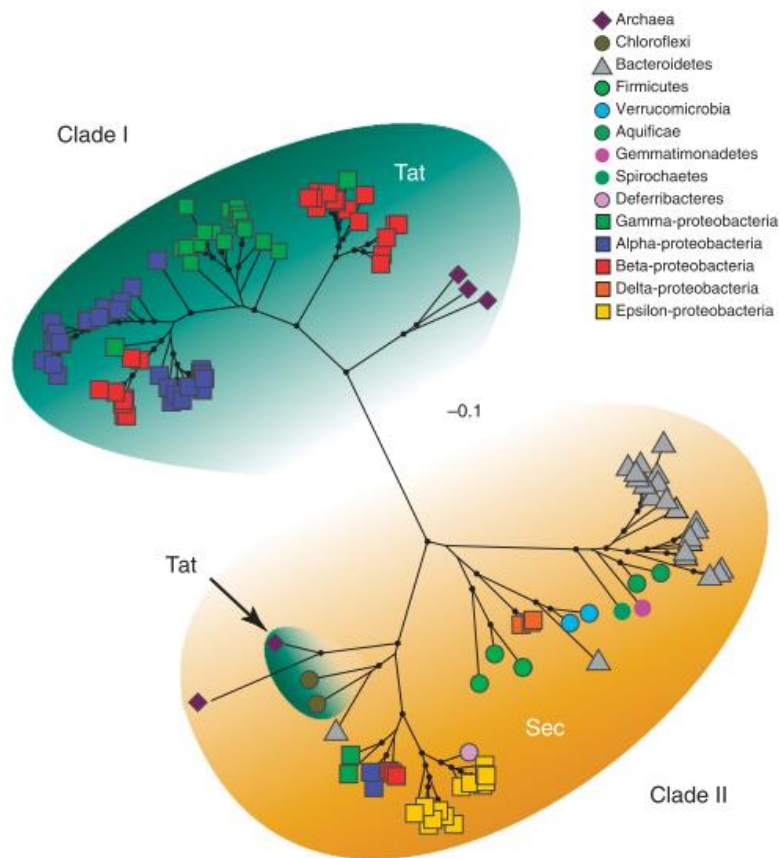


Figura 4. Árbol filogenético creado por el método de máxima verosimilitud a partir de secuencias de aminoácidos de longitud completa del gen *nosZ*. Se indica la distribución del motivo del péptido señal detectado en cada uno de los clados. Los símbolos en las árboles especifican los grupos taxonómicos. [13]

Por otra parte, el estudio de los elementos estructurales de N₂OR ha revelado que los dos clados difieren en la distribución de la señal peptídica, indicando diferencias en el mecanismo de secreción para la translocación de N₂O a través de la membrana celular: Tat para el Clado I y Sec para el Clado II [13]. De hecho, esta distinción ha sido usada para la creación de árboles filogenéticos, tal y como se observa en la Figura 4.

A partir de los conocimientos adquiridos sobre el gen *nosZ* y su filogenia se han hecho estudios para determinar si es posible mitigar las emisiones de N₂O provenientes de los suelos agrícolas.

Por ejemplo, se sabe que los cultivos de soja no sólo emiten N₂O a causa de los fertilizantes y el nitrógeno del suelo, sino que la descomposición de los nódulos después de su cosecha también es una fuente emisora importante. Se comprobó que la inoculación de cepas de *nosZ*⁺⁺ (mutante con mayor actividad de N₂OR) de *B. japonicum* reduce exitosamente las emisiones de N₂O [14].

Para los organismos del Clado I, que tienen la capacidad de reducir completamente a dinitrógeno, la cantidad de N₂O y N₂ emitida depende de factores ambientales como la concentración de NO₃⁻, el pH o la aireación del suelo [11]. De hecho, hay estudios que modifican estas variables físico-químicas con el fin de aumentar la actividad de N₂OR. Es el caso de un estudio [15] en el que se acidifica levemente el pH del suelo mediante la introducción de dolomita, que promueve la transcripción del gen *nosZ*.

Otro ejemplo es la aplicación de *biochar*, un carbón vegetal producido por pirólisis, que juntamente con la modificación del pH y el aumento en la abundancia de genes *nosZ* es capaz de reducir eficazmente las emisiones de N₂O [16].

3 Metodología

3.1 Creación y preparación de la base de datos

Para la creación de una base de datos propia que contenga un conjunto de secuencias proteicas del gen *nosZ* en distintas especies es necesario disponer de una secuencia molde. Esta secuencia molde no es más que una secuencia del gen en cuestión de una especie elegida según un criterio básico: que la secuencia esté bien caracterizada. El objetivo es usarla como punto de partida para encontrar secuencias de proteínas homologas de otras especies.

Para la elección de la secuencia molde ha sido necesario realizar una búsqueda bibliográfica. Se ha optado por elegir la secuencia proteica del gen *nosZ* de *Achromobacter cycloclastes*. El motivo de la elección es que el gen en esta especie se aisló hace ya más de treinta años [17] y poco después se logró clonar y secuenciar [18], por lo que ha sido ampliamente estudiado desde entonces. Se ha obtenido la secuencia en formato FASTA de la base de datos de secuencias de proteínas UniProt [19]. Se puede consultar la secuencia en el [Anexo I](#).

La preparación de la base de datos se ha hecho mediante la herramienta PSI-BLAST (Position-Specific Iterated - Basic Local Alignment Tool), que permite realizar una búsqueda de secuencias similares a la secuencia query (en este caso, la secuencia molde) para identificar regiones de alineamiento local, obteniéndose una matriz de puntuación PSSM (position-specific scoring matrix) construida a partir del alineamiento de secuencias múltiple. Las puntuaciones de la matriz serán usadas para realizar iteraciones en la búsqueda, de forma que la sensibilidad de BLAST aumentará, así como la potencial identificación de nuevas relaciones entre la query y las entradas de las bases de datos.

Una vez introducida la secuencia proteica en formato FASTA dentro de BLAST, se ha seleccionado la “Non-redundant protein sequences (nr)” como base de datos donde realizar la búsqueda y se han mantenido todos los parámetros predeterminados excepto el número de secuencias y el umbral restrictivo, que se han cambiado a 1000 y 0.001, respectivamente. Tras 9 iteraciones, se obtiene convergencia en los resultados.

El siguiente paso es calcular cuantas secuencias de la proteína de interés han sido encontradas, ya que en los resultados aparecen varias. Tras la descarga de los datos completos, se crea un pequeño programa en R para ver cuántos tipos de proteínas distintos hay. En total, hay 20 tipos diferentes, de las cuales se seleccionan las nombradas “nitrous-oxide reductase”, “Nitrous-oxide reductase”, “nitrous-oxide reductase NosZ”, “nitrous oxide reductase” y “nosZ”. Tras el filtrado, se obtienen 222 secuencias.

El filtrado y limpieza de los datos se ha hecho usando R y los paquetes Biostrings y seqinr, el código R usado puede consultarse en el [Anexo II](#).

3.2 Alineamiento múltiple de las secuencias

Para realizar el alineamiento múltiple de las 222 secuencias obtenidas, se usa el software MEGA (Molecular Evolutionary Genetics Analysis), una herramienta para realizar análisis estadísticos de la evolución molecular y para la construcción de árboles filogenéticos. El software incluye MUSCLE (Multiple Sequence Comparison by Log-Expectation) [20], un programa para generar alineamientos múltiples de secuencias, con un alto rango de precisión en comparación con otros programas.

Después de cargar el fichero de las secuencias filtradas preparado en R en formato FASTA a MEGA, se usa el programa MUSCLE. Para el alineamiento múltiple se usan los valores por defecto, que pueden observarse en la siguiente Figura 5.

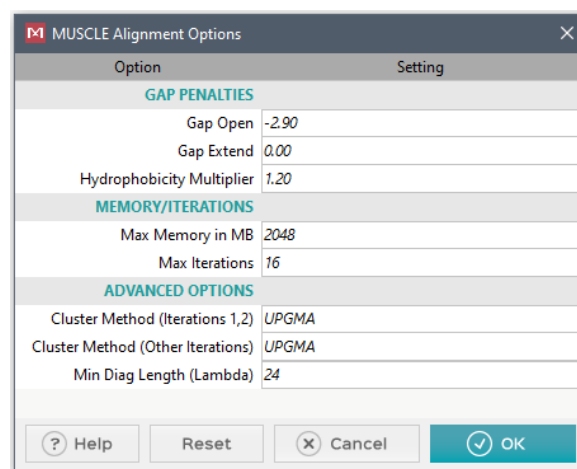


Figura 5. Valores de los parámetros para el alineamiento múltiple de secuencias proteicas en MEGA-X.

Los resultados del alineamiento obtenidos se guardan en un fichero FASTA, el cual ha tenido que ser revisado manualmente ya que ha aparecido algún espacio en blanco al final de las secuencias que se ha tenido que borrar.

3.3 Estudio del modelo evolutivo y la señal filogenética

Para la creación del árbol filogenético es importante disponer del modelo evolutivo y de la señal filogenética. Es, por lo tanto, necesario determinarlos para poder desarrollar el siguiente apartado.

El modelo evolutivo es un parámetro que explica la probabilidad que una secuencia cambie a otra secuencia homóloga con el tiempo, que se calcula según la frecuencia de cada proteína y las tasas de sustituciones. Por otra parte, la señal filogenética es la tendencia de los homólogos a expresar rasgos similares, que es útil para conocer si el alineamiento se ha producido por azar o si realmente las secuencias son homólogas.

Para determinar el modelo evolutivo, se usa el fichero generado en el alineamiento múltiple y se usa la opción Models/Best Fit Protein Models (ML) del software MEGA, sin cambiar los parámetros por defecto:

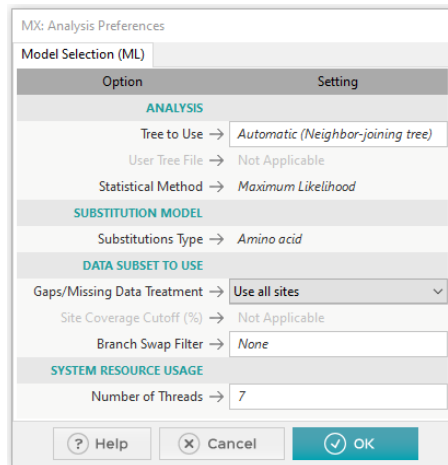


Figura 6. Valores de los parámetros para la determinación del modelo evolutivo en MEGA-X.

Para la comprobación de la señal filogenética se usa TREE-PUZZLE, un programa para la visualización del contenido filogenético de un alineamiento de secuencias. Este método está basado en el análisis “likelihood-mapping”. La selección de los distintos parámetros necesarios para el análisis se puede visualizar en la Figura 7.

Antes del análisis ha sido necesario transformar el formato del fichero de secuencias. Se ha hecho uso de la herramienta EMBOSS Seqret para convertir el fichero FASTA a .phy.

```

GENERAL OPTIONS
b          Type of analysis?      Likelihood mapping
v          Quartet evaluation criterion? Approximate maximum likelihood (ML)
g          Group sequences in clusters? No
n          Number of quartets?    10000 (random choice)
e          Parameter estimates?    Approximate (faster)
x          Parameter estimation uses? Neighbor-joining tree
SUBSTITUTION PROCESS
d          Type of sequence input data? Auto: Amino acids
m          Model of substitution?   Auto: WAG (Whelan-Goldman 2000)
f          Amino acid frequencies?  Estimate from data set
RATE HETEROGENEITY
w          Model of rate heterogeneity? Uniform rate
Quit [q], confirm [y], or change [menu] settings:
  
```

Figura 7. Parámetros seleccionados en TREE-PUZZLE para el análisis de la señal filogenética

3.4 Construcción del árbol filogenético

La construcción del árbol filogenético a partir de las secuencias alineadas se hace con el software MEGA-X.

Se emplea el método de “maximum likelihood” o máxima verosimilitud, que trata de determinar la topología que posee la mayor probabilidad para explicar los cambios en los nucleótidos entre las diferentes secuencias (21).

Para obtener el árbol filogenético es necesario especificar qué modelo evolutivo se ha producido. En el apartado anterior se determinó que el modelo LG es el que mejor se ajusta a los datos.

Además, se usa el método por Bootstrap, con un total de 100 réplicas. Todos estos parámetros se pueden consultar en la siguiente Figura 8.

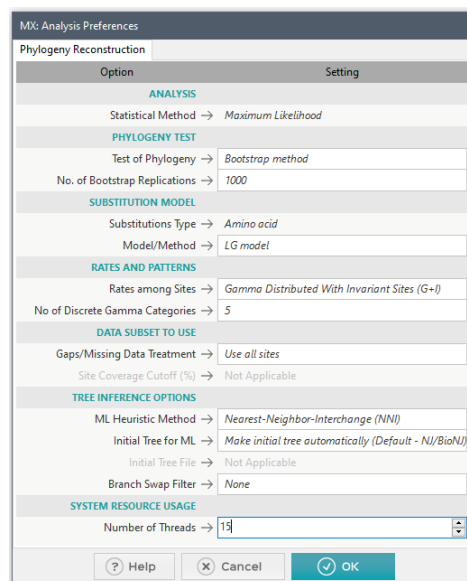


Figura 8. Parámetros para la construcción del árbol filogenético

3.5 Preparación del perfil proteico

Para la determinación del perfil proteico se ha usado el servidor web Phyre2, que se basa en la predicción de estructuras tridimensionales de proteínas mediante el modelado por homología.

A grandes rasgos, su funcionamiento se puede resumir en cuatro etapas (5):

- Recopilación de secuencias homólogas con, por lo menos, un 20% de identidad con la secuencia query. Después se realiza un alineamiento múltiple, que sirve para predecir la estructura secundaria de la proteína. El alineamiento y la estructura predicha se combinan en un Hidden Markov Model (HMM).
- Escaneo del modelo anterior contra una base de datos de modelos HMM con estructuras proteicas conocidas. Los alineamientos con un puntaje más alto se usan para construir un modelo estructural.
- Corrección de los indels en el modelo mediante un modelado en bucle
- Agregación de cadenas de aminoácidos para generar el modelo final.

Para llevar a cabo todas estas etapas es necesario introducir una secuencia de aminoácidos, que será usada como *query*.

La elección de las secuencias se ha hecho basándonos con los resultados del árbol filogenético, en el cual se observan dos ramas claramente diferenciadas: los organismos poseedores del gen Clado I y los poseedores del Clado II. Es por eso que se crean dos perfiles proteicos distintos, una para cada grupo.

En concreto, se ha seleccionado la secuencia de *Aminobacter sp.* para el Clado I, y *Rhodobacteraceae bacterium* para el Clado II. Se pueden consultar sus secuencias proteicas en el **Anexo III**.

4 Resultados

4.1 La base de datos

Usando la secuencia proteica de *Achromobacter cycloclastes* como secuencia molde en BLAST, se obtiene un listado de 1000 secuencias. En total, hay descritas 20 tipos de proteínas distintas, que son:

- Carbohydrate-binding and sugar hydrolysis
- Crystal Structure Of Nitrous Oxide Reductase From P. Denitrificans
- nitrous-oxide reductase
- Nitrous-oxide reductase
- nitrous-oxide reductase NosZ
- nitrous-oxide reductase precursor
- Nitrous-oxide reductase precursor
- nitrous-oxide reductase precursor
- Nitrous-oxide reductase precursor
- nitrous-oxide reductase, TAT-dependent
- nitrous oxide reductase
- nitrous oxide reductase apoprotein
- Nitrousoxide reductase EC 17996
- nosZ
- putative nitrous-oxide reductase NosZ
- RecName: Full=Nitrous-oxide reductase; AltName: Full=N(2)OR; AltName: Full=N₂O reductase; Flags: Precursor
- Resting form of pink nitrous oxide reductase from *Achromobacter Cycloclastes*
- TAT-dependent nitrous-oxide reductase
- TPA: nitrous-oxide reductase
- TPA: TAT-dependent nitrous-oxide reductase

Basándonos en los nombres recomendados para el gen *nosZ* en UniProt, seleccionamos únicamente cinco tipos: “nitrous-oxide reductase”, “Nitrous-oxide reductase”, “nitrous-oxide reductase NosZ”, “nitrous oxide reductase” y “nosZ”. El resultado obtenido es una base de datos de 222 secuencias distintas, que pueden consultarse en el [Anexo IV](#).

En la siguiente Figura 9 se puede observar como las 222 secuencias que conforman la base de datos tienen una longitud muy parecida. Están compuestas, de media, por 640 proteínas. La secuencia más larga tiene 652 y la más corta 584.

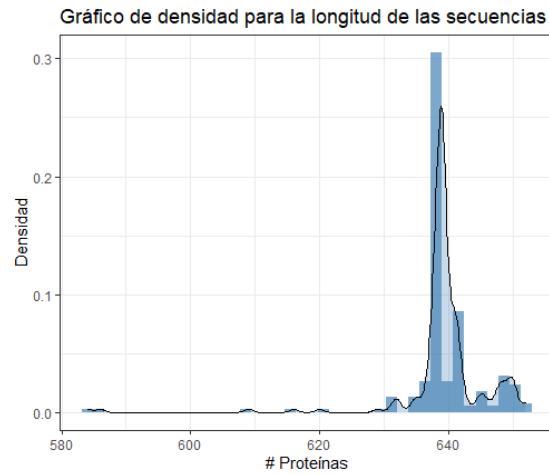


Figura 9. Representación de la frecuencia en el conteo de proteínas para todas las secuencias incluidas en la base de datos

4.2 El alineamiento múltiple

Los resultados del alineamiento múltiple de las 222 secuencias se presentan en la Figura 10.

Se observa como el extremo N-terminal presenta más variaciones entre las secuencias, pero el resto de regiones parecen conservarse bien entre las distintas especies.

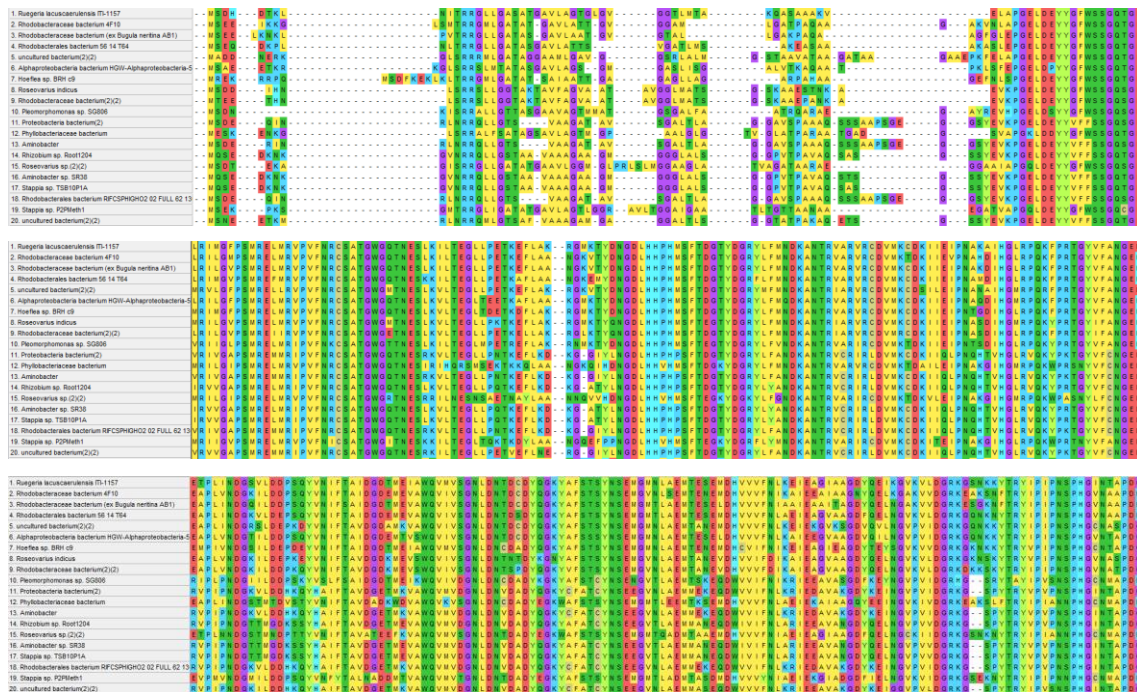


Figura 10. Resultados del alineamiento múltiple para las 15 primeras secuencias. [Continúa en la siguiente página]

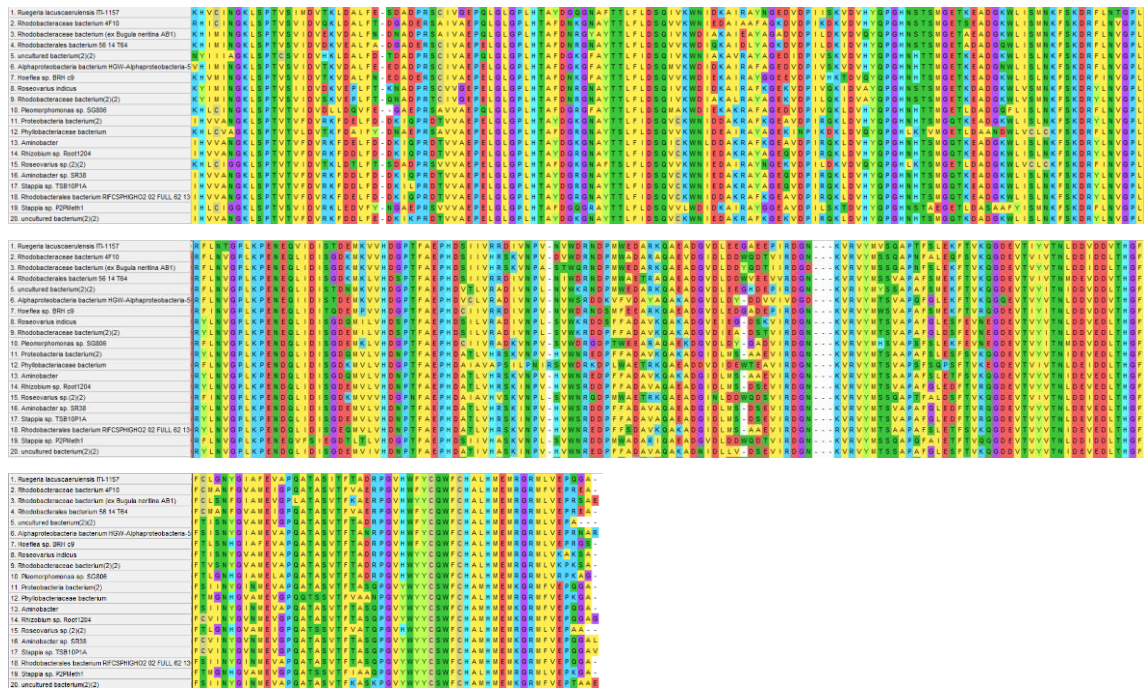


Figura 11. [Continuación]

4.3 El modelo evolutivo y la señal filogenética

Con las secuencias alineadas, se ha determinado los modelos evolutivos que mejor explican nuestros datos. Se pueden observar en la Tabla 1 a continuación.

Model	#Param	BIC	AICc	InL
LG+G+I	443	82520,05317	78153,28089	-38632,2509
LG+G	442	82621,38094	78264,45967	-38688,84655
LG+G+I+F	462	82774,38869	78220,45264	-38646,71496
WAG+G+I	443	82861,49965	78494,72737	-38802,97413
LG+G+F	461	82882,36972	78338,28414	-38706,63725
WAG+G	442	82988,63789	78631,71662	-38872,47503
rtREV+G+I+F	462	83218,43512	78664,49907	-38868,73818
WAG+G+I+F	462	83220,31746	78666,38141	-38869,67934
rtREV+G+F	461	83318,5918	78774,50622	-38924,74829

Tabla 1. Los 10 modelos evolutivos calculados con el software MEGA

Para la elección del mejor modelo nos basamos en el BIC score (Bayesian Information Criterion), ya que hay diversos estudios que demuestran que es el parámetro más apropiado para la selección del modelo con más precisión [22].

Teniendo esto en cuenta, el BIC de menor puntuación corresponde al modelo LG+G+I, que es el modelo que seleccionamos.

Los resultados obtenidos del análisis "likelihood-mapping" se presentan en la siguiente Figura 12. Resultados del análisis "likelihood-mapping"

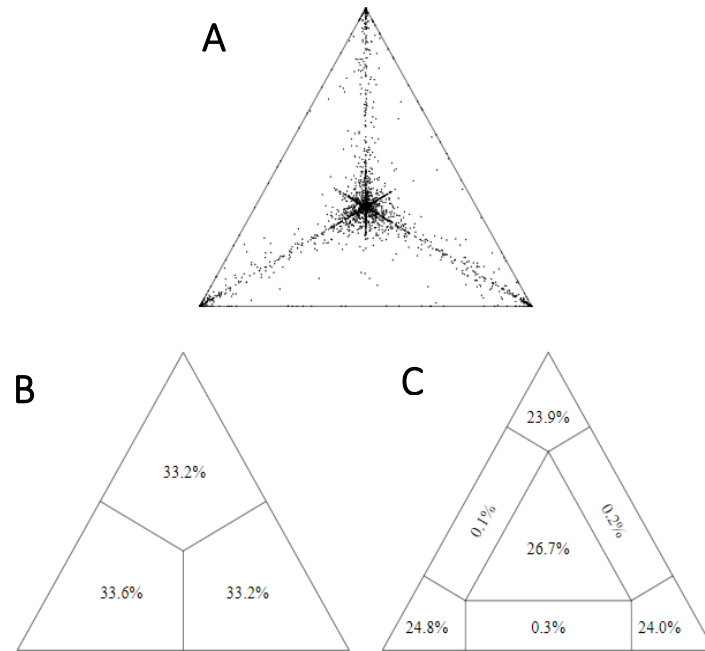


Figura 12. Resultados del análisis "likelihood-mapping"

El primer triángulo (A) muestra cada alineamiento de secuencias, mientras que los triángulos de abajo representan la atracción de los datos para cada una de las tres tipologías de árboles (B) y las ocupaciones en porcentaje de las áreas de atracción divididas en siete tipos (C).

Para determinar si la señal filogenética es correcta nos fijamos en la distribución de los puntos del triángulo A: la nube de puntos se distribuye entre el centro y las esquinas. Esto significa que hay una buena señal.

De hecho, los triángulos A y B reafirman esa idea, ya que la proporción se distribuye uniformemente entre las 3 áreas (triángulo B), y tiende a distribuirse hacia las esquinas (triángulo C).

4.4 El árbol filogenético

El árbol filogenético se ha construido usando el alineamiento múltiple de las secuencias, y usando el modelo evolutivo seleccionado en el apartado anterior.

El árbol resultante se puede observar en la Figura 13.

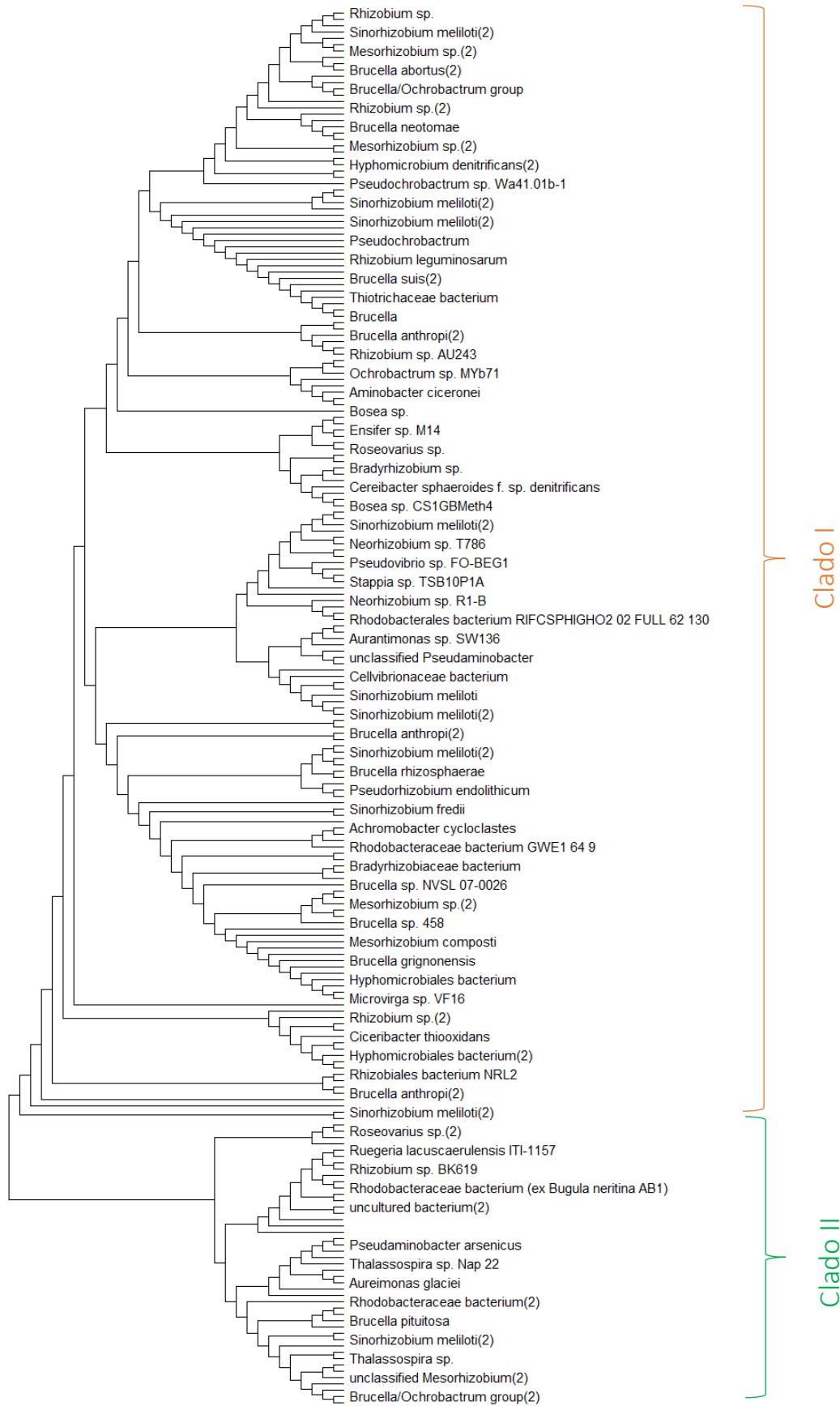


Figura 13. Árbol filogenético de las 222 secuencias proteicas. Se ha calculado usando máxima verosimilitud con Bootstrapping

En el árbol aparecen dos ramificaciones principales que se corresponden con los dos grupos o clados distintos.

Por una parte hay el Clado I, que agrupa 162 de los microorganismos presentes en la base de datos (el 73%) y el Clado II, formado por 60 (el 27%).

Tal y como se ha dicho en los apartados anteriores, los organismos del Clado II son muy diversos, es por eso que nos centraremos en el Clado I.

Para hacer una comprobación de la taxonomía de los organismos presentes en el Clado I se prepara un pequeño script en R, que puede consultarse en el [Anexo V](#).

Los resultados obtenidos de la comprobación taxonómica dejan claro que el fílum que forma la base de datos es el Proteobacteria. En cuanto a la clase, las Alphaproteobacteria son predominantes (el 97% de los datos), siendo las Gammaproteobacterias residuales (3%).

La orden taxonómica más abundante en el Clado I es Hyphomicrobiales, con una gran diferencia al resto: Rhodobacterales, Rhodospirillales, Thiotrichales, Cellvibrionales y Xanthomonadales. Los porcentajes de ellas son 85.2%, 8.0%, 2.5%, 0.6%, 0.6% y 0.6% respectivamente.

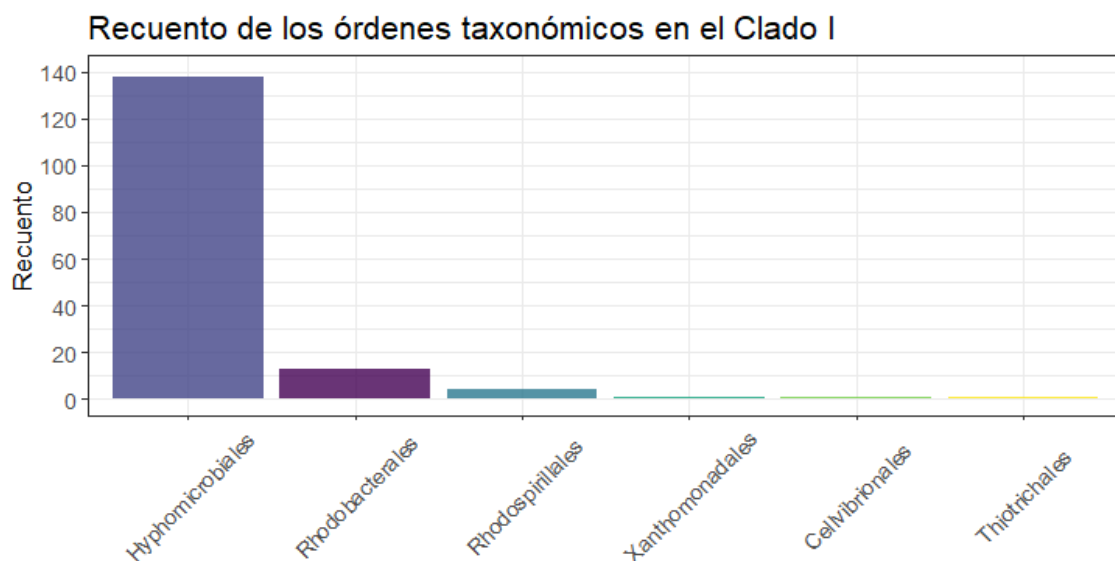


Figura 14. Recuento de las órdenes taxonómicas presentes en el Clado I del árbol filogenético.

También se ha hecho una comprobación de las familias taxonómicas presentes. En total, se han obtenido 17 familias distintas, entre las cuales destaca, Brucellaceae, Rhizobiaceae y Phyllobacteriaceae.

La siguiente Tabla 2 muestra los resultados obtenidos.

Familia	Recuento	Porcentaje (%)
Brucellaceae	51	31,5
Rhizobiaceae	45	27,8
Phyllobacteriaceae	23	14,2
Rhodobacteraceae	8	4,9
Boseaceae	4	2,5
Stappiaceae	3	1,9
Thalassospiraceae	3	1,9
Methylobacteriaceae	3	1,9
Roseobacteraceae	2	1,2
Bradyrhizobiaceae	2	1,2
Kaistiaceae	1	0,6
Xanthomonadaceae	1	0,6
Beijerinckiaceae	1	0,6
Aurantimonadaceae	1	0,6
Cellvibrionaceae	1	0,6
Thiotrichaceae	1	0,6
Azospirillaceae	1	0,6

Tabla 2. Familias taxonómicas presentes en el Clado I del árbol filogenético construido en el apartado 4.3.

4.5 El perfil de la proteína

4.5.1 Perfil para el Clado I

Para la creación del perfil proteico del Clado I se ha introducido la secuencia de *Aminobacter sp.* En el servidor web Phyre2, el cual devuelve una lista clasificada de modelos basada en diferentes plantillas, ordenadas según la puntuación en el alineamiento múltiple.

En nuestro caso, hay 7 plantillas con una confianza del 100%, es decir, con una alta probabilidad que la secuencia introducida y la plantilla obtenida sean homólogos. Además, el porcentaje de identidad obtenido en todas ellas también es moderadamente alto, oscilando entre el 69 y 61%.

Se pueden consultar las 7 platillas obtenidas en la Figura 15.

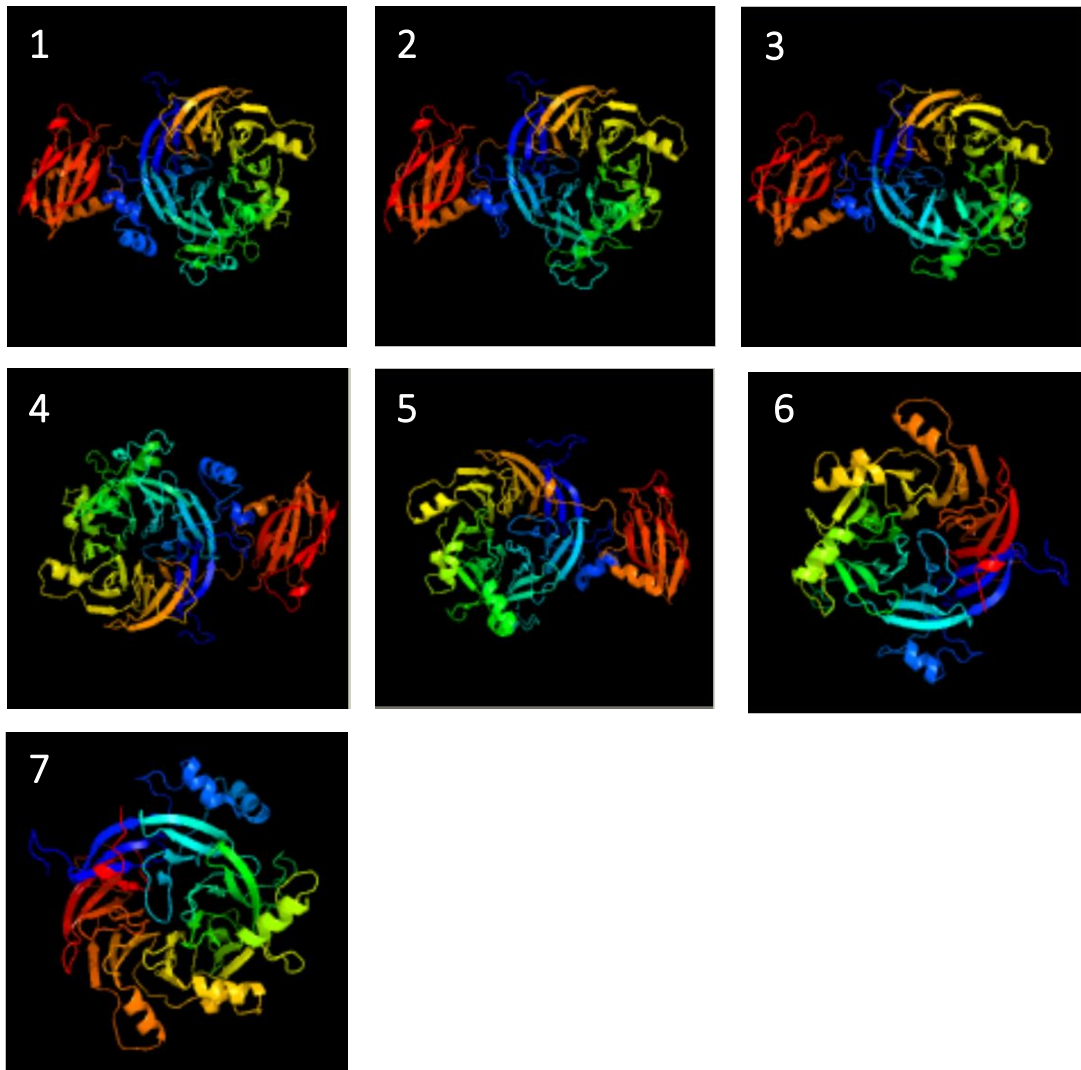


Figura 15. Plantillas con un 100% de confianza obtenidas durante el proceso de modelado por homología del perfil proteico para *Aminobacter* sp. (Clado I).

Todas las plantillas anteriores han sido satisfactoriamente detectadas como molécula PDB óxido nitroso reductasa. Otra información adicional para cada una:

- Plantilla 1: forma unida al inhibidor de óxido nitroso reductasa de 2 ciclocastos de *Achromobacter*
- Plantilla 2: estructura cristalina de *Pseudomonas nautica*
- Plantilla 3: óxido nitroso reductasa de *Shewanella denitrificans*
- Plantilla 4: estructura cristalina de *P. Denitrificans*
- Plantilla 5: óxido nitroso reductasa de *Pseudomonas stutzeri*, en formato cristalino con dos sustratos
- Plantilla 6: dominio N-terminal de óxido nitroso reductasa
- Plantilla 7: dominio N-terminal de óxido nitroso reductasa

A partir de la plantilla 1 se modela la estructura tridimensional de la proteína, con un 100% de confianza y una cobertura del 91%. La información proporcionada es la que aparece en la Figura 16.

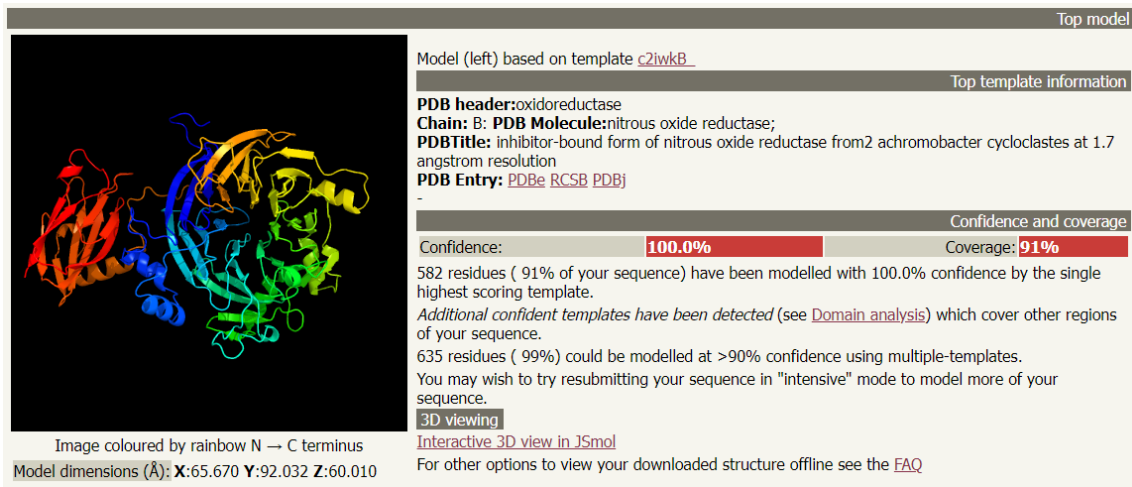


Figura 16. Resultados obtenidos de la modelización para predecir la estructura proteica de N₂OR. Este ha sido seleccionado como el mejor modelo.

Existe una herramienta en Phyre2 para analizar la secuencia proteica usada y el modelo de la proteína generado.

De todas las opciones disponibles, se usan tres: Alignment confidence, ProQ2 Quality Assessment [23] y Mutational sensitivity.

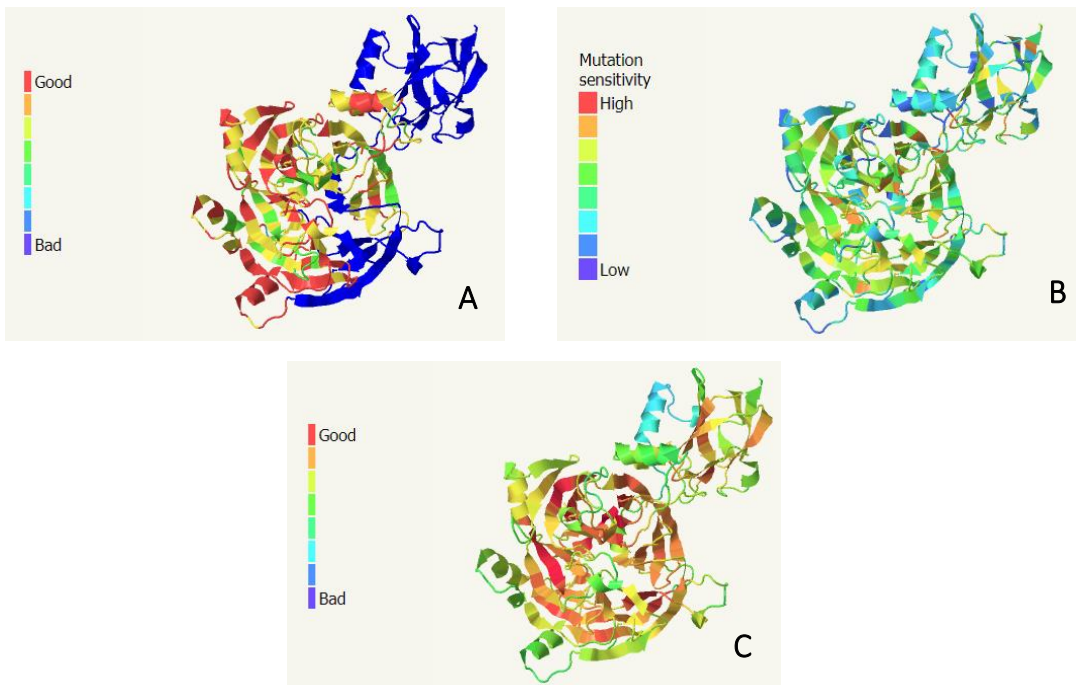


Figura 17. La figura A muestra la confianza del alineamiento, es decir, la fiabilidad de la alineación de la secuencia query. La figura B muestra la sensibilidad mutacional, que sirve para predecir si las mutaciones en la proteína pueden tener un efecto funcional o fenotípico. La figura C es la evaluación de la calidad ProQ2, un algoritmo de evaluación de la calidad del modelo.

4.5.2 Perfil para el Clado II

Para la creación del perfil proteico del Clado II se ha introducido la secuencia de *Rhodobacteraceae bacterium*.

El formato de los resultados obtenidos es el mismo que en el subapartado anterior. Se presentan las 7 plantillas con un 100% de confianza, aunque con un rango de identidad más amplio, del 58% al 75%. Los resultados se pueden visualizar en la Figura 18.

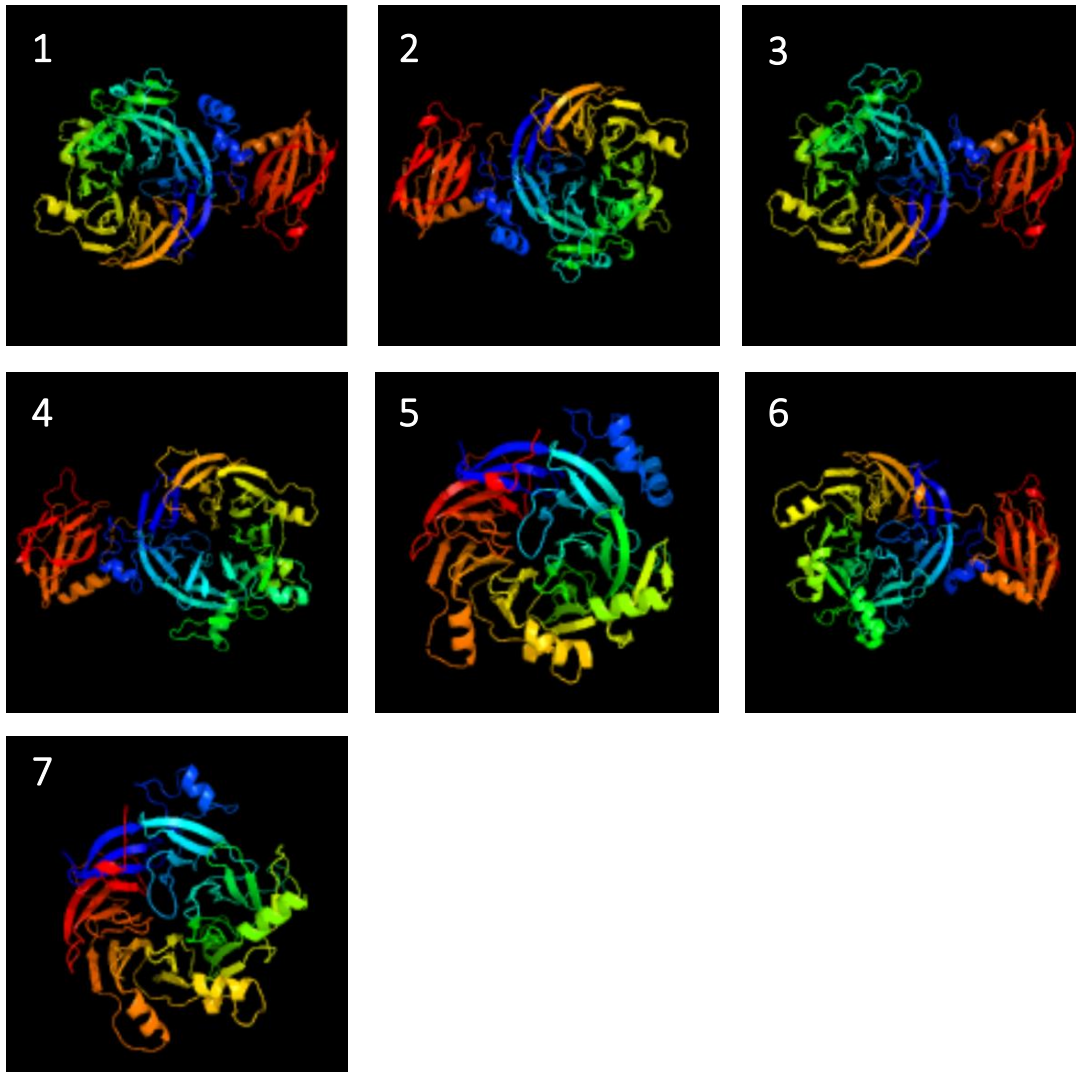


Figura 18. Plantillas con un 100% de confianza obtenidas durante el proceso de modelado por homología del perfil proteico para *Rhodobacteraceae bacterium* (Clado II).

Todas las plantillas anteriores han sido satisfactoriamente detectadas como molécula PDB óxido nítrico reductasa. Otra información adicional para cada una:

- Plantilla 1: estructura cristalina de *P. Denitrificans*
- Plantilla 2: forma unida al inhibidor de óxido nítrico reductasa de 2 ciclocastos de *Achromobacter*
- Plantilla 3: estructura cristalina de *Pseudomonas nautica*
- Plantilla 4: óxido nítrico reductasa de *Shewanella denitrificans*

- Plantilla 5: dominio N-terminal de óxido nítrico reductasa
- Plantilla 6: óxido nítrico reductasa de *Pseudomonas stutzeri*, en formato cristalino con dos sustratos
- Plantilla 7: dominio N-terminal de óxido nítrico reductasa

A partir de la plantilla 1 se modela la estructura tridimensional de la proteína, con un 100% de confianza y una cobertura del 99%. La información proporcionada es la que aparece en la Figura 19.

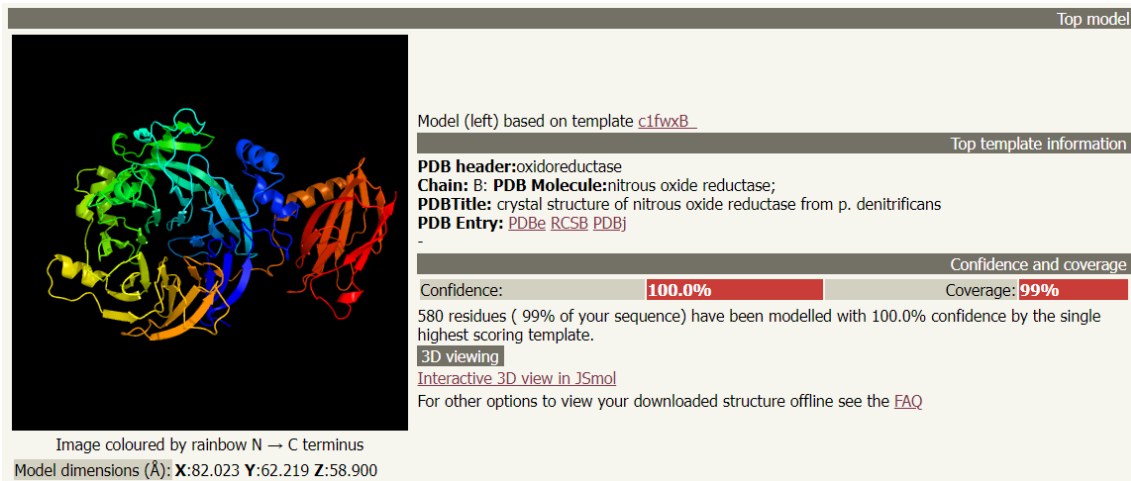


Figura 19. Resultados obtenidos de la modelización para predecir la estructura proteica de N₂OR. Este ha sido seleccionado como el mejor modelo.

Los resultados de los análisis Alignment confidence, ProQ2 Quality Assessment y Mutational sensitivity son:

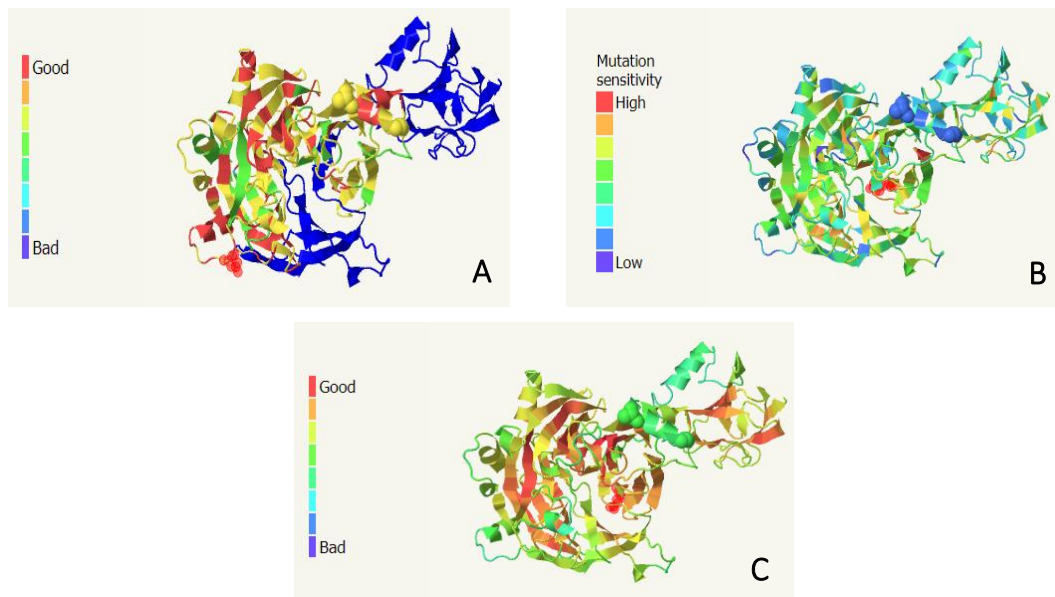


Figura 20. Representación de la confianza del alineamiento (A), la sensibilidad mutacional (B) y la evaluación de la calidad (C).

5 Discusión

La preparación de la base de datos usando una base de datos pública como es la del NCBI ha permitido recopilar convenientemente un conjunto de secuencias usando como base tan sólo una secuencia *query*. Con los resultados que se han ido encontrando a lo largo del trabajo, se corrobora la buena ejecución de ello.

Con el alineamiento múltiple de las secuencias se observa que hay una gran parte de la secuencia, sobre todo la región central, bien conservada entre los distintos organismos. Y según la señal filogenética obtenida, este alineamiento no ha sido producido por azar sino por la similitud encontrada entre los homólogos.

En el árbol filogenético se obtiene una clara diferenciación entre los dos Clados, de la cual se ha hablado extensamente a lo largo de trabajo. Lo que sí parece destacable es la menor cantidad de organismos pertenecientes al Clado II, hecho que podría deberse a que estos organismos no han sido históricamente tan estudiados [12], pero también a que los aminoácidos de las secuencias entre los dos clados difieren ligeramente [24]. Probablemente el haber usado una secuencia *query* de un organismo del Clado I haya determinado la baja coincidencia con organismos del Clado II.

De los perfiles proteicos elaborados podemos observar como la confianza e identidad son altos, por lo que parecen ser unos modelos robustos. De hecho, los resultados de la evaluación de calidad ProQ2 reafirman esa idea, aunque sí es cierto que la calidad del modelo es más alta en las regiones centrales de la estructura proteica. Tiene sentido, ya que tanto en el alineamiento múltiple realizado con MEGA como en la representación de la confianza del alineamiento de Phyre2 se ve claramente como son esas regiones centrales las que se conservan mejor. En cambio, los extremos N-terminales presentan más variaciones entre organismos.

Tal y como se ha dicho antes, las secuencias de los dos Clados tienen diferencias entre ellas aunque las dos sean capaces de catalizar la misma reacción. Si se comparan los resultados de la sensibilidad mutacional para el Clado I y II se observa como la mayor sensibilidad se encuentra en la región central y no en los extremos. Posiblemente sea gracias a esto que ambos Clados conserven la misma función reductora a pesar de sus diferencias.

En cuanto a la estructura modelada, podemos afirmar que la proteína tiene dos dominios: el N-terminal 7-bladed beta-propeller que contiene el centro catalítico de CuZ, y el C-terminal con un pliegue de cupredoxina. Los resultados coinciden con otros estudios previos [25].

6 Conclusiones

Los resultados obtenidos en el estudio coinciden y complementan estudios previos realizados en materia de filogenia de la óxido nitroso reductasa. Se concluye, por lo tanto, que es posible ampliar los conocimientos sobre dicha enzima usando bases de datos públicas y programas de código abierto o gratuitos.

Los objetivos planteados para el trabajo se han cumplido en su totalidad, aunque en algunos casos la metodología pensada inicialmente ha sido cambiada para adaptarse a las necesidades que han ido surgiendo a lo largo del estudio. Los principales cambios han sido en cuanto a qué software usar para cada tarea. Queda claro que hay una gran variedad y cantidad de herramientas bioinformáticas robustas disponibles, como los programas, softwares y servidores web usados en este estudio.

Es necesario recalcar la importancia del estudio de los grupos taxonómicos poseedores del gen *nosZ*, en especial los del Clado II, así como su estructura y requerimientos ambientales, ya que son conocimientos útiles para una futura línea de investigación en la mitigación de las emisiones del gas de efecto invernadero N₂O.

7 Glosario

- Alineamiento múltiple: Resultado del alineamiento de tres o más secuencias biológicas, ya sean secuencias de DNA, RNA o aminoácidos. Es un método usado para comparar secuencias.
- Bootstrapping: Metodología estadística basada en el remuestreo mediante la reposición aleatoria de los datos partiendo de una muestra inicial. Es útil para disminuir el sesgo.
- FASTA: formato de fichero plano usando para la representación de secuencias biológicas.
- HMM: Siglas de Hidden Markov Model, un modelo estadístico comúnmente usado para la modelación de secuencias biológicas.
- Indel: acrónimo de inserción y deleción, que hace referencia a mutación puntual de nucleótidos en una secuencia.
- Homodímero: Unión de dos moléculas proteicas iguales.
- Matriz PSSM: Siglas de Position-Specific Score Matrix, una matriz para la representación de motivos en secuencias biológicas.
- Metagenómica: Estudio de los genomas de poblaciones de distintos organismos microbianos
- Metaloenzimas: Proteína que contiene un ion metálico en su estructura.
- NCBI: Siglas del National Center for Biotechnology Information
- *nosZ*: gen que codifica la óxido nitroso reductasa
- N₂OR: óxido nitroso reductasa, enzima que cataliza la reducción de N₂O a N₂ durante el proceso de desnitrificación.
- Secuencias homólogas: Se dice de dos o más secuencias de DNA, RNA o proteínas que son similares porque tienen un mismo antepasado evolutivo.

8 Bibliografía

1. Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389-402. doi: 10.1093/nar/25.17.3389.
2. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. Disponible en: <https://www.r-project.org/>
3. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol.* 2018;35(6):1547-9. doi: 10.1093/molbev/msy096.
4. Schmidt HA, Strimmer K, Vingron M, Von Haeseler A. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics.* 2002;18(3):502-4. doi: 10.1093/bioinformatics/18.3.502.
5. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc.* 2015;10(6):845-58. doi: 10.1038/nprot.2015.053
6. Thompson RL, Lassaletta L, Patra PK, Wilson C, Wells KC, Gressent A, et al. Acceleration of global N₂O emissions seen from two decades of atmospheric inversion. *Nat Clim Chang.* 2019;9:993-998. doi: 10.1038/s41558-019-0613-7
7. Intergovernmental Panel on Climate Change (IPCC). Atmospheric Chemistry and Greenhouse Gase. En: Ehhalt D, Lead MP, Dentener F, Derwent R, Dlugokencky E, Holland E, et al., editores. *Climate Change: The Scientific Basis.* University of Cambridge; 2001. p. 241-80.
8. Wrage N, Lauf J, Del Prado A, Pinto M, Pietrzak S, Yamulki S, et al. Distinguishing sources of N₂O in European grasslands by stable isotope analysis. *Rapid Commun Mass Spectrom.* 2004;18(11):1201-7. doi: 10.1002/rcm.1461.
9. Alvarez L, Bricio C, Blesa A, Hidalgo A, Berenguer J. Transferable Denitrification Capability of *Thermus thermophilus*. *Appl Environ Microbiol.* 2014;80(1):19. doi: 10.1128/AEM.02594-13
10. Fujita K, Chan JM, Bollinger JA, Alvarez ML, Dooley DM. Anaerobic purification, characterization and preliminary mechanistic study of recombinant nitrous oxide reductase from *Achromobacter cycloclastes*. *J Inorg Biochem.* 2007;101(11-12):1836-44. doi: 10.1016/j.jinorgbio.2007.06.029.

11. Shan J, Sanford RA, Chee-Sanford J, Ooi SK, Löffler FE, Konstantinidis KT, et al. Beyond denitrification: The role of microbial diversity in controlling nitrous oxide reduction and soil nitrous oxide emissions. *Glob Chang Biol.* 2021;27(12):2669-83. doi: 10.1111/gcb.15545
12. Chee-Sanford JC, Connor L, Krichels A, Yang WH, Sanford RA. Hierarchical detection of diverse Clade II (atypical) *nosZ* genes using new primer sets for classical- and multiplex PCR array applications. *J Microbiol Methods.* 2020;172:105908. doi: 10.1016/j.mimet.2020.105908
13. Jones CM, Graf DRH, Bru D, Philippot L, Hallin S. The unaccounted yet abundant nitrous oxide-reducing microbial community: a potential nitrous oxide sink. *ISME J* 2013 72. 2012;7(2):417-26. doi: 10.1038/ismej.2012.125
14. Itakura M, Uchida Y, Akiyama H, Hoshino YT, Shimomura Y, Morimoto S, et al. Mitigation of nitrous oxide emissions from soils by *Bradyrhizobium japonicum* inoculation. *Nat Clim Chang.* 2012;3(3):208-12. doi: 10.1038/nclimate1734
15. Shaaban M, Wu Y, Khalid MS, Peng Q an, Xu X, Wu L, et al. Reduction in soil N₂O emissions by pH manipulation and enhanced *nosZ* gene transcription under different water regimes. *Environ Pollut.* 2018;235:625-31.
16. Aamer M, Chattha MB, Mahmood A, Naqve M, Hassan U, Shaaban M, et al. Rice Residue-Based Biochar Mitigates N₂O Emission from Acid Red Soil. *Agronomy.* 2021;1-13. doi: 10.1016/j.envpol.2017.12.066
17. Hulse CL, Averill BA. Isolation of a high specific activity pink, monomeric nitrous oxide reductase from *Achromobacter cycloclastes*. *Biochem Biophys Res Commun.* 1990;166(2):729-35. doi: 10.1016/0006-291x(90)90870-s
18. McGuirl MA, Nelson LK, Bollinger JA, Chan YK, Dooley DM. The *nos* (nitrous oxide reductase) gene cluster from the soil bacterium *Achromobacter cycloclastes*: Cloning, sequence analysis, and expression. *J Inorg Biochem.* 1998;70(3-4):155-69. doi: 10.1016/s0162-0134(98)10001-6
19. Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, Ahmad S, et al. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* 2021;49(D1):D480-9. doi: 10.1093/nar/gkaa1100
20. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 2004;32(5):1792-7. doi: 10.1093/nar/gkh340
21. Martínez A, González A. Aplicaciones de la bioinformática en la elaboración de filogenias moleculares. Departamento de Biología Celular y Molecular Universidade da Coruña;
22. Luo A, Qiao H, Zhang Y, Shi W, Ho SY, Xu W, et al. Performance of criteria for selecting evolutionary models in phylogenetics: A comprehensive study

- based on simulated datasets. *BMC Evol Biol.* 2010;10(1):1-13. doi: 10.1186/1471-2148-10-242
23. Ray A, Lindahl E, Wallner B. Improved model quality assessment using ProQ2. *BMC Bioinformatics.* 2012;13(1):1-12. doi: 10.1186/1471-2105-13-224
 24. Yoon S, Nissen S, Park D, Sanford RA, Löffler FE. Nitrous Oxide Reduction Kinetics Distinguish Bacteria Harboring Clade I NosZ from Those Harboring Clade II NosZ. *Appl Environ Microbiol.* 2016;82(13):3793. doi: 10.1128/AEM.00409-16
 25. Brown K, Tegoni M, Prudêncio M, Pereira AS, Besson S, Moura JJ, et al. A novel type of catalytic copper cluster in nitrous oxide reductase. *Nat Struct Biol.* 2000;7(3):191-5. doi: 10.1038/73288

9 Anexos

Anexo I

```
>sp|P94127|NOSZ_ACHCY Nitrous-oxide reductase OS=Achromobacter
cycloclastes OX=223 GN=nosZ PE=1 SV=2
MESKEHKGLSRRALFSATAGSAILAGTVGPAALSLGAAGLATPARAATGADGSVAPGKLDD
YYGFWSSGQGTGEMRILGIPSMRELMRVPVFNRCATGWGQTNESIRIHQRTMTEKTKKQL
AANGKKIHDNGDLHHVHMSFTDGKYDGRYLFMNDKANTRVARVRCDVMKTDAILIPNA
KGIHGMRPQKWPRSNYVFCNGEDEAPLVNDGSTMTDVATYVNIIFTAVDADKWEVAWQ
VKVSGNLDNCDADYEGKWFSTSYNSEMGMTLEEMTKSEMDHVVFVFNIAEIEKAIKAGQ
YEEINGVKVVDGRKEAKSLFTRYIPIANNPHGCNMAPDRKHLCVAGKLSPTVTVLDVTKFDA
LFYDNAEPRSAVVAEPELGLGPLHTAFDGRGNAYTSLFLDSQVVKWNIDEAIRAYAGEKINPI
KDKLDVQYQPGHLKTVMGETLDAANDWLVLCLCKFSKDRFLNVGPKPENDQLIDISGDKM
VLVHDGPTFAEPHDAIAVSPSILPNIRSVWDRKDPLWAETRKQAEAEVDIDEWTEAVIRD
GNKVRVYMTSVAPSFSQPSFTVKEGDEVTVIVTNLDEIDDLTHGFTMGNHGVAMEVGPQ
QTSSVTFVAANPGVYWYYCQWFCHALHMEMRGRMFVEPKGA
```

Anexo II

Código de R para el filtrado, limpieza y preparación de la base de datos.

```
# Script para calcular las secuencias -----
# Cargar datos
listado = read.csv("../datos/alignment_descr.csv", header = T)

# Seleccionar columnas y renombrarlas
listado = listado[,c(1, 2)]
names(listado) = c('descripcion', 'especie')

# Limpieza de la columna descripcion
listado$proteina = sub("\\[.+", "", listado$descripcion)
listado$proteina = as.factor(trimws(listado$proteina, 'r'))

# Recuento del tipo de columnas
summary(listado$proteina)

# Seleccion de los tipos de proteínas
proteinas = c('nitrous-oxide reductase', 'Nitrous-oxide reductase', 'nitrous-oxide reductase NosZ',
              'nitrous oxide reductase', 'nosZ')

# Aquí hay solamente las secuencias de interés
listado_filtro = listado[listado$proteina %in% proteinas,]
listado_filtro = listado_filtro[-c(3)]

# Script para filtrar los datos -----
# Cargar los datos
seq = Biostrings::readBStringSet("../datos/fasta_seq.txt")

seq = data.frame(seq[,1])
names(seq)[1] = 'secuencia'
seq$nombre = rownames(seq)
rownames(seq) = NULL

# Se fusionan con el listado total
seq = cbind(seq, listado)

seq = seq[seq$proteina %in% proteinas, ]

seq = seq[,c(1, 4)]

# Se guardan los resultados
seqinr::write.fasta(as.list(seq$secuencia), names = seq$especie, file = 'seq.fasta')
```


Anexo III

>MBR27429.1 nitrous-oxide reductase [Rhodobacteraceae bacterium]
MDEYYGFWSSGQSGEMRILGVPSMRELMRVPVFNRC SATGWGQTNESLKILTQDMLPA
TKDYLA AQGKTVHDNGDLHHVHMSFTEGKYDGRYLFMNDKANTRIARVRC DVMKPDRII
DIPNAKAIHGLRPQKWPRNTNYVFANG EDEVPMVNDGKTLDDPSKYVNFYTA INADEMTV
AWQVMVSGNLDNTDADYDGKYAFSTSYNSEMGMTLP EMMEAEMDHVVIFNLAEIEKGI
EAGDYIELNGVKVIDGRKGSNKYTRYVPIPN SPHGINMAPDRKHLCVAGKLSPTVSVLDVT
KFDALFEDASLDPRS VVAEPELGLGPLHTCFD GKGNAFTTLFLDSQVMMWNLEKAI AAYN
GEAVDPILT KKVHYPGHNSTTMGETLEADGK WYLSMNKFSKDRFLN VGPLKPENEQLF
SIDGEELSLVHDGPTFAEPHDSILVHRSV VNPKSVWDRQDPMWADARAQAEADGVDLDD
WQDTIIRDGT KVRVYMTSQAPEFSLSQFTVK QGDEVTVYVTNLDEIDDLTHGFTLNNHGVA
MEVAPQATASVTFTA AKPGVYWYYCQWFCHALH MEMRGRMFVEPTGA

>WP_192321664.1 nitrous-oxide reductase [Aminobacter sp. SR38]
MSDERINRLNRRQLLGT SVAAGATAVSGALTL AGGAVSPAAAQSSSAAPSGEGGSYEVKPG
ELDEYYVFFSSGQSGEVRIVGAPSMREMMRIPVFNRC SATGWGQTNESRKVLTEGLLPNT
KEFLKDKGGIYLNGLHHPHSFTDGT YDGRYLFANDKANTRVCRIRLDVMKCDKIIQLPNQ
HTVHGLRVQKYPKTGYVFCNGEDRVPIPN DGKVLDDHKQYHAIFTAVDGETMKVAWQV
MVDGNLDNVDADYQ GKYCFATCYNSEEGVNLAE MMEKEQDWVIFNLKRIEDAVAKGD
YKEINGVPVIDGRK GSPYTRYVVPN SPHGINTAPDGIHV VANGKLSPTVTVFDVRKFDEL F
DDKIQRD TVVAEPELGLGPLHTAYD GKGNAFTTLFIDSQICKWNLDDAKRAF KGEAVDPIR
QKLDVHYQPGHNHTSMGQTK EADGKWLISLNKFSKDRYLN VGPLKPENDQLIDISGDQM
VLVHDNPTFAEPH DATLVHRSKVN PVHVWNRED PFFADAVKQAKADGIDLMSAAEVIRD
GNKVRVYMTS AAPAFSLETFSVKQGDEVTVYVTN IDEVEDLTHGFSIINYGINMEVAPQATA
SVTFTASQPGVYWYYCSW FCHAMHMEMKGRMFVEPQGA

Anexo IV

Listado de las especies y su descripción incluidas en la base de datos

especie	proteina
Ruegeria lacuscaerulensis ITI-1157	nitrous-oxide reductase
Rhodobacteraceae bacterium 4F10	nitrous-oxide reductase
Rhodobacteraceae bacterium (ex Bugula neritina AB1)	nitrous-oxide reductase
Rhodobacterales bacterium 56_14_T64	nitrous-oxide reductase
uncultured bacterium	nitrous-oxide reductase
Alphaproteobacteria bacterium HGW-Alphaproteobacteria-5	nitrous-oxide reductase
Hoeflea sp. BRH_c9	nitrous-oxide reductase
Roseovarius indicus	nitrous-oxide reductase
Rhodobacteraceae bacterium	nitrous-oxide reductase
Pleomorphomonas sp. SG806	nitrous-oxide reductase
Proteobacteria bacterium	nosZ
Phyllobacteriaceae bacterium	nitrous-oxide reductase
Aminobacter	nitrous-oxide reductase
Rhizobium sp. Root1204	nitrous-oxide reductase
Roseovarius sp.	nitrous-oxide reductase
Aminobacter sp. SR38	nitrous-oxide reductase
Stappia sp. TSB10P1A	nitrous-oxide reductase
Rhodobacterales bacterium RIFCSPHIGHO2_02_FULL_62_130	nitrous-oxide reductase
Stappia sp. P2PMeth1	nitrous-oxide reductase
uncultured bacterium	nitrous-oxide reductase
Stappia sp. TSB10GB4	nitrous-oxide reductase
Aminobacter lissarensis	nitrous-oxide reductase
uncultured bacterium	nitrous-oxide reductase
Rhodobacteraceae bacterium	nitrous-oxide reductase
Ensifer mexicanus	nitrous-oxide reductase
Thalassospira sp. Nap_22	nitrous-oxide reductase
Proteobacteria bacterium	nitrous-oxide reductase
Kaistia sp. SCN 65-12	nitrous-oxide reductase
Alphaproteobacteria bacterium BRH_c36	nitrous-oxide reductase
Thalassospira tepidiphila	nitrous-oxide reductase
Rhodobacteraceae bacterium GWE1_64_9	nitrous-oxide reductase
Rhizobium sp. 63-7	nitrous-oxide reductase
Pseudaminobacter arsenicus	nitrous-oxide reductase
Neorhizobium sp. T786	nitrous-oxide reductase
Ensifer sp. M14	nitrous-oxide reductase
Sinorhizobium fredii	nitrous-oxide reductase

Brucellaceae	nitrous-oxide reductase
Achromobacter cycloclastes	nitrous-oxide reductase
unclassified Pseudaminobacter	nitrous-oxide reductase
Rhizobium sp. Khangiran2	nitrous-oxide reductase
Ciceribacter sp. F8825	nitrous-oxide reductase
Rhodobacter sp. CACIA14H1	nitrous-oxide reductase
Ensifer psoraleae	nitrous-oxide reductase
Brucella/Ochrobactrum group	nitrous-oxide reductase
Brucella/Ochrobactrum group	nitrous-oxide reductase
Mesorhizobium sp. B2-6-4	nitrous-oxide reductase
Falsochrobactrum shanghaiense	nitrous-oxide reductase
Pseudochrobactrum sp. AO18b	nitrous-oxide reductase
unclassified Ensifer	nitrous-oxide reductase
Brucella oryzae	nitrous-oxide reductase
Cereibacter sphaeroides f. sp. denitrificans	nitrous oxide reductase
Rhizobium sp.	nitrous-oxide reductase
Thalassospira sp.	nitrous-oxide reductase
Ensifer adhaerens	nitrous-oxide reductase
Rhizobiaceae bacterium	nitrous-oxide reductase
Rhizobium sp.	nitrous-oxide reductase
Mesorhizobium sp. YM1C-6-2	nitrous-oxide reductase
Rhodobacteraceae bacterium HLUCCA12	nitrous-oxide reductase NosZ
Brucella pecoris	nitrous-oxide reductase
Rhizobium	nitrous-oxide reductase
Agrobacterium sp. a22-2	nitrous-oxide reductase
Rhizobiales bacterium NRL2	nitrous-oxide reductase
unclassified Mesorhizobium	nitrous-oxide reductase
Roseovarius sp.	nitrous-oxide reductase
Microvirga massiliensis	nitrous-oxide reductase
Pseudochrobactrum sp. B5	nitrous-oxide reductase
Roseovarius sp.	nitrous-oxide reductase
Pseudorhizobium endolithicum	nitrous-oxide reductase
Paramesorhizobium deserti	nitrous-oxide reductase
Pseudovibrio sp. FO-BEG1	Nitrous-oxide reductase
Hyphomicrobium denitrificans	nitrous-oxide reductase
Hyphomicrobium denitrificans	nitrous-oxide reductase
Pseudoceanicola sp.	nitrous-oxide reductase
Brucella anthropi	nitrous-oxide reductase
Mesorhizobium sp.	nitrous-oxide reductase
Neorhizobium	nitrous-oxide reductase
Bradyrhizobium sp.	nitrous-oxide reductase
Neorhizobium sp. R1-B	nitrous-oxide reductase

Brucella anthropi	nitrous-oxide reductase
Rhizobium/Agrobacterium group	nitrous-oxide reductase
Brucella/Ochrobactrum group	nitrous-oxide reductase
Rhizobium sp.	nitrous-oxide reductase
Brucella anthropi	nitrous-oxide reductase
Stappia indica	nitrous-oxide reductase
Brucella/Ochrobactrum group	nitrous-oxide reductase
Thalassospira sp.	nitrous-oxide reductase
Aureimonas glaciei	nitrous-oxide reductase
Brucella cytisi	nitrous-oxide reductase
Brucella tritici	nitrous-oxide reductase
Arenimonas sp.	nitrous-oxide reductase
Rhizobium naphthalenivorans	nitrous-oxide reductase
Aminobacter ciceronei	nitrous-oxide reductase
Bradyrhizobium japonicum	nosZ
Ensifer sp. BR816	nitrous-oxide reductase
Brucella	nitrous-oxide reductase
Pseudochelatococcus lubricantis	nitrous-oxide reductase
Brucella tritici	nitrous-oxide reductase
Brucella lupini	nitrous-oxide reductase
Brucella anthropi	nitrous-oxide reductase
Brucella/Ochrobactrum group	nitrous-oxide reductase
Pseudochrobactrum	nitrous-oxide reductase
Brucella anthropi	nitrous-oxide reductase
Alphaproteobacteria bacterium HGW- Alphaproteobacteria-1	nitrous-oxide reductase
Brucella anthropi	nitrous oxide reductase
Fuscovulum blasticum	nitrous-oxide reductase
Pseudorhizobium halotolerans	nitrous-oxide reductase
Rhodobacterales bacterium 59_46_T64	nitrous-oxide reductase
Ciceribacter thiooxidans	nitrous-oxide reductase
Rhizobium sp.	nitrous-oxide reductase
Pseudochrobactrum sp. Wa41.01b-1	nitrous-oxide reductase
Mesorhizobium sp. Root157	nitrous-oxide reductase
Rhodobacteraceae bacterium	nitrous-oxide reductase
Ochrobactrum sp. MYb71	nitrous-oxide reductase
Brucella rhizosphaerae	nitrous-oxide reductase
Ensifer alkalisoli	nitrous-oxide reductase
Pseudochrobactrum asaccharolyticum	nitrous-oxide reductase
Rhizobium sp. AU243	nitrous-oxide reductase
Brucella grignonensis	nitrous-oxide reductase
Brucella neotomae	nitrous-oxide reductase

unclassified Mesorhizobium	nitrous-oxide reductase
Mesorhizobium sp.	nitrous-oxide reductase
Sinorhizobium sp. M4_45	nitrous-oxide reductase
Rhizobium selenitireducens	nitrous-oxide reductase
Brucella thiophenivorans	nitrous-oxide reductase
Ochrobactrum sp. J50	nitrous-oxide reductase
Brucella pseudintermedia	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Brucella oryzae	nitrous-oxide reductase
Microvirga rosea	nitrous-oxide reductase
Brucella anthropi	nitrous-oxide reductase
Brucella daejeonensis	nitrous-oxide reductase
Bosea thiooxidans	nitrous-oxide reductase
Ochrobactrum sp. MR28	nitrous-oxide reductase
Ochrobactrum sp. SFR4	nitrous-oxide reductase
Mesorhizobium	nitrous-oxide reductase
Brucella grignonensis	nitrous-oxide reductase
Rhizobium sp. H41	nitrous-oxide reductase
Brucella pituitosa	nitrous-oxide reductase
Rhizobiales bacterium 32-66-11	nitrous-oxide reductase
Mesorhizobium sp.	nitrous-oxide reductase
Mesorhizobium sp.	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Ochrobactrum sp. POC9	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Roseovarius sp. BRH_c41	nitrous-oxide reductase
Mesorhizobium sp.	nitrous-oxide reductase
Mesorhizobium sp. M00.F.Ca.ET.216.01.1.1	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Aurantimonas sp. SW136	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Mesorhizobium sp. B2-3-15	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
unclassified Mesorhizobium	nitrous-oxide reductase
Alphaproteobacteria bacterium HGW-Alphaproteobacteria-6	nitrous-oxide reductase

Rhizobiaceae bacterium	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Falsochrobactrum sp. TDYN1	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
uncultured bacterium	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Hyphomicrobiales bacterium	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Corticibacterium populi	nitrous-oxide reductase
Sinorhizobium sp. NP1	nitrous oxide reductase
Rhodobacterales bacterium 32-67-9	nitrous-oxide reductase
Brucella sp. 6810	nitrous-oxide reductase
Brucella	nitrous-oxide reductase
Bosea sp. CS1GBMeth4	nitrous-oxide reductase
Cellvibrionaceae bacterium	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Brucella sp. 458	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Bosea sp.	nitrous-oxide reductase
Rhodobacteraceae bacterium	nitrous-oxide reductase
Mesorhizobium sp.	nitrous-oxide reductase
Sinorhizobium meliloti	nitrous-oxide reductase
Ochrobactrum sp. CNPMS2088	nitrous-oxide reductase
Brucella sp. 2280	nitrous-oxide reductase
Mesorhizobium sp.	nitrous-oxide reductase
Bosea sp.	nitrous-oxide reductase
Brucella sp. NVSL 07-0026	nitrous-oxide reductase
Mesorhizobium sp. 3P27G6	nitrous-oxide reductase
Mesorhizobium composti	nitrous-oxide reductase
Brucella sp. BO2	nitrous-oxide reductase
Brucella vulpis	nitrous-oxide reductase
Bosea sp. BIWAKO-01	nitrous-oxide reductase
Brucella	nitrous-oxide reductase
Mesorhizobium sp.	nitrous-oxide reductase
Brucella inopinata	nitrous-oxide reductase
Mesorhizobium tianshanense	nitrous-oxide reductase
unclassified Mesorhizobium	nitrous-oxide reductase
Rhizobium leguminosarum	nitrous-oxide reductase
Rhizobium anhuiense	nitrous-oxide reductase
Rhizobium leguminosarum	nitrous-oxide reductase
Mesorhizobium loti	nitrous-oxide reductase

Brucella abortus	nitrous-oxide reductase
Hyphomicrobiales bacterium	nitrous-oxide reductase
Brucella suis	nitrous-oxide reductase
unclassified Brucella	nitrous-oxide reductase
Brucella abortus	nitrous-oxide reductase
Microvirga sp. VF16	nitrous-oxide reductase
Brucella suis	nitrous-oxide reductase
Bradyrhizobiaceae bacterium	nitrous-oxide reductase
Brucella	nitrous-oxide reductase
Brucella anthropi	nitrous-oxide reductase
Brucella sp. 63/311	nitrous-oxide reductase
Brucella ovis	nitrous-oxide reductase
Microvirga lupini	nitrous-oxide reductase
Brucella abortus	nitrous-oxide reductase
Brucella microti	nitrous-oxide reductase
Rhizobium gallicum	nitrous-oxide reductase
Rhizobium sp. BK619	nitrous-oxide reductase
Thiotrichaceae bacterium	nitrous-oxide reductase
Brucella sp. BCCN 84-3	nitrous-oxide reductase
Skermanella aerolata	nitrous-oxide reductase

Anexo V

Script usado para la comprobación de la taxonomía del Clado I

```
# Función para encontrar la taxonomía
datos = data.frame()

for (especies in clado1$especie) {
  info = taxonomy(especies, output = 'classification')

  if (!is.na(info)[1] == TRUE) {

    phylum = info[info$rank == 'phylum', ][,1]
    class = info[info$rank == 'class', ][,1]
    order = info[info$rank == 'order', ][,1]
    family = info[info$rank == 'family', ][,1]
    specie = as.character(especies)

    phylum = ifelse(length(phylum) == 0, NA, phylum)
    class = ifelse(length(class) == 0, NA, class)
    order = ifelse(length(order) == 0, NA, order)
    family = ifelse(length(family) == 0, NA, family)
    specie = ifelse(length(specie) == 0, NA, specie)

    resultado = data.frame(t(c(phylum, class, order, family, specie)))

    datos = data.frame(rbind(datos, resultado), stringsAsFactors = F)

    message('[CORRECTE]: ', especies)

  } else {
    message('[ERROR]: ', especies)
  }
  rm(info, phylum, class, order, family, specie, resultado)
}
```