Universitat Oberta de Catalunya

uoc.edu

Determinación del perfil de la enzima óxido nítrico reductasa

Estudiante: Carlos Navarro Marcos
Máster en Bioinformática y Bioestadística
Bioinformática i Bioestadística Área 4

Consultora: Paloma María Pizarro Tobías
Nombre Profesor/a responsable de la asignatura: Antoni Pérez Navarro

Entrega: 24/12/2021

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Determinación del perfil de la enzima óxido nítrico reductasa*
Nombre del autor:	*Carlos Navarro Marcos*
Nombre del consultor/a:	*Paloma María Pizarro Tobías*
Nombre del PRA:	*Antoni Pérez Navarro*
Fecha de entrega (mm/aaaa):	12/2021
Titulación:	*Máster en Bioinformática y Bioestadística*
Área del Trabajo Final:	*Bioinformática i Bioestadística Área 4*
Idioma del trabajo:	Castellano
Número de créditos:	15
Palabras clave	*Óxido nítrico reductasa, nor, filogenia*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

El objetivo del proyecto fue establecer el perfil (o perfiles) de la enzima óxido nítrico reductasa (*nor*) ya que los perfiles disponibles actualmente no son adecuados, y la obtención de perfiles con capacidad discriminante permitirían generar nuevos conocimientos para identificar la enzima en genomas ya secuenciados o en secuencias obtenidas *de novo*, obtener nuevas cepas bacterianas con propiedades interesantes para su uso en agricultura y medio ambiente, o para estudios de biodiversidad, que darán lugar a nuevos conocimientos en relación a la *nor*, y permitirá a la comunidad científica adquirir una mayor comprensión de los mecanismos de desnitrificación, lo que en último lugar permitirá desarrollar estrategias viables para reducir las emisiones de N_2O. El enfoque y metodología propuestos en este proyecto para cumplir con los objetivos del proyecto, está basado en los trabajos realizados Udaondo y col. y éstos se basan en aprovechar la gran cantidad de información disponible existente (secuencias) para establecer el perfil de la enzima mediante la construcción o elaboración de perfiles generalizados PROSITE. En el árbol filogenético se observaron tres tipos de *nor* (cNOR, vNOR, qNOR+qCuNOR) que se corresponden con la literatura. Se observó que los perfiles generalizados obtenidos para cada clase de *nor* tenían capacidad para identificar y discriminar entre tipos de *nor* en el genoma de un organismo e igualmente identificar la proteína en bases de datos se secuencias no anotadas. Sin embargo, no se

pudo realizar una validación más exhaustiva por lo que se considera necesario realizar más trabajos en un futuro.

Abstract (in English, 250 words or less):

The objective was to establish the profile (or profiles) of the enzyme nitric oxide reductase (*nor*) because the profiles currently available are not adequate, and obtaining profiles with discriminant capacity would allow to identify the enzyme in genomes already sequenced or in sequences obtained *de novo*, to obtain new bacterial strains with interesting properties for use in agriculture and the environment, or for biodiversity studies, which will give rise to new knowledge in relation to *nor*, and will allow the scientific community to acquire a better understanding of denitrification mechanisms, which will ultimately allow the development of viable strategies to reduce N_2O emissions. The approach and methodology proposed in this project to meet the project objectives is based on the work carried out by Udaondo et al. and these are based on analyzing the large amount of existing information available (sequences) to establish the profile of the enzyme through the construction or elaboration of generalized PROSITE profiles. In the phylogenetic tree obtained in this project, three types of *nor* were observed (cNOR, vNOR, qNOR + qCuNOR) that corresponded to the information available in the literature. It was observed that the generalized profiles obtained for each class of *nor* had the capacity to identify and discriminate between types of *nor* in the genome of an organism and could identify the protein in databases of non-annotated sequences. However, a more exhaustive validation could not be performed, so it is considered further studies are required.

Índice

1. Resumen **1**

2. Introducción **1**
2.1 Contexto y justificación del Trabajo 1
2.2 Objetivos del Trabajo 4
2.3 Enfoque y método seguido 5
2.4 Planificación del Trabajo 6
2.5 Breve sumario de productos obtenidos 8
2.6 Breve descripción de los otros capítulos de la memoria 9

3. Estado del arte **11**

4. Metodología **15**

5. Resultados **23**

6. Discusión **40**

7. Conclusiones **45**

8. Glosario **47**

9. Bibliografía **48**

10. Anexos **54**

Lista de figuras

Figura 2.1.1. Emisiones totales de N_2O.

Figura 2.1.2. Proceso de desnitrificación.

Figura 2.2.1. Esquema de la metodología propuesta en este Trabajo [18].

Figura 2.4.1. Diagrama de Gantt del Proyecto.

Figura 4.1. Base de datos *Uniprot*.

Figura 4.2. Secuencias (2208) abiertas en MEGA.

Figura 4.3. Secuencias similares (no se eliminan).

Figura 4.4. Secuencias divergentes (se eliminan).

Figura 4.5. Ejemplo para el criterio nº 2.

Figura 4.6. Oracle VM VirtualBox.

Figura 4.7. EMBOSS para la conversión de formatos.

Figura 4.8. Archivo en formato.msf.una vez asignados los pesos con el comando *pfw*

Figura 5.1. Esquema de las secuencias utilizadas.

Figura 5.2. Árbol filogenético obtenido para *nor* (*unrooted)*.

Figura 5.3. Árbol filogenético obtenido para *nor* (circular sin considerar las distancias*)*.

Figura 5.4. Árbol filogenético obtenido para *nor* (circular*)*.

Figura 5.5. Árbol filogenético obtenido para *nor* (*unrooted*) utilizando las secuencias con la parte final eliminada.

Figura 5.6. Árbol filogenético obtenido para *nor* (*unrooted*) con el parámetro *fast*.

Figura 6.1. Enzimas cNOR y qNOR.

Lista de tablas

Tabla 2.4.1. Dedicación a las Tareas del Proyecto.

Tabla 5.1. Resumen de las secuencias utilizadas.

Tabla 5.2. Resultados obtenidos. *iqtree -s Alineamiento_3.fas -m MF -safe.*

Tabla 5.3. Estadísticas del árbol utilizado para determinar el mejor modelo.

Tabla 5.4. Comparativa de los organismos de cada Grupo.

Tabla 5.5. Resultados obtenidos mediante *pfsearch* qNOR+qCuNOR vs *Swiss-Prot.*

Tabla 5.6. Resultados obtenidos mediante *pfsearch* cNOR vs *Swiss-Prot.*

Tabla 5.7. Resultados obtenidos mediante *pfsearch* vNOR vs *Swiss-Prot* (únicamente resultados con *Score* > 20).

Tabla 5.8. M*atches* y *scores* obtenidos al utilizar los diferentes perfiles en cada uno de los genomas descargados.

Tabla 5.9. M*atches* de proteínas hipotéticas para el perfil qNOR+qCuNOR.

Tabla 5.10. M*atches* de proteínas hipotéticas para el perfil cNOR.

Tabla 5.11. M*atches* de proteínas hipotéticas para el perfil vNOR.

Tabla 5.12. Proteínas identificadas en común.

1 Resumen

En los últimos años, el interés por el óxido nitroso (N_2O) ha crecido debido a que se trata de un gas con un potente efecto invernadero con una permanencia en la estratosfera de aproximadamente 150 años. Más de dos tercios del gas producido tiene su origen en los procesos de desnitrificación y nitrificación de algunas bacterias y hongos en los suelos. La producción de NO_2 ha crecido exponencialmente en el último siglo debido a la intensificación de la agricultura y a la mayor utilización de nitrógeno para fertilizar los cultivos. Por otro lado, la ganadería intensiva también contribuye en las emisiones totales de N_2O. En los procesos de desnitrificación intervienen cuatro enzimas para reducir el ion NO_{3-} a N_2, sin embargo, este proceso no siempre se completa o se realiza de una forma eficiente, lo que provoca emisiones de los productos intermedios, como el N_2O. Una de las enzimas que participan en la desnitrificación es la óxido nítrico reductasa (*nor*), responsable de uno de los procesos intermedios de la desnitrificación, y se encarga de convertir el óxido nítrico (NO) a N_2O. La información disponible en la literatura en relación al perfil de esta enzima puede no ser completa e igualmente, los perfiles de la óxido nítrico reductasa disponibles en las bases de datos a día de hoy no tienen poder discriminatorio y no son capaces de distinguir entre las clases de ésta. Por lo tanto, el **objetivo** de este proyecto consiste en establecer el perfil (o perfiles) de la enzima óxido nítrico reductasa (*nor*). El enfoque y metodología propuestos en este proyecto para cumplir con los objetivos del proyecto, está basado en los trabajos realizados Udaondo y col. y éstos se basan en aprovechar la gran cantidad de información disponible existente (secuencias) para establecer el perfil de la enzima mediante la construcción o elaboración de perfiles generalizados PROSITE. Para ello, en primer lugar, de obtuvieron un total de 1.543 secuencias de proteínas de la base de datos *Uniprot* anotadas con actividad oxido nítrico reductasa. A partir de estas secuencias, se elaboró un árbol filogenético para obtener un alineamiento múltiple para cada tipo o clase de *nor* para la elaboración de los perfiles generalizados mediante las herramientas *pftools*. En el árbol filogenético se observaron tres tipos de *nor* (cNOR, vNOR, qNOR+qCuNOR) que se corresponden con la literatura. Se observó que los perfiles generalizados obtenidos para cada clase de *nor* tenían capacidad para identificar y discriminar entre tipos de *nor* en el genoma de un organismo e igualmente identificar la proteína en bases de datos se secuencias no anotadas. Sin embargo, no se pudo realizar una validación de los perfiles más exhaustiva.

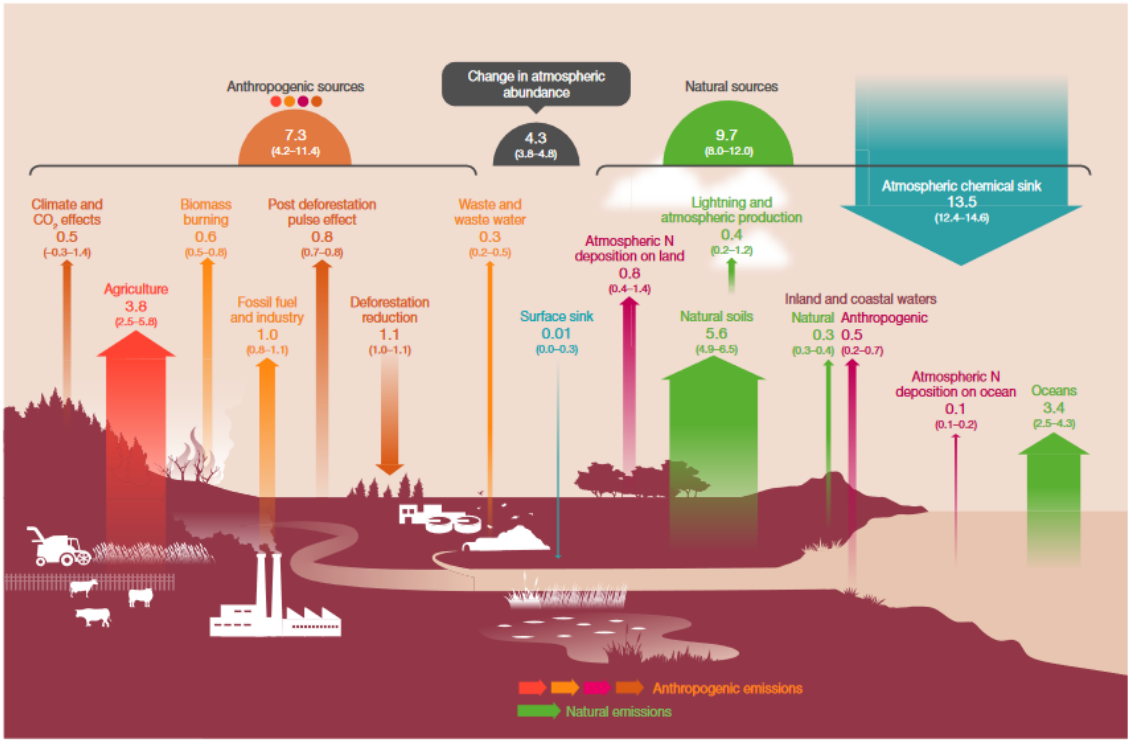
2 Introducción

2.1 Contexto y justificación del Trabajo

Este TFM consistió en identificar perfiles para la enzima óxido nítrico reductasa (*nor*), implicada en la emisión de óxido nítrico, precursor del óxido nitroso y molécula señal implicada en numerosos procesos metabólicos. Actualmente no se disponen de perfiles de esta proteína que permitan discriminar en función de la clase de la proteína. A partir de este proyecto, se podrán identificar diferentes clases de *nor* con capacidad para discriminar y clasificar a los microorganismos y enzimas. Los perfiles identificados en este trabajo podrían sentar las bases de futuros trabajos para el aislamiento de nuevas cepas bacterianas con propiedades interesantes para su uso en agricultura y medio ambiente, o en estudios de biodiversidad y el desarrollo de nuevas enzimas con interés científico o tecnológico.

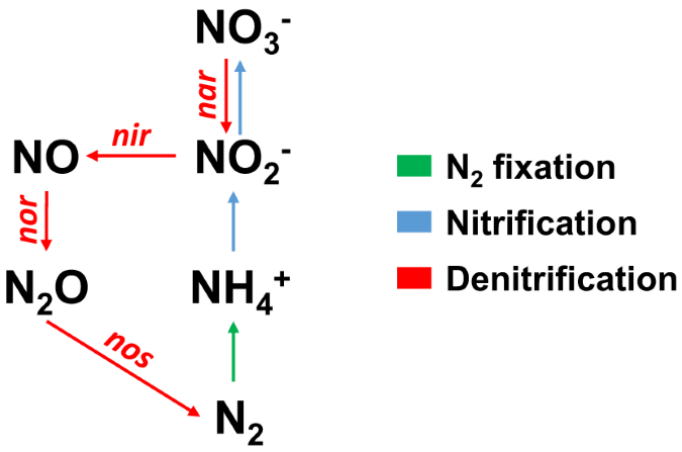
En los últimos años, el interés por el óxido nitroso (N_2O) ha crecido debido a que se trata de un gas con un potente efecto invernadero con una permanencia en la estratosfera de aproximadamente 150 años [1-3]. Aunque este gas únicamente representa alrededor del 0,03% de las emisiones totales de gases de efecto invernadero, tiene un potencial contaminante 300 veces mayor que el dióxido de carbono (CO_2) [4] y este gas, en unidades equivalentes, representa el 10% de las emisiones totales [4]. Más de dos tercios del gas producido tiene su origen en los procesos de desnitrificación y nitrificación de algunas bacterias y hongos en los suelos [3,5]. La producción de NO_2 ha crecido exponencialmente en el último siglo debido a la intensificación de la agricultura y a la mayor utilización de nitrógeno para fertilizar los cultivos. Por otro lado, la ganadería intensiva también contribuye en las emisiones totales de N_2O [2,3]. En la Figura 2.1.1. puede verse un esquema de las emisiones por actividad. Hoy en día, existe un gran interés en reducir las emisiones de N_2O para combatir los efectos del cambio climático, sin embargo, todavía se desconocen muchos aspectos y mecanismos relacionados con la producción de este gas, lo que dificulta el desarrollo de estrategias eficaces.

Figura 2.1.1. Emisiones totales de N_2O [6].



Parece poco probable que alguna vez sea posible desarrollar prácticas agrícolas que eliminen por completo las emisiones de N_2O de los desnitrificantes del suelo. Sin embargo, debería ser posible mitigar las emisiones de N_2O mediante el estudio de la microbiología de la desnitrificación para el desarrollo de estrategias que permitan modular las bacterias desnitrificantes [7]. En los procesos de desnitrificación (Figura 2.1.2.) intervienen cuatro enzimas para reducir el ion NO_3^- a N_2, sin embargo, este proceso no siempre se completa o se realiza de una forma eficiente, lo que provoca emisiones de los productos intermedios, como el N_2O. Una de las enzimas que participan en la desnitrificación es la óxido nítrico reductasa (*nor*), responsable de uno de los procesos intermedios de la desnitrificación, y se encarga de convertir el óxido nítrico (NO) a N_2O. El NO es una citotoxina potente y las bacterias deficientes en *nor* mueren debido a este intermedio tóxico [8,9]. Igualmente, el NO puede contribuir en la formación de ozono, uno de los gases con mayor efecto invernadero [10]. Las óxido nítrico reductasas más conocidas son la cNor y la qNor que utilizan citocromo c / cupredoxinas o quinonas como socios redox inmediatos y ambas pertenecen a la superfamilia de hemo-cobre oxidasas, sin embargo, existen más tipos [11]. Muchos de los estudios de esta enzima se han realizado únicamente en *Paracoccus denitrificans* por lo que la información disponible en la literatura puede no ser completa [12] e igualmente, los perfiles de la óxido nítrico reductasa disponibles en las bases de datos a día de hoy no tienen poder discriminatorio y no son capaces de distinguir entre las clases de ésta [13].

Figura 2.1.2. Proceso de desnitrificación (www.nature.com/articles/s41598-019-55408-z).



Por esta razón, existe una necesidad real de establecer un perfil de la óxido nítrico reductasa capaz de discriminar en función de la clase, ya que esto permitirá identificar la enzima en genomas ya secuenciados o en secuencias obtenidas *de novo*, obtener nuevas cepas bacterianas con propiedades interesantes para su uso en agricultura y medio ambiente, o para estudios de biodiversidad, que darán lugar a nuevos conocimientos en relación a la *nor,* y permitirá a la comunidad científica adquirir una mayor comprensión de los mecanismos de desnitrificación, lo que en último lugar permitirá desarrollar estrategias viables para reducir las emisiones de N_2O.

2.2 Objetivos del Trabajo

El único **objetivo** de este proyecto consiste en establecer el perfil (o perfiles) de la enzima óxido nítrico reductasa (*nor*), implicada en la emisión de óxido nítrico, precursor del óxido nitroso (uno de los principales gases de efecto invernadero) y molécula señal implicada en numerosos procesos metabólicos.

Para alcanzar dicho objetivo, se prevén los siguientes **objetivos específicos**:

- Establecer relaciones filogenéticas entre las secuencias disponibles en las bases de datos anotadas como *nor* o *nitric oxide reductase*.
- Elaborar un árbol filogenético a partir de las relaciones entre las secuencias obtenidas e identificar las diferentes clases de la enzima óxido nítrico reductasa.
- Obtener perfiles generalizados de la enzima óxido nítrico reductasa que permitan discriminar entre las diferentes clases existentes.
- Validar los diferentes perfiles de la enzima óxido nítrico reductasa mediante bases de datos anotadas y librerías metagenómicas.

2.3 Enfoque y método seguido

Una posible estrategia consistiría en la secuenciación de un gran número de microorganismos, con el objetivo de establecer el perfil de la enzima *nor*, sin embargo, esta estrategia no es viable de ningún modo, ya que requería de una gran inversión económica y una gran cantidad de trabajo y análisis.

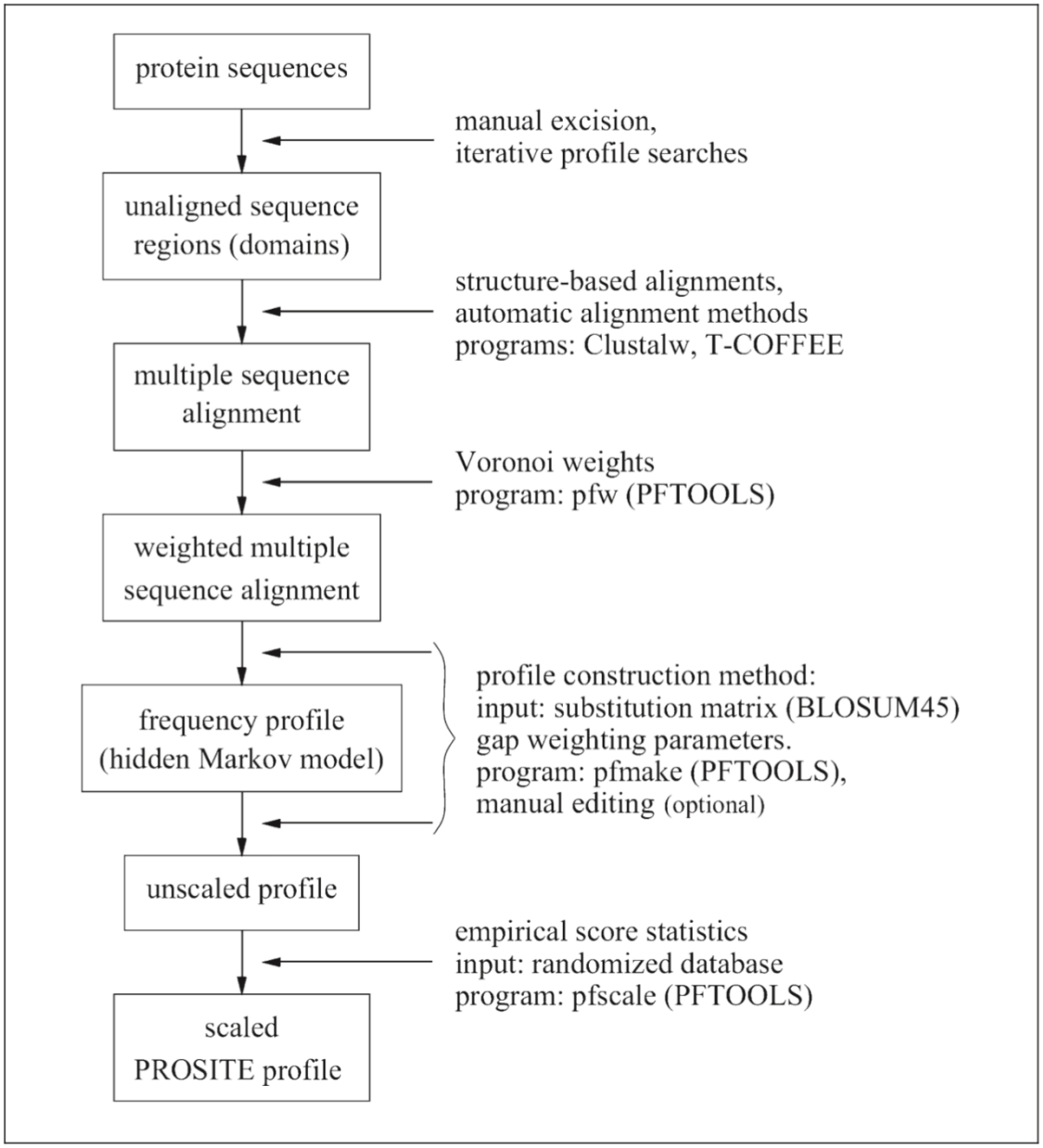
Otra posible estrategia sería la secuenciación de microorganismos obtenidos mediante cultivos y la elaboración de árboles filogenéticos de forma similar a la descrita por Casciotti y Ward [14] y la posterior obtención de los perfiles. Igualmente, esta estrategia tendría un coste elevado y requeriría de una gran cantidad de análisis, si bien estos serían económicamente viables. Desafortunadamente, esta estrategia se limita a los microorganismos estudiados.

A día de hoy, existe un gran número de microorganismos genotipados cuyas secuencias están almacenadas en bases de datos accesibles a toda la comunidad científica de forma gratuita. Una estrategia sencilla, práctica y económicamente viable para establecer el perfil de la enzima óxido nítrico reductasa podría hacer uso de dichas bases de datos y de la gran cantidad de trabajo realizado hasta la fecha.

El enfoque y metodología propuestos en este proyecto para cumplir con los objetivos del proyecto, está basado en los trabajos realizados Udaondo y col. [13]. En resumen, ésta se basa en aprovechar la gran cantidad de información disponible existente (secuencias) para establecer el perfil de la enzima mediante la construcción o elaboración de perfiles generalizados PROSITE, ampliamente utilizados para la generación de nueva información e investigación. En su trabajo, Udaondo y col. [13] demostraron que esta metodología es una solución viable y práctica para caracterizar perfiles de enzimas con poder discriminante capaces de detectar la proteína objetivo en secuencias. Igualmente, esta estrategia se puede ejecutar utilizando únicamente software libre, lo que reduce el coste. Será necesario adaptar la metodología descrita al presente trabajo, lo que arroja cierta incertidumbre al proyecto, ya que ésta puede no ser adecuada para cumplir con los objetivos del proyecto.

Esta metodología se basa en el método original de Gribskov y col. [15] con algunas modificaciones [16,17]. El desarrollo de un perfil para una proteína implica lógicamente varios pasos, como se muestra en la Figura 2.2.1. La primera parte, y quizás la más crítica, es la generación de un buen alineamiento múltiple a partir de secuencias completas.

Figura 2.2.1. Esquema de la metodología propuesta en este Trabajo [18].



La estrategia propuesta se basa principalmente en tres etapas:

- Elaboración del árbol filogenético
- Desarrollo de los perfiles
- Validación de los perfiles

2.4 Planificación del Trabajo

El proyecto está compuesto por diversas actividades interrelacionadas, divididas en dos Hitos donde los resultados de las Tareas anteriores son necesarios para las Tareas posteriores. Se trata de un proyecto sistemático ya que se lleva a cabo según un plan, manteniendo un registro tanto del proceso como de los resultados cuya finalidad está bien definida.

Para cumplir con el objetivo general del proyecto, así como con los objetivos específicos asociados a dicho objetivo, se han propuesto las siguientes Tareas que se describen a continuación:

- **Tarea 1. Estudio de la situación actual y elaboración de los protocolos experimentales**
  - Duración: 5/10/21 – 20/12/21
  - Esta Tarea consiste en realizar una revisión bibliográfica en relación a la temática del proyecto (Estado del Arte), y en la redacción y actualización, en caso de que fuera necesario debido a desviaciones técnicas o temporales, de los protocolos experimentales a utilizar durante el proyecto.
- **Tarea 2. Desarrollo de los trabajos experimentales**
  - Duración: 6/10/21 – 20/12/21
  - Esta Tarea consiste en la ejecución de los trabajos experimentales necesarios para cumplir tanto con el objetivo general del proyecto como con los objetivos específicos.

Esta tarea esta a su vez formada por tres Subtareas:

- *Subtarea 2.1. Elaboración del árbol filogenético*
  - *Duración: 6/10/21 – 03/11/21*
  - Obtención de secuencias
  - Filtrado de las secuencias
  - Elaboración del árbol filogenético
- Subtarea 2.2. *Establecimiento del perfil de la proteína*
  - *Duración: 04/11/21 – 23/1//21*
  - Valoración y estudio del árbol filogenético
  - Obtención de los diferentes perfiles
- *Subtarea 2.3. Validación de los perfiles*
  - *Duración: 30/11/21 –* 20/12/21

Por otro lado, como trabajos asociados a la ejecución del proyecto se encuentran las siguientes tareas:

- R1. Elaboración del primer informe parcial – 8/11/21
- R2. Elaboración del segundo informe parcial – 9/12/21
- R3. Elaboración de la memoria – 5/10/21 – 24/12/21
- R4. Elaboración de la presentación – 27/12/21 – 3/1/22

- R5. Defensa pública – 13/1/21 – 21/01/22

En la Tabla 2.4.1. se muestra el tiempo dedicado a las Tareas definidas anteriormente, así como a las tareas de elaboración de los informes parciales de cada Hito (R1 y R2), elaboración de la memoria (R4), elaboración de la presentación (R4) y defensa pública (R5). En la Figura 2.4.1. se muestra el diagrama de Gantt del proyecto.

Tabla 2.4.1. Dedicación a las Tareas del Proyecto.

Tarea	Descripción	Duración (días)
Tarea 1	Revisión bibliográfica y elaboración protocolos	55
Tarea 2. Desarrollo de los trabajos experimentales		
Subtarea 2.1.	Árbol filogenético	19
Subtarea 2.2.	Obtención perfiles	11
Subtarea 2.3.	Validación	15
R1	Primer informe	-
R2	Segundo informe	-
R3	Memoria	59
R4	Presentación	6
R5	Defensa pública	7

Como ya se ha comentado, el proyecto consta de dos Hitos:

El **primer Hito** finaliza el 8/11/21 e incluye la finalización de las siguientes Tareas:

- Subtarea 2.1. Elaboración del árbol filogenético
- Subtarea 2.2. Parcial – Valoración y estudio del árbol filogenético.

El **segundo Hito** finaliza el 20/12/21 e incluye la finalización de las siguientes Tareas:

- Tarea 1. Estudio de la situación actual y elaboración de los protocolos experimentales.
- Subtarea 2.2. Establecimiento del perfil de la proteína
- Subtarea 2.3. Validación de los perfiles

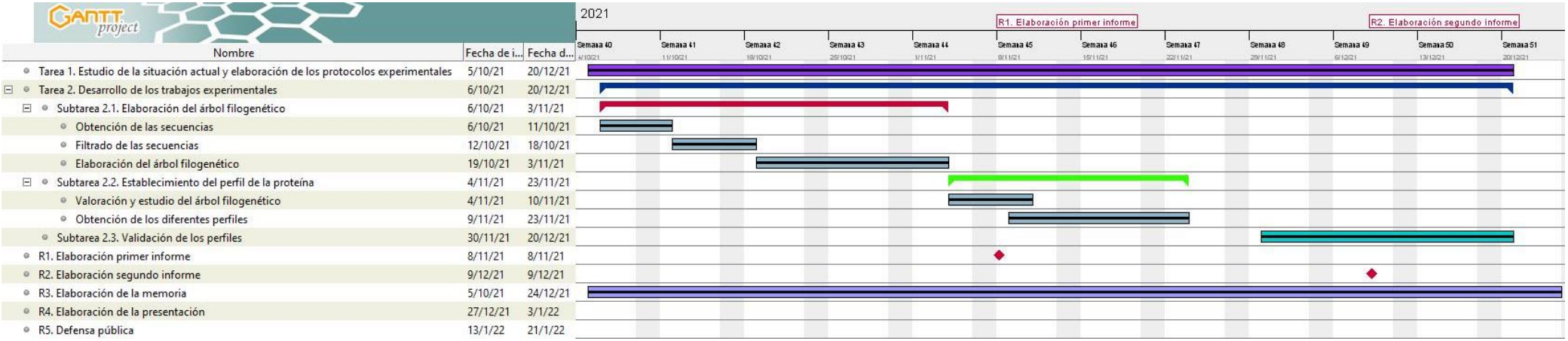
2.5 Breve sumario de contribuciones y productos obtenidos

El objetivo de este proyecto era establecer el perfil (o perfiles) de la enzima óxido nítrico reductasa (*nor*). Como resultado, se han construido tres perfiles generalizados para los tres tipos de *nor* identificadas en la bibliografía. Dichos perfiles se pueden encontrar en los Anexos a esta memoria. Estos perfiles, una vez validados con mayor profundidad, podrían subirse a PROSITE (https://prosite.expasy.org/) para que otros usuarios pudieran utilizarlos.

2.6 Breve descripción de los otros capítulos de la memoria

- Estado del arte: Este capítulo incluirá un estudio de la situación actual del tema sobre el que desarrollará el TFM (*nor*).
- Metodología: En este capítulo se describirá en detalle la metodología utilizada en las diferentes tareas experimentales realizadas durante el proyecto.
- Resultados: En este capítulo se presentarán de forma detallada los resultados obtenidos durante la ejecución del proyecto.
- Discusión: En este capítulo se realizará un análisis crítico y objetivo de los resultados obtenidos.
- Conclusiones: Este capítulo recogerá las conclusiones derivadas de los resultados del proyecto y se valorará el grado de cumplimiento de los objetivos del proyecto.
- Glosario: Este capítulo contendrá la definición de los términos y acrónimos más relevantes utilizados dentro de la Memoria.
- Bibliografía: En este capítulo se citarán los trabajos y materiales utilizados para la ejecución del proyecto.
- Anejos: Recopilación de información complementaria útil para la comprensión de la memoria (secuencias p.ej.).

Figura 2.4.1. Diagrama de Gantt del Proyecto.



Nombre	Fecha de i...	Fecha d...
Tarea 1. Estudio de la situación actual y elaboración de los protocolos experimentales	5/10/21	20/12/21
Tarea 2. Desarrollo de los trabajos experimentales	6/10/21	20/12/21
Subtarea 2.1. Elaboración del árbol filogenético	6/10/21	3/11/21
Obtención de las secuencias	6/10/21	11/10/21
Filtrado de las secuencias	12/10/21	18/10/21
Elaboración del árbol filogenético	19/10/21	3/11/21
Subtarea 2.2. Establecimiento del perfil de la proteína	4/11/21	23/11/21
Valoración y estudio del árbol filogenético	4/11/21	10/11/21
Obtención de los diferentes perfiles	9/11/21	23/11/21
Subtarea 2.3. Validación de los perfiles	30/11/21	20/12/21
R1. Elaboración primer informe	8/11/21	8/11/21
R2. Elaboración segundo informe	9/12/21	9/12/21
R3. Elaboración de la memoria	5/10/21	24/12/21
R4. Elaboración de la presentación	27/12/21	3/1/22
R5. Defensa pública	13/1/22	21/1/22

3 Estado del arte

3.1 Óxido nítrico reductasa

La óxido nítrico reductasa (*nor*) es una enzima de membrana que interviene en la desnitrificación, reduciendo el óxido nítrico (NO) a óxido nitroso (N_2O)[19]. Esta clase de enzimas pertenece a la superfamilia hemo-cobre oxidasa (HCuO). A día de hoy, se conocen tres tipos de *nor* bacterianas; cNOR, qNOR y qCuNOR, que difieren en el donante de electrones, el número de subunidades y la composición del centro de transferencia de electrones [19-21]. Como ya se ha comentado, las *nor* son miembros de la superfamilia de hemo-cobre oxidasa y generalmente se subdividen en cNOR oxidantes del citocromo c y qNOR oxidantes de quinol. Las cNOR están formadas por las subunidades NorB con actividad catalítica, que alberga el sitio activo binuclear heme b3-hierro no hemo (FeB), y NorC con un hemo c que acepta electrones. Las qNOR, por el contrario, son enzimas de una sola subunidad (denominadas NorZ). En estas subunidades, la parte C-terminal es homóloga a la subunidad NorB y la N-terminal a la subunidad NorC, donde la N-terminal retiene el pliegue del citocromo c aunque el hemo c está ausente [22].

La cNOR se ha observado en varios organismos desnitrificantes, como *Paracoccus denitrificans* [23,24], *Halomonas halodenitrificans* [25-26], *Pseudomonas nautica* [27], *Pseudomonas stutzeri* [28,29] y *Pseudomonas aeruginosa* [30]. Por otro lado, se han observado qNOR en organismos como *Wautersia eutropha* [31] y *Pyrobaculum aerophilim* [32]. La tercera clase de *nor* (qCuNOR) se ha observado en la bacteria gram-positiva *Bacillus azotoformans*; este tercer tipo de *nor* tiene la forma NorBC (subunidades B+C), pero tiene un sitio CuA en lugar de un hemo- en la subunidad NorC. Curiosamente, también tiene una extensión N-terminal en NorB, que se cree que forma un sitio de unión al quinol para que la enzima pueda derivar electrones de dos fuentes diferentes [33].

La flavorubredoxina es un miembro de la familia de las flavoproteínas de tipo A. Las flavoproteínas de tipo A son una gran familia de enzimas, muy extendidas entre las bacterias y las arqueas, ya sean anaerobios estrictos o facultativos [34,35]. La principal característica distintiva es la unidad central común en todas ellas, construida por dos módulos estructurales independientes [36]. En un trabajo de Gomes y col. [37], se demostró que la flavoproteína purificada en *E.coli* tenía una elevada actividad NO reductasa, y propusieron dicha enzima como un nuevo grupo de proteínas con actividad NO reductasa.

Microorganismos como *E. coli, Moorella thermoacetica, Paracoccus denitrificans, Pseudomonas nautica, Thermotoga maritima o Trichomonas vaginalis* desarrollaron sus propios mecanismos de defensa expresando enzimas como con actividad óxido nítrico reductasa, flavohemoglobinas y proteínas de flavodiiron (FDP) [24,29,38-42]. Las proteínas flavodiiron son una clase de enzimas microbianas, purificadas por primera vez a principios de la

década de 1990 a partir de la bacteria anaeróbica reductora de sulfato Desulfovibrio gigas e identificadas como rubredoxina: oxidorreductasa de oxígeno [43]. Estas proteínas fueron posteriormente renombradas por Wasserfallen y col. [34] como flavoproteínas de tipo A y en 2003 como FDP, debido a la constitución de sus centros redox centrales [41]. Hoy en día, se sabe que lss FDP tienen, en diferentes grados, actividad NO reductasa y O_2 reducatasa; la unidad monomérica de FDP se caracteriza por un núcleo de dos dominios, que es común en todos los FDP analizados y que definen esta familia de enzimas [36,44-46]. Se ha observado que la longitud de dichas enzimas es de alrededor de 825 secuencias [47]. Los FDP de clase B contienen un dominio de rubredoxina C-terminal, el tipo más simple de proteínas FeS, con un centro FeCys4, y por esta razón se han denominado saborubredoxinas (FlRd, ca 475 aminoácidos, también llamado norV). Los FDP de clase B están restringidos al filo de Proteobacteria, de las clases beta, delta y gamma, éstas han sido estudiada en el microorganismo *E. coli* [35,45] donde se ha demostrado que confiere resistencia al NO en condiciones anaeróbicas [40,48,49]. En el microorganismo *Escherichia coli*, la FlRd tiene de hecho una actividad óxido nítrico reductasa significativa y una actividad O_2 reductasa nula.

3.2 PROSITE

PROSITE es una colección anotada de motivos dedicados a la identificación de dominios y familias de proteínas. Los motivos utilizados en PROSITE son patrones o perfiles, obtenidos mediante alineamientos múltiples de secuencias homólogas. Esto permite a dichos motivos identificar relaciones distantes entre secuencias que habrían pasado inadvertidas en el caso de realizar únicamente alineamientos de secuencias por pares. Estos perfiles tienen sus propias ventajas y problemas que definen su área de aplicación óptima.

PROSITE, inicialmente una "firma" o base de datos de perfiles, fue creada en 1988 por Amos Bairoch. La primera versión de PROSITE estuvo disponible en PC / Gene en marzo de 1988 y contenía 58 patrones. Cada perfil iba acompañado de un resumen que describía el dominio o familia de proteínas correspondiente. PROSITE se desarrolló en paralelo con Swiss-Prot y ambas bases de datos se beneficiaron entre sí; se identificaron multitud patrones anotando familias de proteínas en Swiss-Prot. Los patrones se utilizaron luego para poblar Swiss-Prot con nuevos miembros de la familia. Por esta razón, PROSITE suscito un gran interés en la comunidad científica [50]. Aquí puede observarse la efectividad de este sistema para la identificación de proteínas.

En 1994, Philipp Bucher introdujo en PROSITE los "perfiles generalizados" como nuevos descriptores de motivos [51]. Todos estos métodos, son más o menos, una valoración estadística de una alineación de múltiples secuencias; se usan puntuaciones específicas de posición para aminoácidos y penalizaciones específicas de posición para abrir y extender una inserción o deleción. Los "perfiles generalizados", en comparación con los perfiles anteriores [52], utilizan

una sintaxis más rigurosa para los estados de inserción, eliminación y coincidencia. Dado que la sintaxis de "perfil generalizado" es muy similar a la del perfil de HMM, casi todas las puntuaciones de "perfil generalizado" se pueden asignar a los parámetros de HMM utilizados por HMMER [15]. Actualmente, casi todas las nuevas entradas de PROSITE son perfiles. Desde su creación, PROSITE ha permitido generar nueva información de gran calidad en relación a dominios, familias y sitios funcionales de multitud de proteínas.

3.3 Perfiles generalizados

Como indica Sigrist y col. en su trabajo [18], en algunos casos, la secuencia de una proteína desconocida está relacionada con otra proteína cuya estructura sí se conoce, de tal forma que es posible detectar similaridades mediante el alineamiento de las secuencias por pares. Sin embargo, estas relaciones también se pueden identificar mediante la aparición de residuos en las secuencias, dichos residuos se conocen de diversas formas como perfil, motivo, o firma. Estos motivos, típicamente están formados por entre 10 y 20 aminoácidos, y surgen porque los residuos específicos y las regiones que se cree o han demostrado ser importantes para la función biológica de un grupo de proteínas se conservan. Estas regiones o residuos biológicamente significativos son generalmente:

- Sitios catalíticos de enzimas.
- Sitios de unión de grupos prostéticos (hemo, piridoxal-fosfato, biotina, etc.).
- Aminoácidos involucrados en la unión de un ión metálico.
- Cisteínas involucradas en enlaces disulfuro.
- Regiones involucradas en la unión de una molécula (ADP / ATP, GDP / GTP, calcio, ADN, etc.) u otra proteína.

Como la secuencia de motivos biológicamente significativos se conserva evolutivamente, una alineación múltiple de éstas puede reducirse a una expresión consenso denominada expresión o perfil. Cada posición de dicho perfil puede estar ocupada por cualquier residuo de un conjunto específico de residuos aceptables y, además, puede repetirse un número variable de veces dentro de un rango especificado. En posiciones estrictamente conservadas solo se acepta un aminoácido particular, mientras que en otras posiciones se pueden aceptar varios aminoácidos con propiedades fisicoquímicas similares. También es posible definir qué aminoácidos son incompatibles con una posición dada, y los residuos conservados pueden separarse mediante espacios de longitudes variables.

Una expresión regular es cualitativa; o coincide o no. No existe un umbral por encima del cual consideremos la coincidencia como estadísticamente significativa. Sin embargo, es posible evaluar la precisión de los patrones

PROSITE gracias al número de coincidencias obtenidas al escanear la base de datos como SWISS-PROT o al escanear otras bases de datos (ver más abajo).

Las ventajas de los perfiles son su fácil interpretación y el hecho de que éstos perfiles se centran en los residuos más conservados. Dado que estos residuos a menudo son importantes para la función biológica de la familia o el dominio de proteínas, existe la posibilidad de realizar investigación adicional centrandose únicamente en ellos, por lo que la generación de éstos perfiles podría contribuir a la generación de nuevos datos y estudios, así como para el desarrollo de aplicaciones. Otra ventaja de los perfiles es que el escaneo de una base de datos de proteínas se puede realizar en un tiempo razonable con la mayoría de equipos.

El poder discriminatorio de los perfiles se debe a las capacidades intrínsecas del propio perfil, así como a los métodos utilizados para la construcción de perfiles. Los perfiles son descriptores de motivos cuantitativos que proporcionan pesos numéricos para cada posible coincidencia entre el residuo de una secuencia y una posición del perfil. El procedimiento automático utilizado para obtener perfiles a partir de alineaciones múltiples es capaz de asignar pesos a los residuos que aún no se han observado, gracias a la utilización de otras herramientas, como las matrices de substitución. Por el contrario, este mismo sistema, no permite elaborar suposiciones sobre qué residuos aún no detectados podrían observarse en el futuro.

Los perfiles generalizados [53] utilizados en PROSITE son una extensión de los perfiles introducidos por Gribskov y col. [54]. Son estructuras lineales similares a secuencias que consisten en puntuaciones o pesos. El formato de perfil utilizado en PROSITE comprende campos para los denominados parámetros accesorios que definen el método de búsqueda que se utilizará para un dominio particular. Permiten especificar valores de corte apropiados, diferentes modos de normalización de puntuaciones e instrucciones sobre cómo tratar coincidencias parcialmente superpuestas.

4 Metodología

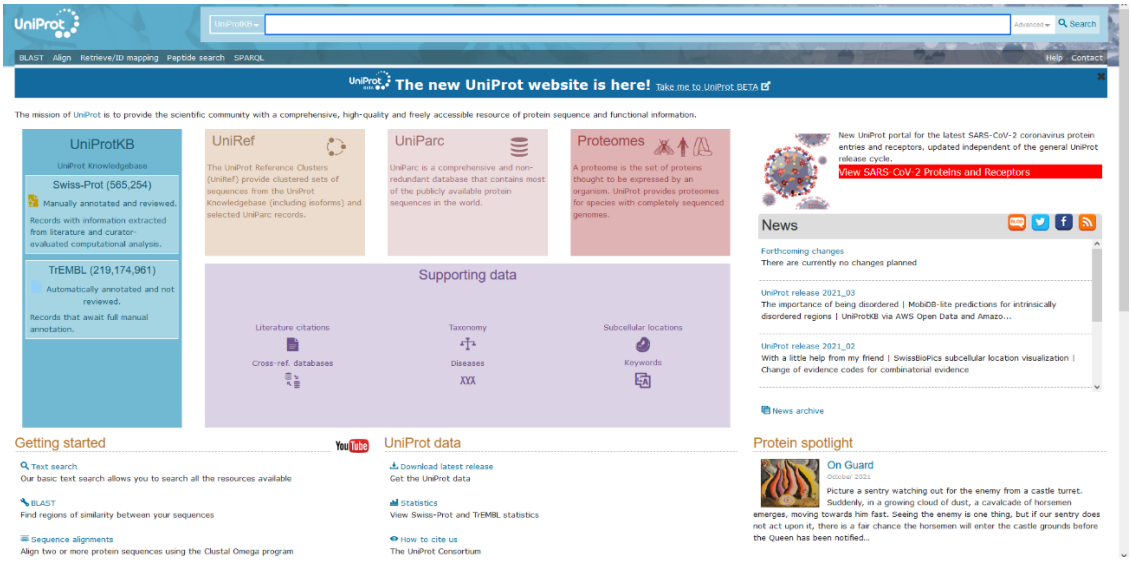
4.1 Obtención de secuencias

En primer lugar, se accedió a la base de datos *Uniprot* (Figura 4.1.; https://www.uniprot.org/), y se descargaron en formato *Excel* las secuencias obtenidas mediante el siguiente criterio de búsqueda:

taxonomy:bacteria goa:("nitric oxide reductase activity") NOT norc

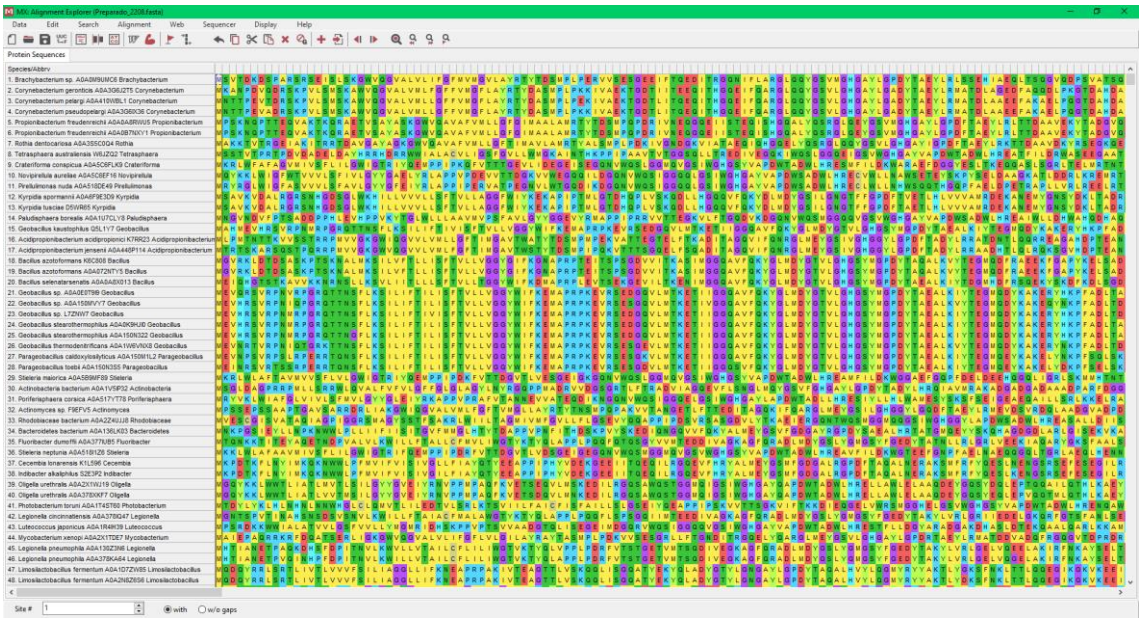
Se descargaron un total de 2.363 secuencias. La lista de secuencias se puede encontrar en el **Anexo 1** adjunto a esta memoria. A continuación, en dicho *Excel* se eliminaron las secuencias cuya secuencia de aminoácidos era demasiado larga o demasiada corta. Igualmente, se realizó un filtrado manual de las secuencias, donde aquellas que contenían palabras como "*uncultured*" o "*fragment*" se eliminaron. Por último, las 2.208 secuencias restantes se ordenador de mayor a menor en función de la longitud (*length*) de la secuencia de aminoácidos. Esta lista puede encontrarse en el **Anexo 2** adjunto a esta memoria.

Figura 4.1. Base de datos *Uniprot*.



La lista de 2.208 secuencias se convirtió a formato FASTA mediante el código Python que puede encontrarse en los Anexos de esta memoria. Una vez obtenido el archivo FASTA, las secuencias se alinearon mediante MUSCLE [55] con el software informático MEGA (Figura 4.2.; MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms [56]). Primeramente, se realizó un alineamiento con los parámetros por defecto (*Gap Open:* -2,90*; Gap Extend:* 0,00*; Hydrophobicity Multiplier:* 1,20*; Cluster Method (Iterations 1,2):* UPGMA*; Cluster Method (Other Iterations):* UPGMA*; Min Diag Length (Lambda):* 24) y dos iteraciones.

Figura 4.2. Secuencias (2208) abiertas en MEGA X.



Una vez finalizado el primer alineamiento (**Anexo 3**), se procedió a una inspección visual y filtrado manual de las secuencias. El criterio utilizado para el filtrado de las secuencias fue el siguiente:

1. Como se ha comentado anteriormente, las secuencias estaban ordenadas de mayor a menor tamaño, por lo tanto, se puede razonar que las secuencias con un tamaño parecido sean similares. Por lo tanto, durante la inspección visual, aquellas secuencias muy divergentes (teniendo en especial consideración a los *gaps*) a las secuencias próximas fueron eliminadas (Figuras 4.3. y 4.4.).

2. Por otro lado, si se detectaba un grupo de tamaño considerable (n > 5) de secuencias divergentes según el primer criterio, similares y próximas entre ellas, éstas no se eliminaban (Figura 4.5.).

Figura 4.3. Secuencias similares (no se eliminan).

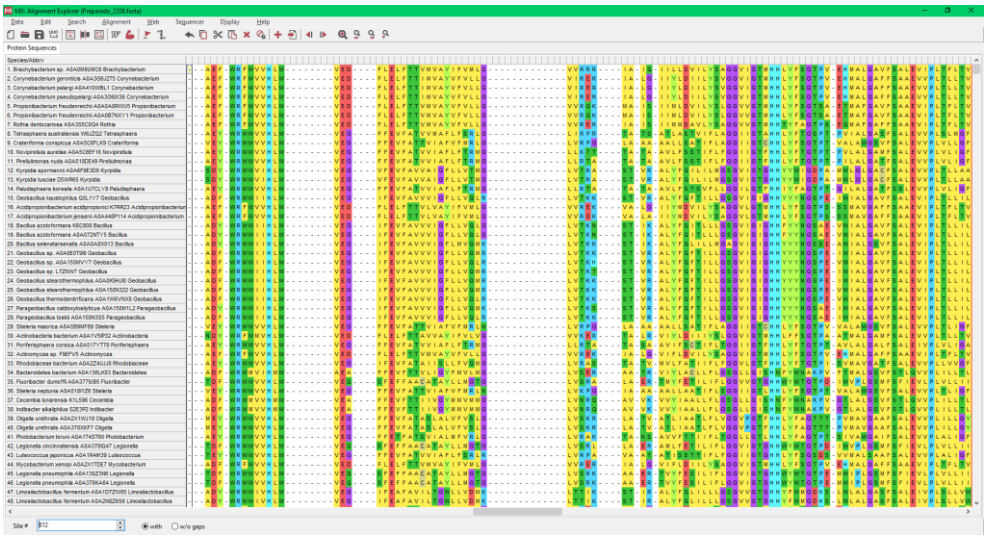


Figura 4.4. Secuencias divergentes (se eliminan).

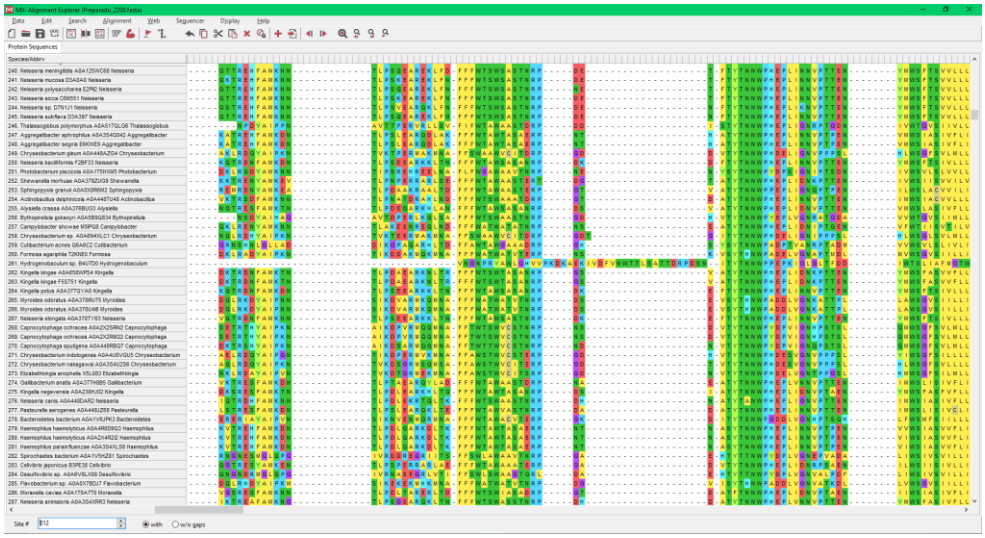
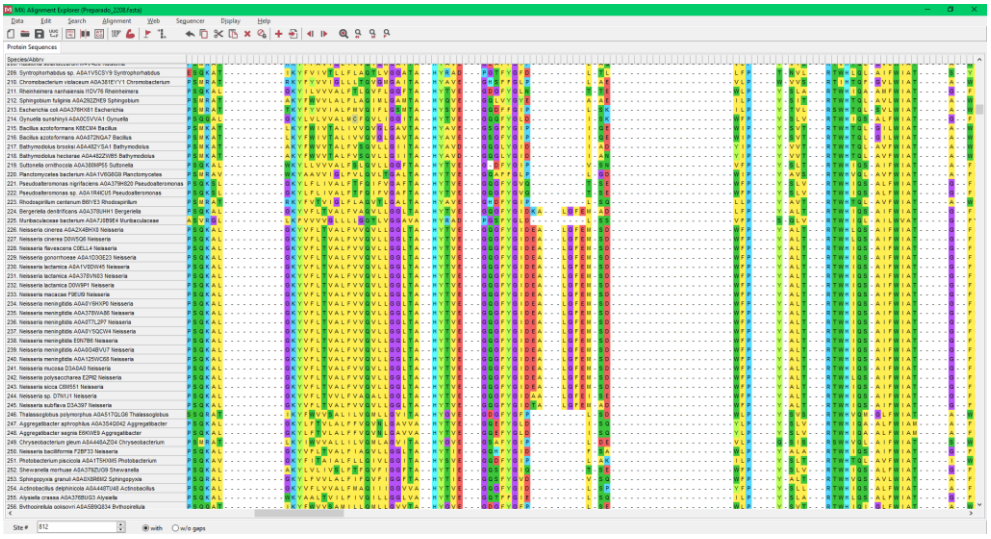


Figura 4.5. Ejemplo para el criterio nº 2.



Una vez realizado este filtrado manual, se procedió a la eliminación automatizada de las secuencias duplicadas, para ello, se utilizó la aplicación web que puede encontrarse en este enlace:

https://arn.ugr.es/srnatoolbox/helper/removedup/

A continuación, las secuencias restantes volvieron a alinearse siguiendo los mismos criterios descritos anteriormente y volvió a realizarse una inspección visual y filtrado manual (**Anexo 4**). Finalizado el segundo alineamiento, se disponía de un total de 1.543 secuencias con 1.726 posiciones. Por último, se realizó un tercer alineamiento (final) con los parámetros por defecto y 16 iteraciones (**Anexo 5**). El resultado de este último alineamiento fue utilizado para la elaboración del árbol filogenético de la proteína *nor*.

4.2 Elaboración del árbol filogenético

Para la elaboración del árbol filogenético, en primer lugar, se realizó una búsqueda para encontrar el mejor modelo mediante IQ-TREE [57,58], utilizando el comando:

iqtree -s Alineamiento_3.fas -m MF -safe

La opción *-safe* se utilizó para evitar errores, ya que, en este caso, los datos utilizados (secuencias) podían considerarse que tenían un gran tamaño. Una vez obtenido el mejor modelo, se procedió a la elaboración del árbol filogenético con *ultrafast bootstrap* (UFBoot; [59]) con IQ-TREE, mediante el comando:

iqtree -s Alineamiento_3.fas -m VT+F+R10 -bb 1000 -nt AUTO -nstop 50

Debido a la escasa potencia del equipo informático utilizado, inicialmente se utilizaron parámetros que reducían el tiempo de computación. Mediante ensayo y error, finalmente se utilizó el comando descrito arriba para la elaboración del árbol filogenético, ya que parecía ser una opción equilibrada entre tiempo de computación y calidad de los resultados. Se ha demostrado que los parámetros por defecto de IQ-TREE son adecuado en la mayoría de las situaciones [57], sin embargo, la utilización de estos parámetros requería de mucho tiempo de computación. Se observó que se podía reducir el tiempo fijando el parámetros *-nstop* a 50 (por defecto 100). Nguyen y col. [57] sugieren que éste es un parámetro importante, pero indican que es especialmente importante en árboles donde se trabaja con muchas secuencias cortas. En este trabajo se consideró que las secuencias no eran cortas, por lo que se consideró aceptable reducir este parámetro en el contexto de este trabajo. Por último, todos los árboles filogenéticos obtenidos (durante ensayo – error) arrojaron resultados muy similares, lo que sugiere que el árbol obtenido con *-nstop* 50 es adecuado y utilizar los parámetros por defecto no arrojaría resultados diferentes. A raíz de este proceso, se obtuvo un Árbol consenso, que se utilizó en las siguientes actividades.

Por último, el árbol consenso obtenido en el análisis anterior se visualizó mediante el software Interactive Tree Of Life (iTOL) v5 [60].

4.3 Determinación de las secuencias para cada grupo

Se obtuvieron las etiquetas de las secuencias para cada clade (grupo/rama) de la representación gráfica en Tree Of Life (iTOL) v5 [60] de forma manual. Una vez obtenidas las etiquetas, mediante el código Python que puede verse en los Anexos de esta memoria, se compararon las referencias (presentes en las etiquetas) con las referencias del archivo *Excel* descargado originalmente de *Uniprot* para recuperar las secuencias de aminoácidos y el organismo. Los tres archivos para los Grupos Azul, Rojo y Verde se pueden encontrar en los **Anexos 7, 8 y 9**, respectivamente, adjuntos a este documento. Además, el código Python también creaba el archivo en formato FASTA para su posterior alineamiento. Dichos documentos pueden encontrarse en los **Anexos 10, 11** y **12** adjuntos.

Con la información recuperada mediante el código Python (organismo y secuencia de aminoácidos para cada grupo), se consultó la base de datos de NCBI (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?name=XX), y para cada secuencia de cada grupo se determinó el Orden y la Familia, y se comparó la información.

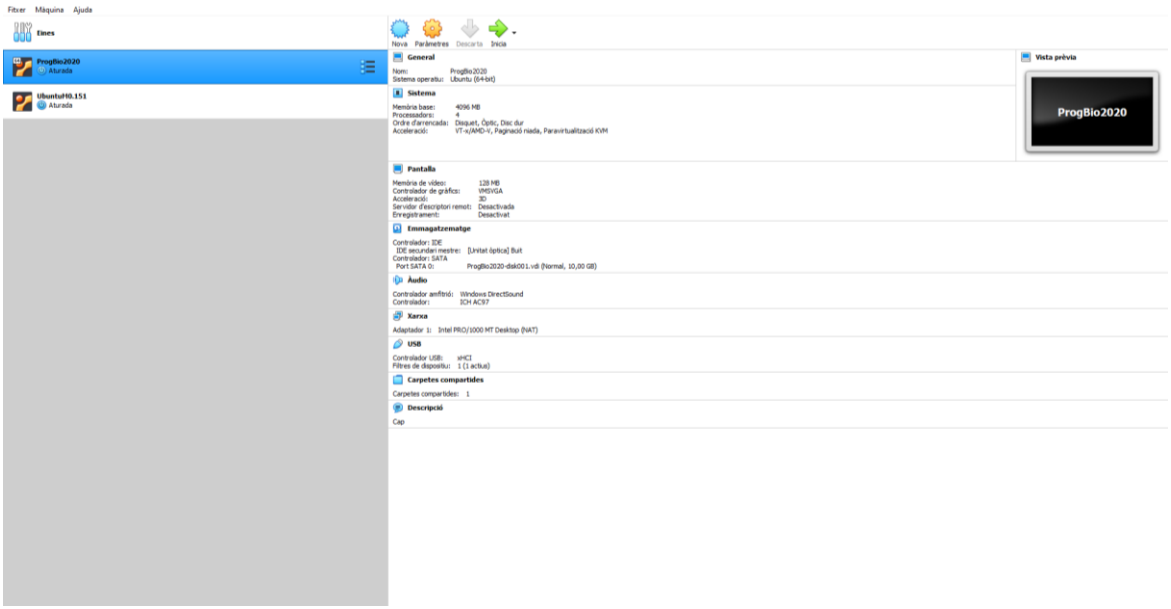
Por último, las secuencias en formato FASTA para cada grupo se alinearon de forma independiente mediante MUSCLE [55] utilizando el software informático MEGA X [56]. Se utilizaron los parámetros por defecto y se realizaron 32 iteraciones. Las secuencias alineades pueden verse en los **Anexos 13, 14** y **15.**

4.4 Obtención de perfiles

Para la elaboración de los perfiles, se utilizó la metodología propuesta por PROSITE (*https://prosite.expasy.org/prosuser.html;* Apartado II.B.) mediante el método clásico desarrollado por Gribskov y col. [15]. que utiliza un fichero con múltiples secuencias alineadas como *input* y utiliza una tabla para convertir las frecuencias de los residuos en pesos, aplicando las modificaciones de Luethy y col. [16].

Para ello, se utilizó el paquete de herramientas *pftools* (https://bio.tools/pftools). Se trata de una colección de programas para construir, calibrar y buscar secuencias biológicas con perfiles generalizados. En primer lugar, como éste no era compatible con el OP Windows (o daba problemas), se utilizó una máquina virtual Linux (Oracle VM VirtualBox; Figura 4.6.).

Figura 4.6. Oracle VM VirtualBox.



Una vez en la máquina virtual, para la instalación y gestión del paquete, se utilizó *Miniconda* con el comando

conda install pftools

Miniconda es una versión reducida de *Conda*, un sistema *open-source* de gestión de paquetes. Una vez instalado el paquete *pftools*, fue necesario modificar las etiquetas de las secuencias utilizadas, ya que éstas debían de ser idénticas. Para esto, se utilizó el código Python utilizado anteriormente con ligeras modificaciones. El código puede encontrarse en los Anexos de esta memoria. Una vez obtenidas las secuencias de cada clase de *nor* con las etiquetas modificadas, se realizó el alineamiento y filtrado de las secuencias para cada clase de *nor*, siguiendo la metodología descrita anteriormente (alineamiento 2 iteraciones-filtrado manual - alineamiento 64 iteraciones – filtrado manual). Estos alineamientos se pueden encontrar en los **Anexos 16, 17 y 18** para las clases Azul, Rojo y Verde, respectivamente (qNOR+qCuNOR, cNOR y vNOR, respectivamente).

A continuación, los archivos con formato FASTA obtenidos tras el alineamiento, se convirtieron a formato MSF (*GCG MSF* (*multiple sequence file*) *file format*) a través de la web (https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/; Figura 4.7.), tal y como se aconseja en los manuales de *pftools.* Gracias a este formato, se añade un peso a cada secuencia, necesario para la elaboración de perfiles generalizados (Figura 4.8.). Los archivos obtenidos pueden encontrase en los **Anexos 19, 20 y 21 (21.1. archivo completo, ver más adelante)** para las clases qNOR+qCuNOR, cNOR y vNOR, respectivamente.

Figura 4.7. EMBOSS para la conversión de formatos.

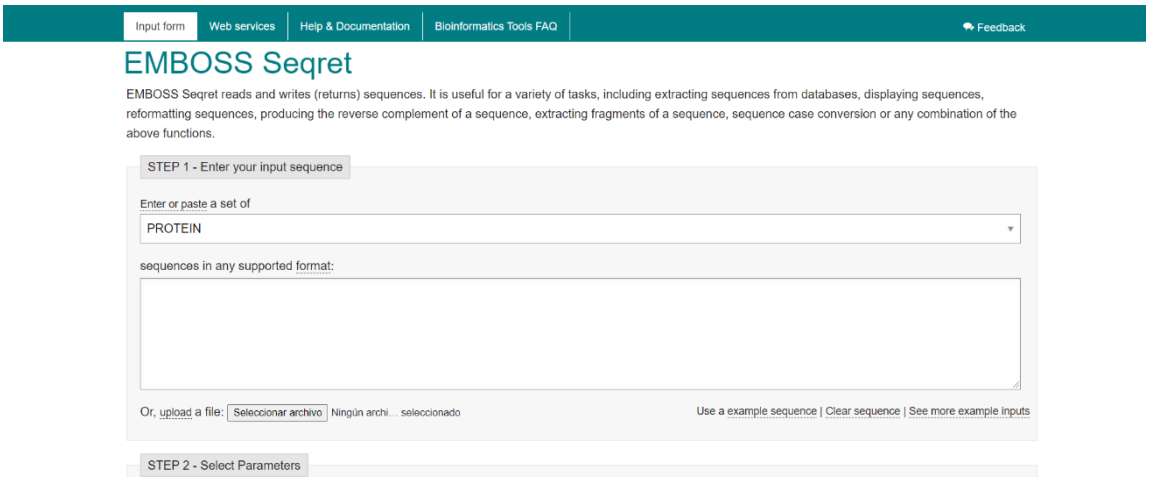
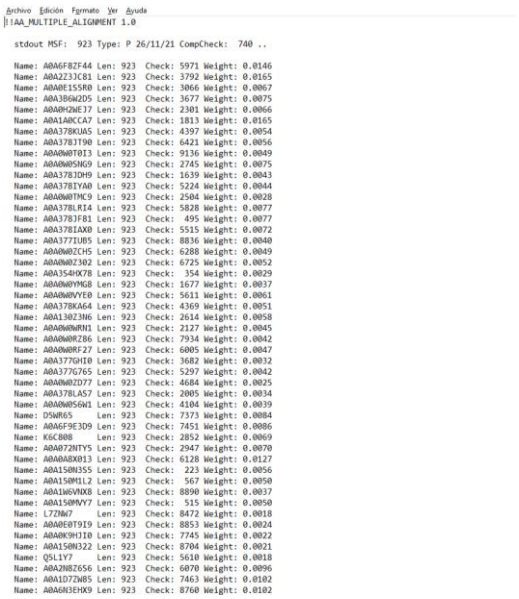


Figura 4.8. Archivo en formato.msf.una vez asignados los pesos con el comando *pfw*.



Una vez los archivos MSF para cada clase de *nor* estuvieron listos, se modificaron los pesos mediante el comando de *pftools, pfw* con los parámetros por defecto (-N 100, -W 1, -X 0.5), ya que estos se consideraron adecuados. El commando *pfw* calcula nuevos pesos para las secuencias individuales en un alineamiento múltiple mediante la metodología propuesta por Sibblad y Argos [61]. En el caso de las secuencias para vNOR, debido al tamaño de éste, y que al utilizar el comando *pfw* daba error, dado que una gran parte de las secuencias que contenían eran muy similares, se eliminaron algunas de éstas de forma aleatoria para reducir el tamaño. Los archivos con los nuevos pesos pueden verse en los **Anexos 22, 23 y 24** para las clases qNOR+qCuNOR, cNOR y vNOR, respectivamente.

Por último, para la elaboración de los perfiles generalizados, se utilizó el comando de *pftools*, *pfmake* con los parámetros por defecto y utilizando la matriz

blosum45 ya que es la propuesta en la metodología empleada. Este comando genera un perfil *prosite* a partir de un archivo de alineamientos múltiples, utilizando los métodos descritos por Gribskov y col. [15], Luethy y col. [16] y Thompson y col. [17].

4.5 Validación de los perfiles

Para la validación de los perfiles, es decir, para determinar si son capaces de discriminar entre los diferentes tipos de *nor* se utilzó el comando *pfsearch* del paquete de herramientas *pftools*. La instalación de dichas herramientas se ha descrito anteriormente.

Dicho comando compara un perfil con una base de datos y como resultado se obtiene una lista de secuencias que coinciden con dicho perfil. La validación de los perfiles mediante la búsqueda en bases de datos es una de las metodologías propuestas por PROSITE.

Primeramente, se pretendía utilizar los perfiles en la base de datos completa *Uniref100* (https://www.uniprot.org/downloads#unireflink). Debido al gran tamaño del archivo, y a que fue necesario trabajar con una máquina virtual, ésta se tuvo que almacenar en un disco duro externo. Probablemente, debido a esto, y al potatismo del equipo informático utilizado, el análisis en la base de datos completa requería de mucho tiempo y, por lo tanto, dicho análisis fue inviable.

Se plantearon otras opciones entre las que se incluía trabajar con bases de datos de menor tamaño (*Uniref90* y *Uniref50*), así como utilizar otras herramientas como *pfsearchV3* [62], sin embargo, ninguna de estas opciones fue viable debido al tiempo, a los resultados obtenidos, y a la potencia computacional. Por esta razón, se plantearon alternativas menos sofisticadas para realizar una validación inicial de los diferentes perfiles:

Se descargó la base de datos completa de *Swiss-Prot* (revisada) y se comparó con los tres perfiles. Es de esperar, que como mínimo, los perfiles sean capaces de buscar y discriminar entre las secuencias utilizadas para su construcción.

Por otro lado, se descargó el genoma de 4 organismos bien documentados, dónde la clase de *nor* es conocida (*Paracoccus denitrificans* (cNOR), *Pyrobaculum aerophilum* (qNOR), *Bacillus azotoformans* (qCuNOR) y *E.coli* (vNOR)) de la base de datos del NCBI (https://www.ncbi.nlm.nih.gov/genome/), y se realizó la búsqueda en cada uno de los genomas con cada uno de los tres perfiles para determinar si son capaces de discriminar entre clases.

Por último, se descargó de la base de datos del NCBI (https://www.ncbi.nlm.nih.gov/protein/) secuencias de proteína mediante el criterio de búsqueda:

((hypothetical protein) AND soil metagenome) AND bacteria

mediante el programa *NCBI Mass Sequence Downloader* [63]. Se utilizaron estos criterios de búsqueda ya que el interés principal reside en el estudio de los microorganismos en suelo para la reducción de emisiones.

En todos los casos, se utilizaron los parámetros por defecto de *pfsearch*.

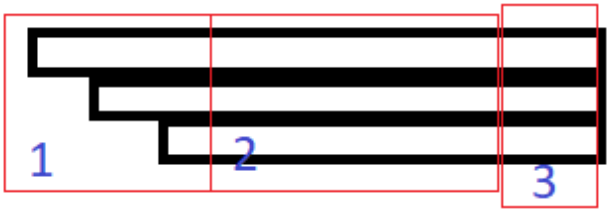
5 Resultados

5.1 Obtención de secuencias

Como resultado del proceso descrito, se obtuvo un listado de secuencias alineadas que se utilizó para la elaboración del árbol filogenético de la proteína *nor*. Dicho alineamiento consistía en 1.543 secuencias (de las 2.363 originales) y 1.726 posiciones y éste se encuentra en el **Anexo 5** adjunto a esta memoria.

Por otro lado, también se realizó, de forma paralela, otro alineamiento en el que la parte final de las secuencias, previo al segundo alineamiento, se eliminó ya que ésta parecía muy variable (Parte 3 de la Figura 5.1.). De igual modo, se elaboró un árbol filogenético a partir de este alineamiento, que se incluyó en este trabajo únicamente con fines comparativos, por lo que no se considera necesario entrar en detalle.

Figura 5.1. Esquema de las secuencias utilizadas.



5.2 Elaboración del árbol filogenético

El resumen de las secuencias utilizadas para la elaboración del Árbol consenso se muestra en la Tabla 5.1.

Tabla 5.1. Resumen de las secuencias utilizadas.

Ítem	Valor
Input data	1.543 secuencias con 1.726 sitios (aminoácidos)
Número de sitios constantes	118 (6,8%)
Número de sitios invariantes	118 (6,8%)
Número de sitios informativos (parsimonia)	1.363
Número de patrones distintos	1.713

Los resultados obtenidos en la determinación del mejor modelo se pueden ver en la Tabla 5.2. El criterio de selección fue la puntuación obtenida en BIC (*Bayerian Information Criteria)*. Como puede verse, el modelo elegido fue VT+F+R10.

Tabla 5.2. Resultados obtenidos.
iqtree -s Alineamiento_3.fas -m MF -safe.

Modelo	BIC
VT+F+R10	686616.8538
VT+R10	686706.9452
VT+F+R9	686708.9271
VT+R9	686797.7610
VT+F+R8	686892.7770

Las estadísticas del árbol utilizado para determinar el mejor modelo se pueden ver en la Tabla 5.3.

Tabla 5.3. Estadísticas del árbol utilizado para determinar el mejor modelo.

Ítem	Valor
Log-verosimilitud del árbol	-331619.1953 (e.s. 9854.6521)
Log-verosimilitud sin restricciones	-12844.9164
Número de parámetros libres	3.120
AIC	669478.3906
AICc	20144518.3906
BIC	686493.5037
Longitud total de las ramas	335.5467
Suma de las longitudes internas de las ramas	157.7339 (47.0%)

Por último, la representación gráfica (*unrooted,* circular sin considerar las distancias y circular) del Árbol consenso obtenido mediante el comando *iqtree -s Alineamiento_3.fas -m VT+F+R10 -bb 1000 -nt AUTO -nstop 50* (Log-verosimilitud -330786.220009) se puede ver en las Figuras 5.2., 5.3. y 5.4., respectivamente. Cada una de las tres principales ramas obtenidas en dicho árbol está coloreada con un color diferente (Azul, Rojo y Verde).

Figura 5.1. Árbol filogenético obtenido para *nor* (*unrooted)*.

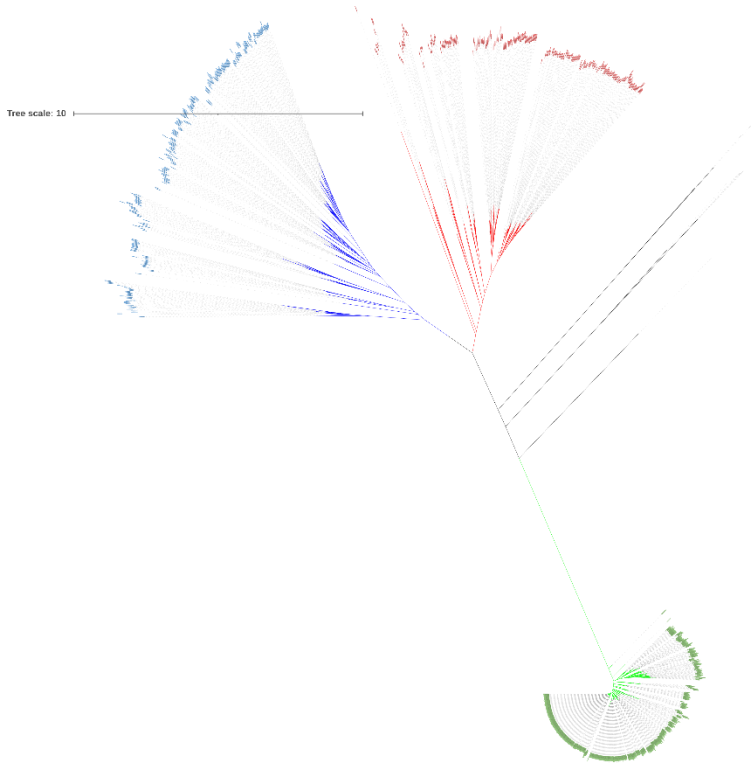


Figura 5.2. Árbol filogenético obtenido para *nor* (circular sin considerar las distancias*)*.

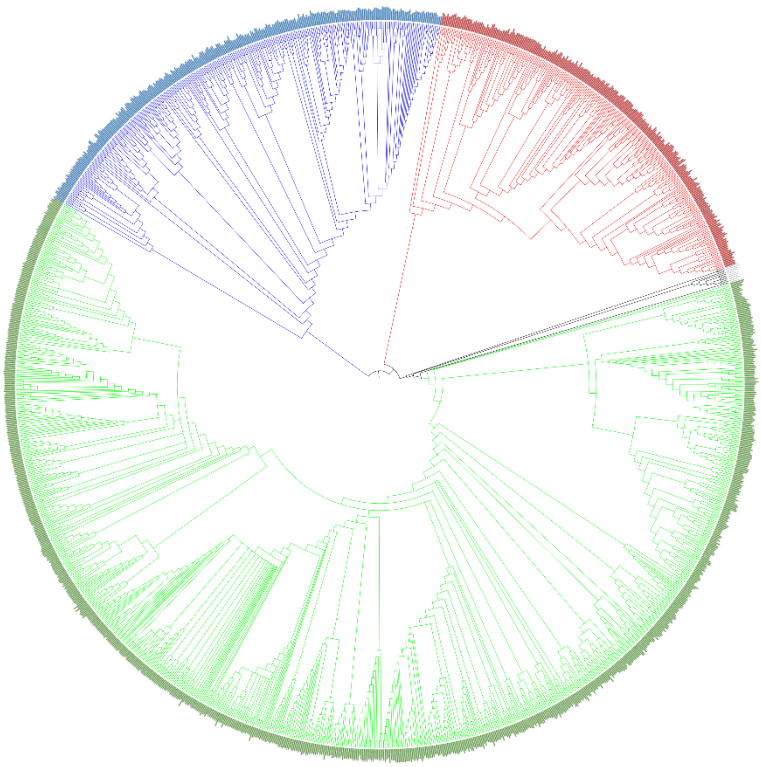
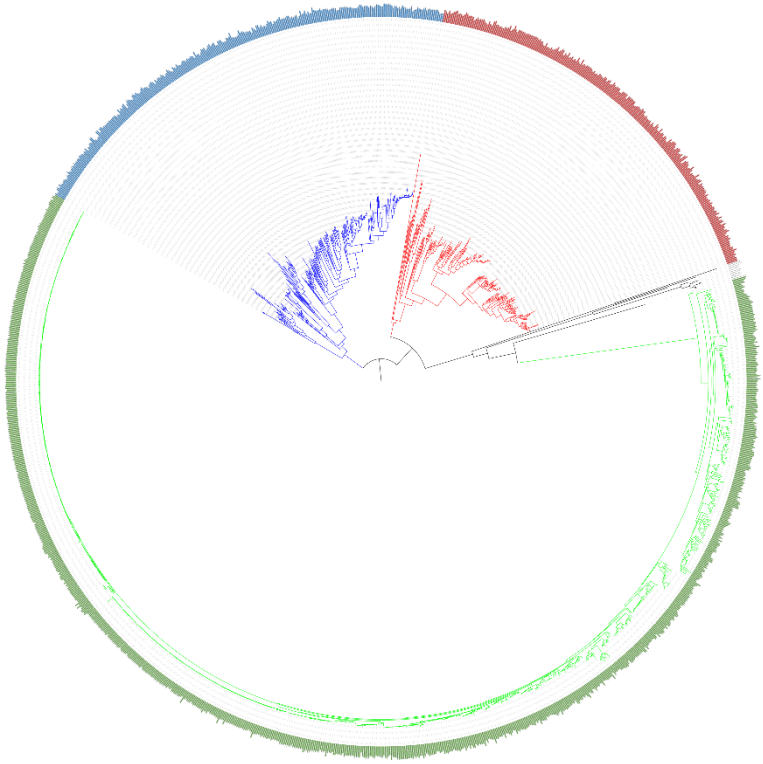


Figura 5.3. Árbol filogenético obtenido para *nor* (circular*)*.



En la Figuras 5.4. se puede ver el árbol obtenido al utilizar las secuencias donde la parte final de éstas fue eliminada. En la Figura 5.5. se puede ver el Árbol consenso al utilizar el parámetro *-fast* de IQ-TREE.

Figura 5.4. Árbol filogenético obtenido para *nor* (*unrooted*) utilizando las secuencias con la parte final eliminada.

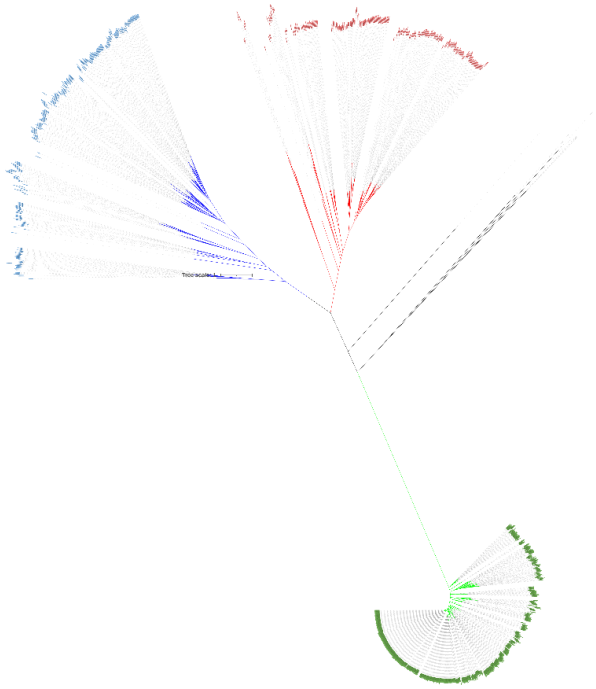
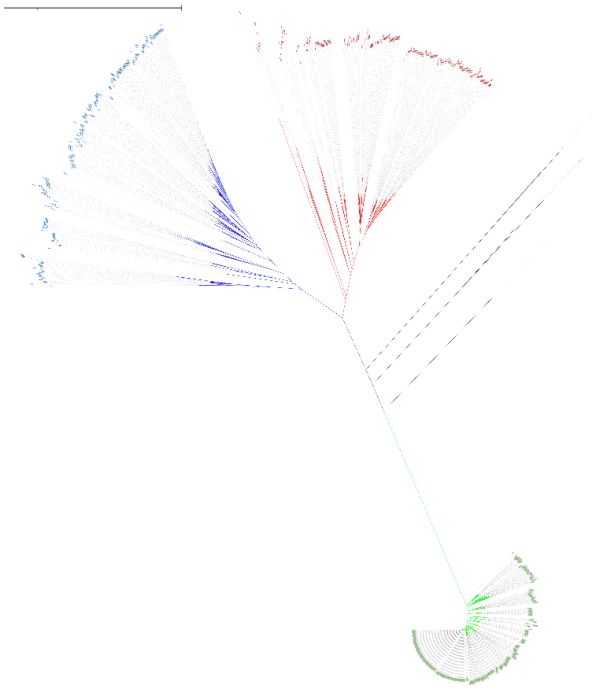


Figura 5.5. Árbol filogenético obtenido para *nor* (*unrooted*) con el parámetro *fast*.



Todas las salidas y resultados del Árbol consenso se encuentran adjuntas a esta memoria (**Anexo 6**).

5.3 Determinación de las secuencias para cada grupo

Como ya se ha comentado, los principales resultados de esta actividad son archivos que se encuentran anexadas. El Grupo Azul estaba formada por 299 secuencias, el Grupo Rojo estaba formada por 264 secuencias y el Grupo Verde por 969.

Por otro lado, en la Tabla 3.1. se muestran los resultados obtenidos en la comparativa, el número de organismos para dicho Orden y Familia en cada grupo. En color y en negrita se indica el Orden, en blanco las Familias con representación para el Orden inmediatamente superior. En esta tabla no se muestran las Familias de todos los Órdenes, únicamente se muestran aquellas en las que se ha observado que hay organismos de un mismo Orden en más de un grupo.

Tabla 5.4. Comparativa de los organismos de cada Grupo.

Organismo	AZUL	ROJO	VERDE	Total general
-		**1**		**1**
Burkholderiales	**2**			**2**
Actinomycetales	**1**			**1**
Aeromonadales			**47**	**47**

Tabla 5.4. Comparativa de los organismos de cada Grupo (continuación).

Alteromonadales	**13**	**4**	**29**	**46**
Alteromonadaceae	2		1	3
Colwelliaceae		3		3
Ferrimonadaceae			4	4
Idiomarinaceae	3			3
Moritellaceae			2	2
Psychromonadaceae			1	1
Shewanellaceae	8	1	21	30
Bacillales	**29**			**29**
Bacteroidales	**1**			**1**
Burkholderiales	**46**	**5**	**1**	**52**
-		2		2
Alcaligenaceae	10			10
Burkholderiaceae	23		1	24
Comamonadaceae	4	2		6
Oxalobacteraceae	9	1		10
Campylobacterales	**2**	**9**		**11**
Campylobacteraceae	2			2
Hydrogenimonaceae		3		3
Sulfurovaceae		2		2
Thiovulaceae		4		4
Cardiobacteriales	**4**			**4**
Cellvibrionales	**1**			**1**
Chromatiaceae	**1**			**1**
Chromatiales		**5**		**5**
Corynebacteriales	**15**			**15**
Cytophagales	**1**	**4**		**5**
Cesiribacteraceae	1			1
Cyclobacteriaceae		2		2
Cyclobacteriaceae;		1		1
Marivirgaceae		1		1
Deinococci		**1**		**1**
Desulfovibrionales	**2**			**2**
Desulfuromonadales		**2**		**2**
Enterobacterales	**1**		**783**	**784**
Eubacteriales		**4**		**4**
Flavobacteriales	**21**	**4**		**25**
Flavobacteriaceae	14	4		18
Flavobacteriia	5			5
Weeksellaceae	2			2
Gemmatales	**1**			**1**

Tabla 5.4. Comparativa de los organismos de cada Grupo (continuación).

Hyphomicrobiales	**4**	**50**		**54**
Beijerinckiaceae		1		1
Boseaceae		2		2
Bradyrhizobiaceae		8		8
Brucellaceae		1		1
Hyphomicrobiaceae		2		2
Methylobacteriaceae		1		1
Phyllobacteriaceae		1		1
Rhizobiaceae		24		24
Rhodobiaceae	1			1
Stappiaceae	3	9		12
Xanthobacteraceae		1		1
Isosphaerales	**3**			**3**
Lactobacillales	**2**			**2**
Legionellales	**28**			**28**
Leptospirales		**2**		**2**
Magnetococcales		**1**		**1**
Methylococcales		**3**		**3**
Micrococcales	**2**			**2**
Moraxellales	**9**			**9**
Myxococcales	**4**	**1**		**5**
-	1			1
Cystobacterineae		1		1
Sorangiineae	3			3
Neisseriales	**33**	**2**		**35**
Chromobacteriaceae	2	2		4
Neisseriaceae	31			31
Nitrosomonadales		**7**		**7**
Nitrosomonadales;		**1**		**1**
Oceanospirillales	**1**	**5**	**1**	**7**
Halomonadaceae		5	1	6
Saccharospirillaceae	1			1
Oscillatoriales	**1**			**1**
Parachlamydiales	**1**			**1**
Pasteurellales	**12**			**12**
Pirellulales	**5**			**5**
Planctomycetales	**12**			**12**
Propionibacteriales	**5**			**5**
Pseudomonadales	**5**	**63**	**24**	**92**
Marinobacteraceae	2	5		7
Neisseriaceae	1			1
Pseudomonadaceae	2	58	24	84
Rhodobacterales		**54**	**1**	**55**

Tabla 5.4. Comparativa de los organismos de cada Grupo (continuación).

Rhodocyclales		**3**		**3**
Rhodospirillales	**3**	**9**		**12**
-	1			1
Acetobacteraceae	1			1
Azospirillaceae		2		2
Rhodospirillaceae	1	5		6
Thalassospiraceae		2		2
Salinisphaerales	**2**	**1**		**3**
Salinisphaeraceae	2	1		3
Selenomonadales		**2**		**2**
Sphingomonadales	**5**			**5**
Syntrophorhabdales	**1**			**1**
Thermales		**1**		**1**
Thermoanaerobacterales		**2**		**2**
Thiotrichales		**2**		**2**
Veillonellales	**2**			**2**
Vibrionales	**2**		**83**	**85**
Xanthomonadales	**1**			**1**
(en blanco)	**15**	**16**		**31**

5.4 Obtención de perfiles

Como resultado del proceso descrito, se obtuvieron tres perfiles para cada una de las clases de *nor* observadas. Estos perfiles pueden verse en los **Anexos 25, 26 y 27.**

Por otro lado, también se construyeron los perfiles con otros parámetros, para determinar qué parámetros eran más adecuados en un futuro, sin embargo, como la validación de éstos requería de mucho tiempo, no fue viable la validación de todos de forma sistemática, y únicamente se optó por el descrito anteriormente.

5.5 Validación de los perfiles

En la Tabla 5.5., 5.6. y 5.7. se muestran las coincidencias al utilizar el perfil de qNOR+qCuNOR, cNOR y vNOR, respectivamente en la base de datos *Swiss-Prot.* Las salidas completas se pueden ver en los **Anexos 28, 29 y 30**, adjuntos a esta memoria.

Tabla 5.5. Resultados obtenidos mediante *pfsearch* qNOR+qCuNOR vs *Swiss-Prot.*

Score	Descripción
49.550	Nitric oxide reductase subunit B OS=Pseudomonas stutzeri
9.490	Uncharacterized protein YqgC OS=Bacillus subtilis (strain 168)
13.650	Cytochrome c oxidase subunit 1 OS=Rhodobacter capsulatus
14.990	Cytochrome c oxidase subunit 1 homolog OS=Agrobacterium tumefaciens (strain T37)
15.990	Cytochrome c oxidase subunit 1 homolog, bacteroid OS=Rhizobium meliloti
12.290	Cytochrome c oxidase subunit 1 homolog OS=Azorhizobium caulinodans (strain ATCC 43989 / DSM 5975 / JCM 20966 / LMG 6465 / NBRC 14845 / NCIMB 13405 / ORS 571)
12.990	Cytochrome c oxidase subunit 1 homolog, bacteroid OS=Bradyrhizobium diazoefficiens (strain JCM 10833 / BCRC 13528 / IAM 13628 / NBRC 14792)
11.260	Cbb3-type cytochrome c oxidase subunit CcoN1 OS=Pseudomonas stutzeri
52.100	Nitric oxide reductase subunit B OS=Pseudomonas aeruginosa (strain ATCC 15692 / DSM 22644 / CIP 104116 / JCM 14847 / LMG 12228 / 1C / PRS 101 / PAO1)

Tabla 5.6. Resultados obtenidos mediante *pfsearch* cNOR vs *Swiss-Prot.*

Score	Descripción
8.800	Probable quinol oxidase subunit 1 OS=Staphylococcus saprophyticus subsp. saprophyticus (strain ATCC 15305 / DSM 20229 / NCIMB 8711 / NCTC 7292 / S-41)
150.970	PSEST Nitric oxide reductase subunit B OS=Pseudomonas stutzeri
12.570	Cytochrome c oxidase polypeptide I+III OS=Thermus thermophilus (strain ATCC 27634 / DSM 579 / HB8)
9.000	Probable cytochrome c oxidase subunit 1-alpha OS=Streptomyces avermitilis (strain ATCC 31267 / DSM 46492 / JCM 5070 / NBRC 14893 / NCIMB 12804 / NRRL 8165 / MA-4680)
11.200	Cytochrome c oxidase subunit 1 OS=Alkalihalobacillus pseudofirmus (strain ATCC BAA-2126 / JCM 17055 / OF4)
23.170	Cytochrome c oxidase subunit 1 OS=Rhodobacter capsulatus
28.160	Cytochrome c oxidase subunit 1 homolog OS=Agrobacterium tumefaciens (strain T37)
27.310	Cytochrome c oxidase subunit 1 homolog, bacteroid OS=Rhizobium meliloti (strain 1021)
22.960	Cytochrome c oxidase subunit 1 homolog OS=Azorhizobium caulinodans (strain ATCC 43989 / DSM 5975 / JCM 20966 / LMG 6465 / NBRC 14845 / NCIMB 13405 / ORS 571)
21.510	Cytochrome c oxidase subunit 1 homolog, bacteroid OS=Bradyrhizobium diazoefficiens (strain JCM 10833 / BCRC 13528 / IAM 13628 / NBRC 14792 / USDA 110)
9.370	Probable cytochrome c oxidase subunit 1-alpha OS=Streptomyces coelicolor (strain ATCC BAA-471 / A3(2) / M145)
24.130	Cbb3-type cytochrome c oxidase subunit CcoN1 OS=Pseudomonas stutzeri
9.330	Cytochrome c oxidase subunit 1 (Fragment) OS=Thermus thermophilus
9.340	Cytochrome c oxidase subunit 1 OS=Synechocystis sp. (strain PCC 6803 / Kazusa)
17.860	Cytochrome c oxidase subunit 1 OS=Thermus thermophilus
150.170	Nitric oxide reductase subunit B OS=Pseudomonas aeruginosa (strain ATCC 15692 / DSM 22644 / CIP 104116 / JCM 14847 / LMG 12228 / 1C / PRS 101 / PAO1)

Tabla 5.7. Resultados obtenidos mediante *pfsearch* vNOR vs *Swiss-Prot* (únicamente resultados con *Score* > 20).

Score	Descripción
72.570	Rubredoxin-oxygen oxidoreductase OS=Desulfovibrio gigas (strain ATCC 19364 / DSM 1382 / NCIMB 9332 / VKM B-1759)
60.500	Putative diflavin flavoprotein A 4 OS=Nostoc sp. (strain PCC 7120 / SAG 25.82 / UTEX 2576)
58.830	Putative diflavin flavoprotein A 5 OS=Nostoc sp. (strain PCC 7120 / SAG 25.82 / UTEX 2576)
57.920	Putative diflavin flavoprotein A 2 OS=Thermosynechococcus elongatus (strain BP-1)
58.440	Putative diflavin flavoprotein A 3 OS=Synechocystis sp. (strain PCC 6803 / Kazusa)
55.880	Diflavin flavoprotein A 1 OS=Synechocystis sp. (strain PCC 6803 / Kazusa)
56.880	Type A flavoprotein fprA OS=Rhodobacter capsulatus
22.490	High molecular weight rubredoxin OS=Moorella thermoacetica (strain ATCC 39073 / JCM 9320)
103.500	Nitric oxide reductase OS=Moorella thermoacetica (strain ATCC 39073 / JCM 9320)
57.790	Type A flavoprotein fprA OS=Rhodobacter capsulatus (strain ATCC BAA-309 / NBRC 16581 / SB1003)
55.490	Putative diflavin flavoprotein A 2 OS=Nostoc sp. (strain PCC 7120 / SAG 25.82 / UTEX 2576)
58.890	Putative diflavin flavoprotein A 4 OS=Synechocystis sp. (strain PCC 6803 / Kazusa)
59.980	Putative diflavin flavoprotein A 6 OS=Nostoc sp. (strain PCC 7120 / SAG 25.82 / UTEX 2576)
249.920	Anaerobic nitric oxide reductase flavorubredoxin OS=Aeromonas hydrophila subsp. hydrophila (strain ATCC 7966 / DSM 30187 / BCRC 13018 / CCUG 14551 / JCM 1027 / KCTC 2358 / NCIMB 9240 / NCTC 8049)
249.680	Anaerobic nitric oxide reductase flavorubredoxin OS=Aeromonas salmonicida (strain A449)
240.500	Anaerobic nitric oxide reductase flavorubredoxin OS=Aliivibrio fischeri (strain ATCC 700601 / ES114)
240.800	Anaerobic nitric oxide reductase flavorubredoxin OS=Aliivibrio salmonicida (strain LFI1238)
251.650	Anaerobic nitric oxide reductase flavorubredoxin OS=Citrobacter koseri (strain ATCC BAA-895 / CDC 4225-83 / SGSC4696)
251.140	Anaerobic nitric oxide reductase flavorubredoxin OS=Escherichia coli O139:H28 (strain E24377A / ETEC
200.190	Anaerobic nitric oxide reductase flavorubredoxin homolog OS=Escherichia coli O157:H7
250.840	Anaerobic nitric oxide reductase flavorubredoxin OS=Escherichia coli (strain K12 / DH10B)
251.230	Anaerobic nitric oxide reductase flavorubredoxin OS=Escherichia coli O9:H4 (strain HS)
250.600	Anaerobic nitric oxide reductase flavorubredoxin OS=Escherichia coli O1:K1 / APEC
250.600	Anaerobic nitric oxide reductase flavorubredoxin OS=Escherichia coli O6:K15:H31 (strain 536 / UPEC)
251.370	Anaerobic nitric oxide reductase flavorubredoxin OS=Escherichia coli O6:H1 (strain CFT073 / ATCC 700928 / UPEC)

Score	Descripción
250.840	Anaerobic nitric oxide reductase flavorubredoxin OS=Escherichia coli (strain ATCC 8739 / DSM 1576 / NBRC 3972 / NCIMB 8545 / WDCM 00012 / Crooks)
250.840	Anaerobic nitric oxide reductase flavorubredoxin OS=Escherichia coli (strain K12)
251.050	Anaerobic nitric oxide reductase flavorubredoxin OS=Escherichia coli (strain SMS-3-5 / SECEC)
250.600	Anaerobic nitric oxide reductase flavorubredoxin OS=Escherichia coli (strain UTI89 / UPEC)
251.390	Anaerobic nitric oxide reductase flavorubredoxin OS=Enterobacter sp. (strain 638)
250.110	Anaerobic nitric oxide reductase flavorubredoxin OS=Klebsiella pneumoniae subsp. pneumoniae (strain ATCC 700721 / MGH 78578)
250.010	Anaerobic nitric oxide reductase flavorubredoxin OS=Pectobacterium atrosepticum (strain SCRI 1043 / ATCC BAA-672
249.670	Anaerobic nitric oxide reductase flavorubredoxin OS=Pectobacterium carotovorum subsp. carotovorum (strain PC1)
251.670	Anaerobic nitric oxide reductase flavorubredoxin OS=Salmonella arizonae (strain ATCC BAA-731 / CDC346-86 / RSK2980
252.570	Anaerobic nitric oxide reductase flavorubredoxin OS=Salmonella choleraesuis (strain SC-B67)
252.100	Anaerobic nitric oxide reductase flavorubredoxin OS=Salmonella paratyphi A (strain ATCC 9150 / SARB42)
252.740	Anaerobic nitric oxide reductase flavorubredoxin OS=Salmonella paratyphi B (strain ATCC BAA-1250 / SPB7)
252.580	Anaerobic nitric oxide reductase flavorubredoxin OS=Salmonella typhi
252.210	Anaerobic nitric oxide reductase flavorubredoxin OS=Salmonella typhimurium (strain LT2 / SGSC1412 / ATCC 700720)
250.160	Anaerobic nitric oxide reductase flavorubredoxin OS=Serratia proteamaculans (strain 568)
251.120	Anaerobic nitric oxide reductase flavorubredoxin OS=Shigella boydii serotype 4 (strain Sb227)
250.750	Anaerobic nitric oxide reductase flavorubredoxin OS=Shigella dysenteriae serotype 1 (strain Sd197)
250.920	Anaerobic nitric oxide reductase flavorubredoxin OS=Shigella flexneri serotype 5b (strain 8401)
250.920	Anaerobic nitric oxide reductase flavorubredoxin OS=Shigella flexneri
251.230	Anaerobic nitric oxide reductase flavorubredoxin OS=Shigella sonnei (strain Ss046)
240.090	Anaerobic nitric oxide reductase flavorubredoxin OS=Vibrio vulnificus (strain CMCP6)
239.870	Anaerobic nitric oxide reductase flavorubredoxin OS=Vibrio vulnificus (strain YJ016)
55.560	Putative diflavin flavoprotein A 2 OS=Synechocystis sp. (strain PCC 6803 / Kazusa)
54.020	Putative diflavin flavoprotein A 3 OS=Nostoc sp. (strain PCC 7120 / SAG 25.82 / UTEX 2576)
49.150	Putative diflavin flavoprotein A 1 OS=Nostoc sp. (strain PCC 7120 / SAG 25.82 / UTEX 2576)
48.280	Putative diflavin flavoprotein A 1 OS=Thermosynechococcus elongatus (strain BP-1)
54.170	Flavo-diiron protein FprA1 OS=Clostridium acetobutylicum (strain ATCC 824 / DSM 792 / JCM 1419 / LMG 5710 / VKM B-1787)
60.810	Flavo-diiron protein FprA2 OS=Clostridium acetobutylicum (strain ATCC 824 / DSM 792 / JCM 1419 / LMG 5710 / VKM B-1787)

En la Tabla 5.8. se pueden ver los resultados (*matches* y *scores*) obtenidos al utilizar los diferentes perfiles en cada uno de los genomas descargados. La totalidad de las salidas pueden encontrase en los **Anexos 31's**.

Tabla 5.8. M*atches* y *scores* obtenidos al utilizar los diferentes perfiles en cada uno de los genomas descargados.

matches *scores*	***Paracoccus denitrificans***	***Pyrobaculum aerophilum***	***Bacillus azotoformans***	***E.coli***
qNOR+qCuNOR	*3* *<50.000*	*2* *9.230 y 85.060*	*2* *221.00 y 196.930*	*0*
cNOR	*3* *<151.000*	*2* *15.940 y 42.290*	*5* *<56.000*	*0*
vNOR	*191* *<11.000*	*58* *<11.000*	*42* *<11.000*	*1* *250.840*

Los resultados obtenidos al comparar los perfiles con la base de datos con proteínas hipotéticas se pueden encontrar en los **Anexos 32, 33 y 34**. En la Tabla 5.9., 5.10. y 5.11. se muestran los *matches* para los perfiles qNOR+qCuNOR, cNOR y vNOR.

Tabla 5.9. M*atches* de proteínas hipotéticas para el perfil qNOR+qCuNOR.

Organismo	***Matches* (n)**
Verrucomicrobia bacterium	74
Actinomycetia bacterium	2
Acidobacteria bacterium	2
Candidatus Rokubacteria bacterium	2
Alphaproteobacteria bacterium	1
Chloroflexi bacterium	1
Acidobacteria bacterium 13_1_40CM_4_58_4	1
Acidobacteria bacterium 13_1_40CM_4_57_6	1
Verrucomicrobia bacterium 13_2_20CM_55_10	1

Tabla 5.10. M*atches* de proteínas hipotéticas para el perfil cNOR.

Organismo	*Matches* (n)
Verrucomicrobia bacterium	111
Chloroflexi bacterium	14
Actinomycetia bacterium	10
Gemmatimonadetes bacterium	4
Acidobacteria bacterium 13_1_40CM_4_58_4	2
Candidatus Dormibacter sp. RRmetagenome_bin12	1
Alphaproteobacteria bacterium	1
Nitrospirae bacterium	1
Candidatus Dormibacteraeota bacterium	1
Chthoniobacterales bacterium	1
Acidobacteria bacterium	1
Chloroflexi bacterium 13_1_20CM_2_59_7	1
Actinobacteria bacterium 13_1_20CM_4_69_9	1
Candidatus Rokubacteria bacterium 13_1_40CM_2_68_8	1
Armatimonadetes bacterium 13_1_40CM_3_65_7	1
Acidobacteria bacterium 13_1_40CM_4_57_6	1
Verrucomicrobia bacterium 13_2_20CM_55_10	1

Tabla 5.11. M*atches* de proteínas hipotéticas para el perfil vNOR.

Organismo	*Matches* (n)
Acidobacteria bacterium	11595
Verrucomicrobia bacterium	4780
Chloroflexi bacterium	4388
Candidatus Rokubacteria bacterium	3685
Gemmatimonadetes bacterium	3203
Deltaproteobacteria bacterium	2196
Actinomycetia bacterium	1759
Gammaproteobacteria bacterium	693
Betaproteobacteria bacterium	533
Alphaproteobacteria bacterium	416
Candidatus Eisenbacteria bacterium	378
Pseudonocardiales bacterium	307
Nitrospirae bacterium	220
Actinobacteria bacterium 13_2_20CM_2_66_6	172
Acidobacteriia bacterium	161
Acidobacteria bacterium 13_2_20CM_58_27	159
Candidatus Dormibacteraeota bacterium	149
Gemmatimonadetes bacterium 13_1_40CM_4_69_8	148
Acidobacteria bacterium 13_1_40CM_65_14	131
Actinobacteria bacterium 13_2_20CM_2_71_6	119
Terrabacteria group bacterium ANGP1	110
Catenulispora sp. 13_1_20CM_3_70_7	109
Candidatus Rokubacteria bacterium 13_1_40CM_68_15	100
Acidobacteria bacterium 13_1_40CM_4_65_8	97

Organismo	*Matches* (n)
Candidatus Rokubacteria bacterium 13_1_40CM_4_67_11	97
Acidobacteria bacterium 13_1_20CM_2_65_9	96
Chthoniobacterales bacterium	93
Acidobacteria bacterium 13_1_40CM_2_68_5	92
Candidatus Rokubacteria bacterium 13_1_40CM_69_27	92
Blastocatellia bacterium AA13	91
Candidatus Rokubacteria bacterium 13_1_20CM_4_68_9	91
Candidatus Rokubacteria bacterium 13_1_40CM_4_69_5	91
Solirubrobacterales bacterium	90
Blastocatellia bacterium	89
Bacteroidetes bacterium	87
Actinobacteria bacterium 13_1_20CM_3_71_11	86
Acidobacteria bacterium 13_1_40CM_4_58_4	84
Ktedonobacter sp. 13_2_20CM_53_11	84
Phenylobacterium sp.	82
Deltaproteobacteria bacterium 13_1_40CM_4_68_19	81
Candidatus Rokubacteria bacterium 13_2_20CM_2_64_8	80
Deltaproteobacteria bacterium 13_1_20CM_2_69_21	79
Acidobacteriia bacterium AA117	77
Acidobacteria bacterium 13_2_20CM_2_57_6	76
Alphaproteobacteria bacterium 13_2_20CM_2_64_7	76
Cyanobacteria bacterium 13_1_40CM_2_61_4	74
Gemmatimonadetes bacterium 13_2_20CM_2_69_23	74
Acidobacteria bacterium 13_1_20CM_3_53_8	73
Candidatus Rokubacteria bacterium 13_1_20CM_4_70_14	71
Acidobacteria bacterium 13_1_40CM_3_65_5	71
Chloroflexi bacterium 13_1_40CM_2_68_14	69
Candidatus Rokubacteria bacterium 13_1_40CM_2_68_8	69
Acidobacteria bacterium 13_1_40CM_56_16	69
Candidatus Rokubacteria bacterium 13_2_20CM_2_70_11	69
Proteobacteria bacterium	68
Acidobacteria bacterium 13_1_20CM_58_21	66
Deltaproteobacteria bacterium 13_1_40CM_3_69_14	66
Actinobacteria bacterium 13_2_20CM_2_72_6	66
Acidobacteria bacterium 13_2_20CM_57_17	65
Acidimicrobiales bacterium	63
Candidatus Melainabacteria bacterium	63
Gemmatimonadetes bacterium 13_1_40CM_69_22	63
Candidatus Rokubacteria bacterium 13_2_20CM_69_15_1	63
Gemmatimonadetes bacterium 13_2_20CM_69_27	62
Planctomycetes bacterium	60
Candidatus Dormibacter sp. RRmetagenome_bin12	59
Acidobacteria bacterium 13_1_20CM_2_60_10	59
Candidatus Rokubacteria bacterium 13_1_20CM_2_68_19	59
Acidobacteria bacterium 13_2_20CM_2_66_4	59

Organismo	*Matches* (n)
Ktedonobacter sp. 13_1_20CM_3_54_15	58
Gemmatimonadetes bacterium 13_1_20CM_4_69_16	58
Gemmatimonadetes bacterium 13_1_20CM_69_28	58
Gemmatimonadetes bacterium 13_1_40CM_66_11	58
Gemmatimonadetes bacterium 13_2_20CM_2_65_7	58
Chloroflexi bacterium 13_1_40CM_4_68_4	57
Gemmatimonadetes bacterium 13_2_20CM_70_9	57
Acidobacteria bacterium 13_1_20CM_2_68_14	56
Gemmatimonadetes bacterium 13_1_20CM_4_66_11	56
Ktedonobacter sp. 13_2_20CM_2_56_8	56
Gammaproteobacteria bacterium 13_2_20CM_66_19	55
Candidatus Rokubacteria bacterium 13_1_20CM_2_70_7	54
Acidobacteria bacterium 13_1_20CM_2_55_15	53
Actinobacteria bacterium 13_1_40CM_4_65_12	52
Chloroflexi bacterium 13_1_40CM_65_17	51
Actinobacteria bacterium 13_1_20CM_2_65_11	50
Candidatus Rokubacteria bacterium 13_1_40CM_4_69_39	50
Candidatus Rokubacteria bacterium 13_2_20CM_69_10	50
Acidobacteria bacterium 13_1_40CM_2_60_7	49
Chloroflexi bacterium 13_1_40CM_3_65_12	48
Cyanobacteria bacterium 13_1_20CM_4_61_6	47
Actinobacteria bacterium 13_1_40CM_2_65_8	47
Gemmatimonadetes bacterium 13_1_40CM_2_70_7	47
Candidatus Rokubacteria bacterium 13_1_40CM_2_68_13	46
Gemmatimonadetes bacterium 13_1_40CM_3_70_6	46
Acidobacteria bacterium 13_2_20CM_57_7	46
Acidobacteria bacterium 13_1_40CM_3_56_11	45
Candidatus Rokubacteria bacterium 13_1_40CM_3_69_38	45
Gemmatimonadetes bacterium 13_1_40CM_4_65_7	45
Verrucomicrobia bacterium 13_2_20CM_54_12	45
Verrucomicrobia bacterium 13_2_20CM_55_10	45
Candidatus Rokubacteria bacterium 13_1_20CM_2_69_58	44
Acidobacteria bacterium 13_1_20CM_3_58_11	44
Ktedonobacter sp. 13_1_20CM_4_53_11	44
Acidobacteria bacterium 13_1_40CM_2_56_11	44
Chloroflexi bacterium 13_1_40CM_2_70_6	44
Gemmatimonadetes bacterium 13_1_40CM_3_69_22	43
Gemmatimonadetes bacterium 13_1_40CM_3_66_12	43
Actinobacteria bacterium 13_1_20CM_3_68_9	42
Armatimonadetes bacterium 13_1_40CM_3_65_7	42
Gemmatimonadetes bacterium 13_1_40CM_70_11	42
Candidatus Eremiobacter sp. RRmetagenome_bin22	41
Bacteroidetes bacterium 13_1_20CM_4_60_6	41
Acidobacteria bacterium 13_1_20CM_2_68_7	41
Acidobacteria bacterium 13_1_40CM_2_68_10	41

Organismo	*Matches* (n)
Acidobacteria bacterium 13_1_40CM_4_69_4	41
Bdellovibrio sp.	40
Delftia sp. 13_1_20CM_4_67_18	39
Actinobacteria bacterium 13_1_20CM_4_69_9	39
Actinobacteria bacterium 13_1_40CM_66_12	39
Verrucomicrobia bacterium 13_1_20CM_4_54_11	38
Chloroflexi bacterium 13_1_20CM_2_70_9	37
Gemmatimonadetes bacterium 13_1_40CM_3_65_8	37
Chloroflexi bacterium 13_1_40CM_4_65_16	37
Gemmatimonadetes bacterium 13_2_20CM_69_8	37
Bradyrhizobium sp.	36
Actinobacteria bacterium 13_1_20CM_4_68_12	36
Verrucomicrobia bacterium 13_1_20CM_3_54_17	36
Nitrospirae bacterium 13_1_40CM_2_62_10	36
Acidobacteria bacterium 13_1_20CM_4_56_7	35
Gemmatimonadetes bacterium 13_1_40CM_70_15	35
Candidatus Rokubacteria bacterium 13_2_20CM_70_12	35
Gemmatimonadetes bacterium 13_1_20CM_2_70_10	34
Verrucomicrobia bacterium 13_1_20CM_54_28	34
Candidatus Rokubacteria bacterium 13_1_40CM_2_70_45	34
Armatimonadetes bacterium 13_1_40CM_64_14	34
Acidobacteria bacterium 13_2_20CM_2_57_12	34
Chromatiales bacterium USCg_Taylor	33
Chloroflexi bacterium 13_1_20CM_54_36	33
Gemmatimonadetes bacterium 13_2_20CM_2_66_5	33
Actinobacteria bacterium 13_2_20CM_68_14	33
Acidobacteriales bacterium 13_2_20CM_55_8	33
Actinobacteria bacterium 13_1_20CM_2_66_18	31
Ignavibacteria bacterium 13_1_40CM_2_61_4	31
Chloroflexi bacterium 13_1_40CM_68_21	31
Chloroflexi bacterium 13_1_40CM_66_19	31
Nitrospirae bacterium 13_2_20CM_2_62_8	31
Beijerinckiaceae bacterium	30
Acidobacteria bacterium 13_1_20CM_4_57_11	30
Acidobacteriales bacterium 13_1_40CM_3_55_5	30
Acidobacteria bacterium 13_1_40CM_4_61_5	30
Acidobacteria bacterium 13_1_40CM_3_55_6	28
Nitrospirae bacterium 13_2_20CM_62_7	28
Candidatus Udaeobacter sp.	27
Deltaproteobacteria bacterium 13_1_40CM_68_24	27
Verrucomicrobia bacterium 13_2_20CM_2_54_15	27
Delftia sp. 13_1_40CM_3_66_6	26
Chloroflexi bacterium 13_1_20CM_2_59_7	26
Nitrospirae bacterium 13_1_40CM_4_62_6	26
Candidatus Rokubacteria bacterium 13_1_40CM_69_96	26

Organismo	*Matches* (n)
Betaproteobacteria bacterium 13_1_20CM_3_63_8	25
Alphaproteobacteria bacterium 13_1_20CM_3_64_12	24
Actinobacteria bacterium 13_1_40CM_2_66_13	24
Actinobacteria bacterium 13_1_40CM_3_66_19	24
Actinobacteria bacterium 13_1_20CM_3_68_10	23
Gemmatimonadetes bacterium 13_1_40CM_4_69_5	23
Ktedonobacter sp. 13_2_20CM_2_54_8	23
Chloroflexi bacterium 13_1_20CM_66_33	22
Nitrospirae bacterium 13_2_20CM_2_61_4	22
Candidatus Eremiobacteraeota bacterium	21
Ktedonobacter sp. 13_1_40CM_4_52_4	21
Candidatus Rokubacteria bacterium 13_2_20CM_69_15_2	21
Verrucomicrobia bacterium 13_2_20CM_2_54_15_9cls	20
Hyphomicrobiales bacterium	19
Chloroflexi bacterium 13_1_20CM_4_66_7	19
Nitrospirae bacterium 13_2_20CM_2_63_8	19
Acidobacteria bacterium 13_1_40CM_4_57_6	18
Ktedonobacter sp. 13_1_20CM_4_53_7	17
Deltaproteobacteria bacterium 13_1_40CM_4_54_4	17
Betaproteobacteria bacterium 13_1_40CM_4_64_4	17
Chloroflexi bacterium 13_1_40CM_67_9	17
Acidobacteria bacterium 13_2_20CM_56_17	17
Myxococcales bacterium 13_1_40CM_2_68_15	16
Candidatus Rokubacteria bacterium 13_1_20CM_70_15	15
Chloroflexi bacterium 13_1_40CM_4_69_19	15
Acidobacteriales bacterium 13_2_20CM_2_55_5	15
Sphingomonas sp.	14
Chloroflexi bacterium 13_1_20CM_50_12	14
Chloroflexi bacterium 13_1_40CM_55_7	14
Ignavibacteria bacterium	13
Candidatus Rokubacteria bacterium 13_1_20CM_4_70_13	13
Acidobacteria bacterium 13_1_40CM_2_64_6	13
Nitrospirae bacterium 13_1_40CM_3_62_11	13
Hydrogenibacillus schlegelii	12
Gemmatimonadetes bacterium 13_1_40CM_3_70_8	12
Nitrospirae bacterium 13_1_40CM_62_7	12
Verrucomicrobia bacterium 13_1_20CM_4_55_9	11
Acidobacteria bacterium 13_1_40CM_2_56_5	11
Chloroflexi bacterium 13_1_40CM_3_70_6	11
Chloroflexi bacterium 13_1_40CM_4_65_13	11
Verrucomicrobia bacterium 13_1_40CM_4_54_4	11
Gemmatimonadetes bacterium 13_1_40CM_2_60_3	9
Gemmatimonadetes bacterium 13_1_40CM_70_12	9
Chlamydiae bacterium	8
Nitrospirae bacterium 13_1_20CM_2_62_14	8

Organismo	*Matches* (n)
Candidatus Cerribacteria bacterium 'Amazon FNV 2010 28 9'	7
Candidatus Saccharibacteria bacterium	7
Chloroflexi bacterium 13_1_20CM_4_66_15	7
Deltaproteobacteria bacterium 13_1_40CM_3_71_4	7
Acidobacteria bacterium 13_1_20CM_2_57_8	6
Gemmatimonas sp. 13_1_20CM_3_60_15	6
Chloroflexi bacterium 13_1_40CM_68_15	6
Candidatus Carbobacillus altaicus	5
Brockia lithotrophica	5
Alphaproteobacteria bacterium 13_1_20CM_4_65_11	5
Gemmatimonadetes bacterium 13_2_20CM_1_70_33	5
Flavobacteriaceae bacterium FS1-H7996/R	3
Actinobacteria bacterium 13_1_20CM_4_66_15	3
Nitrospirae bacterium 13_1_20CM_4_62_6	3
Firmicutes bacterium 13_1_40CM_3_65_11	3
Gemmatimonadetes bacterium 13_2_20CM_1_69_27	3
Thaumarchaeota archaeon	2
Sulfurimonas sp. UBA10385	2
uncultured bacterium	2
Phage 66_12	1
Phage 67_12	1
Candidatus Bathyarchaeota archaeon	1
Armatimonadetes bacterium 13_1_20CM_4_65_7	1
Betaproteobacteria bacterium 13_1_20CM_67_22	1
uncultured Acidobacteria bacterium	1

Para el cuenteo de las tablas anteriores, se utilzó el *script* que puede verse en los Anexos de esta Memoria. Por último, mediante el *script* que puede verse en los Anexos de esta Memoria, se compararon los *matches* de proteínas hipotéticas entre grupos. Los resultados se pueden ver en la Tabla 5.12.

Tabla 5.12. Proteínas identificadas en común.

Intersección	*matches*
qNOR+qCuNOR vs cNOR	80 en común
qNOR+qCuNOR vs vNOR	1
cNOR vs vNOR	2
qNOR+qCuNOR vs cNOR vs vNOR	1

6 Discusión

6.1 Obtención de secuencias

Según la literatura disponible a día de hoy, existen tres tipos de *nor*; cNOR, qNOR y qCuNOR. La cNOR está formada por las subunidades B y C, dónde la subunidad B es la que tiene actividad catalítica. Por otro lado, la qNOR y la qCuNOR están formadas por una única subunidad. Por lo tanto, se decidió

excluir de la búsqueda los resultados de la subunidad C (*NOT norc)* ya que su inclusión podría afectar negativamente a los resultados. Por otro lado, el grupo qCuNOR es “parecido” a qNOR, y todavía se considera “poco representado”. Las proteínas flavodiiron del grupo B (norV) también tienen actividad nítrico oxido reductasa, y por ello también se consideran un tipo de *nor.*

De las secuencias obtenidas en la búsqueda en *Uniprot*, así como de las inspecciones visuales realizadas, se pudo inferir la existencia de 3 grupos *nor* en las secuencias obtenidas; cNOR, qNOR y qCuNOR (dado que son similares y las anotaciones pueden no ser precisas), y norV. Esto coincide con la información disponible en la literatura. Además, se pudo observar que la longitud de las secuencias también coincide con la información disponible, ya que se observaron secuencias de una longitud aproximada de 800 AA, secuencias con una longitud intermedia, y secuencias más cortas, que se podrían corresponder con norV, qNOR (una subunidad más grande) y cNOR (únicamente subunidad B), respectivamente.

6.2 Elaboración del árbol filogenético

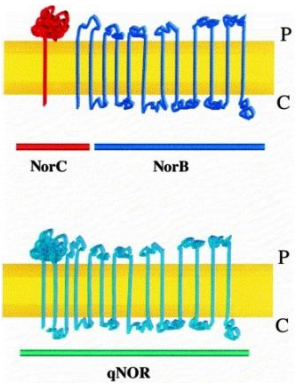
A la hora de elaborar un árbol filogenético es recomendable que que n>>K, pero en este caso, no se cumplió (n = 1.726 y K = 3.120), y esto puede dar lugar a que las estimaciones fueran erróneas [64]. Sin embargo, como el objetivo de este árbol fue clasificar las diferentes secuencias en grupos, se pudo considerar que los errores en las estimaciones no afectan significativamente al principal objetivo, en cuyo caso, aceptaríamos el árbol obtenido.

Los modelos utilizados en aminoácidos están restringidos a pares de proteínas que tienen un grado constante y pequeño de divergencia. El modelo VT obtenido en la búsqueda del mejor modelo, es una extensión del modelo *Dayhoff ,* no está sujeto a estas limitaciones, y permite estimar un modelo de sustitución de aminoácidos a partir de alineaciones de grado variable [65]. Esto concuerda con las secuencias utilizadas, donde las regiones y el grado de alineamiento era variable.

En el Árbol consenso obtenido (así como en el resto de árboles desarrollados), pudo verse claramente tres grupos principales (Azul, Rojo y Verde), lo que se corresponde con la hipótesis inicial, dónde se esperaba que una vez realizado el árbol filogenético, éste representara tres grupos (o tuviera tres ramas) principales que se corresponderían con los grupos cNOR, qNOR+qCuNOR, y vNor.

En la parte superior del árbol filogenético, se observan que los grupos Azul y Rojo están más próximos entre sí, que con el grupo Verde. Según la bibliografía, los grupos cNOR y qNOR+qCuNOR son similares (Figura 6.1.), mientras que el grupo vNor es algo mucho más distinto. Esto concuerda con los resultados obtenidos, donde los grupos Rojo y Azul están próximos y alejados del grupo Verde, suponiendo que la rama Verde se corresponda con vNor. La distancia entre los grupos Azul y Rojo es considerable, lo que pone en evidencia la gran diferencia con vNor.

Figura 6.1. Enzimas cNOR y qNOR [66].



Por otro lado, según la hipótesis inicial, en una de las ramas debería de observarse una división, que debiera corresponderse con qNOR y qCuNOR. Sin embargo, en ninguna de las tres ramas parece observarse ninguna división clara. Esto puede ser debido a dos motivos; qCuNOR no se encuentra representada en dicho árbol, es decir, “no hay” secuencias, o bien, debido a que se utilizaron secuencias de todos los tipos de *nor,* no es posible realizar una distinción precisa entre qNOR y qCuNOR.

Por último, se observaron algunas secuencias aisladas del resto que no se corresponden (*a priori*) con ningún grupo. Están pertenecen a secuencias anotadas como *Candidatus*. Sin entrar en más detalle, es difícil saber si se tratan de errores o un cuarto grupo, pero en este proyecto, dichas secuencias no se utilizaron.

6.3 Determinación de las secuencias para cada grupo

Es importante destacar que el número de secuencias en cada Grupo carece de relevancia, y no debe de interpretarse como qué estos son más abundantes; podría ser que algún tipo de organismo esté mucho más estudiado y se han encontrado más secuencias en la base de datos. Dicho esto, como puede verse, salvo en contadas excepciones, está bien definido qué organismos pertenecen a qué Grupo. Por ejemplo, las secuencias utilizadas en este estudio de los organismos del Orden *Bacillales* pertenecen al grupo Azul, así como los del Orden *Legionellales*, mientras que los del Orden *Enterobacterales* al grupo Verde. Igualmente sucede con las familias. Sin embargo, hay situaciones dónde dicha división no queda clara, como sucede con el Orden *Pseudomonadales*.

Según la literatura, la cNOR se ha observado en varios organismos desnitrificantes, como *Paracoccus denitrificans* [23,24], *Halomonas halodenitrificans* [25-26], *Pseudomonas nautica* [27], *Pseudomonas stutzeri* [28,29] y *Pseudomonas aeruginosa* [30]. Por otro lado, se han observado qNOR en organismos como *Wautersia eutropha* [31] y *Pyrobaculum aerophilim* [32]. La tercera clase de *nor* (qCuNOR) se ha observado en la bacteria gram-positiva *Bacillus azotoformans* [33]. Los FDP de clase B (vNor) están restringidos al filo

de Proteobacteria, de las clases beta, delta y gamma, éstas han sido estudiada en el microorganismo *E. coli* [35,45].

Con estas referencias y los resultados obtenidos, queda bastante claro que el grupo Rojo se corresponde con cNOR, el grupo Azul con qNOR+qCuNOR y el grupo Verde con vNor. Durante el próximo Hito, se realizará un árbol filogenético del grupo Azul para determinar si es posible discriminar entre qNOR y qCuNOR, con el objetivo de obtener una secuencia consenso para cada tipo.

6.4 Obtención de perfiles

Los perfiles generalizados son una extensión de las matrices de puntuación específicas de posición, ya que incluyen puntuaciones específicas para inserciones y eliminaciones. Corresponden a una representación matricial de un alineamiento múltiple que se puede utilizar para buscar secuencias homólogas. Como resultado, se han obtenido tres perfiles para la búsqueda de tres clases de *nor*; cNOR, qNOR y qCuNOR y vNOR.

En este punto, quizás los aspectos más importantes son los parámetros utilizados para la construcción de los perfiles. Se utilizó la matriz BLOSUM (BLOcks of Amino Acid SUbstitution Matrix, o matriz de sustitución de bloques de aminoácidos) ya que se trata de una matriz de sustitución utilizada para el alineamiento de secuencias de proteínas [67]. Por otro lado, existen diferentes tipos de matrices *blosum* (indicadas por un número). Por ejemplo, *blosum80* se usa para alineamientos menos divergentes, mientras que *blosum45* se usa para alineamientos más divergentes. Las puntuaciones dentro de una matriz *blosum* corresponden a log-probabilidades que reflejan, en un alineamiento, el logaritmo de la razón de la probabilidad de la aparición de dos aminoácidos de una forma biológicamente intencionada o aceptada (residuos homológos; este numerador es la probabilidad de la hipótesis que queremos contrastar) y la probabilidad de su aparición por casualidad (el denominador es la probabilidad de la hipótesis nula; [68]). Las matrices se basan en el mínimo porcentaje de identidad de la secuencia de proteína alineada usada al calcularlas (por ejemplo. *blosum*45 correspondería a alineamientos con un mínimo de un 45% de identidad; [69]). A cada posible identidad o sustitución se le asigna una puntuación basada en las frecuencias observadas en el alineamiento de proteínas relacionadas. Se da una puntuación positiva a las sustituciones más probables, mientras que corresponde una puntuación negativa para sustituciones menos probables.

A raíz de la observación de las secuencias obtenidas, y para utilizar la misma matriz de substitución para los tres perfiles, y facilitar la operación, se utilizó la matriz *blosum45*. Por otro lado, a la hora de seleccionar los parámetros para está matriz, como se indica en la documentación que se puede consultar en:

https://manpages.debian.org/testing/pftools/pfmake.1.en.html

los parámetros por defecto funcionan de forma adecuado cuando se utiliza la matriz *blosum45.* Por esta razón, se utilizaron los parámetros por defecto.

6.5 Validación de los perfiles

Como puede verse al utilizar el perfil obtenido para qNOR+qCuNOR, las secuencias obtenidas en la base de datos *SwissProt* atienden a los términos *reductase subunit B* y a *c oxidase subunit*, pero únicamente las subunidades B obtienen puntuaciones altas. Al utilizar el perfil para cNOR se obtienen resultados similares, pero más cuantiosos. Esto puede ser debido a las similitudes entre ambos tipos de *nor* donde em ambos casos tienen una subunidad B. Por el contrario, para la proteína vNOR se obtienen resultados mucho más específicos y, a primera vista, precisos (términos como *flavoprotein, nitric oxide reductase,* etc.). Parece ser, que en este caso, los perfiles has sido capaces de discriminar entre qNOR+qCuNOR+cNOR vs vNOR, dado que las diferencias entre éstas son mayores.

Por otro lado, como se ha visto al utilizar los diferentes perfiles en diferentes genomas donde el tipo de *nor* está bien documentado, sí se ha observado una gran capacidad para discriminar. Para *Paracoccus denitrificans*, dónde se ha encontrado la *nor* tipo cNOR, observamos que al utilizar el perfil para cNOR, se obtiene un *match* con una elevada puntuación, mientras que cuando se utilizan otros perfiles, se obtienen los mismos o más *matchs* pero con *scores* muchos más bajas. Lo mismo se observa en *Pyrobaculum aerophilum* y *Bacillus azotoformans* dónde se han detectado *nor* tipo qNOR y qCuNOR, en este caso, al utilizar el perfil qNOR+qCuNOR se obtienen los *matches* con mayor puntuación. Por último, en cuanto al perfil vNOR utilizado en *E.coli*, se observa claramente la capacidad de discriminar de ésta.

Por último, al buscar *matches* en una base de datos de proteínas hipotéticas del metagenoma del suelo, con los perfiles cNOR y qNOR+qCuNOR se observan que los organismos de las proteínas detectadas son similares. En el caso de cNOR, se observa que ha detectado un mayor número de proteínas de *Chloroflexi bacterium* y *Actinomycetia bacterium*, en comparación. En cuanto a vNOR, se ve claramente que es capaz de detectar la proteína, y que los microorganismos detectados se corresponden con lo descrito en la bibliografía. Sin embargo, gran parte de los muchos *matches* observados (tanto con *SwissProt* como con las proteínas hipotéticas) tienen puntuaciones pequeñas, por lo que habría que utilizar en este caso, valores de corte grandes.

A partir de los resultados obtenidos, los perfiles pueden discriminar entre vNOR y el resto de *nor*. En cuanto a cNOR vs qNOR+qCuNOR, parece observarse cierta capacidad discriminante si se entra en detalle (comparando puntuaciones de los *matches*), lo que sugiere que sería posible discriminar entre estos tipos si se realiza posteriormente una inspección visual. En algunos casos, esto puede ser adecuado, pero si se busca automatizar un proceso, es muy probable que estos perfiles no fueran suficientes.

7 Conclusiones

7.1 Conclusiones

En primer lugar, en este proyecto se han cumplido con todos los objetivos específicos propuestos:

- Establecer relaciones filogenéticas entre las secuencias disponibles en las bases de datos anotadas como *nor* o *nitric oxide reductase*.
- Elaborar un árbol filogenético a partir de las relaciones entre las secuencias obtenidas e identificar las diferentes clases de la enzima óxido nítrico reductasa.
- Obtener perfiles generalizados de la enzima óxido nítrico reductasa que permitan discriminar entre las diferentes clases existentes.
- Validar los diferentes perfiles de la enzima óxido nítrico reductasa mediante bases de datos anotadas y librerías metagenómicas.

Por otro lado, y en relación al objetivo principal, establecer el perfil (o perfiles) de la enzima óxido nítrico reductasa (nor), como se ha comentado anteriormente en la memoria, un buen alineamiento y la elaboración del árbol filogenético a partir de las secuencias completas es quizás la parte más importante. En este proyecto, se puede considerar, que, independientemente del modo utilizado, esta etapa se ha realizado de forma satisfactoria. Por otro lado, la validación de los perfiles generalizados desarrollados mediante dichos alineamientos se realizó de forma muy superficial. Pese a ello, se pudo observar una clara capacidad para identificar y discriminar entre tipos de *nor* en diferentes organismos y en secuencias no anotadas. Los resultados obtenidos en este proyecto animan al optimismo, pero será necesario realizar más validaciones para terminar de verificar los perfiles obtenidos. Sin embargo, y dicho esto, se puede considerar que también se ha cumplido con el objetivo principal del proyecto, por lo tanto, los perfiles generalizados obtenidos en este proyecto podrían utilizarse en futuras líneas de investigación para identificar la enzima en genomas ya secuenciados o en secuencias obtenidas de novo, obtener nuevas cepas bacterianas con propiedades interesantes para su uso en agricultura y medio ambiente, o para estudios de biodiversidad, que darán lugar a nuevos conocimientos en relación a la nor, y permitirá a la comunidad científica adquirir una mayor comprensión de los mecanismos de desnitrificación, lo que en último lugar permitirá desarrollar estrategias viables para reducir las emisiones de N_2O.

Finalmente, la metodología propuesta y los *scripts* elaborados, podrían utilizarse, con equipos informáticos más potentes, para realizar este trabajo de una forma más precisa y realizar una validación más adecuada.

7.2 Líneas de futuro

Como ya se ha comentado, futuras líneas de investigación podrían incluir la realización de este mismo trabajo con equipos informáticos más potentes para obtener resultados más robustos, la validación de los perfiles mediante la secuenciación de organismos y posterior identificación mediante los perfiles generalizados obtenidos, o la preparación de los perfiles para incluirlos en la base de datos de PROSITE.

7.3 Seguimiento de la planificación

Durante el proyecto, se han producido desviaciones técnicas y temporales principalmente a causa de problemas con el equipo utilizado. Sin embargo, se puede considerar que se ha seguido con la planificación tal y como estaba previsto y se han realizado los trabajos previstos, con algunas modificaciones en los protocolos de trabajo.

8 Glosario

- **Alineamiento**: representación y comparación de dos o más cadenas de ADN, ARN, estructuras primarias de proteínas donde se evidencian zonas de similitud en un formato de matriz
- **Clade**: conjunto de ramificaciones obtenidas a partir de un corte de un árbol filogenético
- **FASTA**: formato de fichero informático basado en texto, utilizado para representar secuencias bien de ácidos nucleicos, bien de péptido, y en el que los pares de bases o los aminoácidos se representan usando códigos de una única letra
- **Match**: resultado positivo al comparar un perfil generalizado con una secuencia
- **nor**: nitrico óxido reductasa
- **Perfil generalizado**: matrices de puntuación específicas de posición, ya que incluyen puntuaciones específicas para inserciones y eliminaciones
- **Python**: lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código
- **Score**: puntuación obtenida (inserciones y eliminaciones) obtenidas al comparar un perfil generalizado con una secuencia
- **Script**: lista de comandos que ejecuta un determinado programa o motor de secuencias de comandos.

9 Bibliografía

1. Okereke, G.U. Growth yield of denitrifiers using nitrous oxide as a terminal electron acceptor. World J. Microbiol. Biotechnol. 9, 59-62, 1993.
2. Zumft, W.G. Cell biology and molecular basis of denitrification. Microbiol. Mol. Biol. Rev. 533–616, 61, 1997.
3. Lassey, D. y Harvey, M. Nitrous oxide: the serious side of laughing gas. NIWA Science 10–1, 15, 2007.
4. Bates, B.C. et al. Climate Change and Water (Technical Paper of the Intergovernmental Panel on Climate Change), IPCC Secretariat, 2008.
5. Kimochi, Y., Inamori, Y., Mizuochi, M., Xu, K-Q., Matsumara, M. Nitrogen removal and N2O emission in a fullscale domestic wastewater treatment plant with intermittent aeration. J. Ferment. Bioeng. 202-206, 86, 1998.
6. Tian, H. et al. A comprehensive quantification of global nitrous oxide sources and sinks. Nature 248–271, 586, 2020.
7. Thomson, A.J., Giannpoulos, G., Pretty, J., Baggs, E.M., Richardson, D.J. Biological sources and sinks of nitrous oxide and strategies to mitigate emissions. Phil. Trans. R. Soc. B. 1157–1168, 367, 2012.
8. de Boer, A.P., van der Oost, J., Reijnders, W.N., Westerhoff, H.V., Stouthamer, A.H., van Spanning R.J. Mutational analysis of the nor gene clúster which encodes nitric-oxide reductase from Paracoccus denitrificans. Eur. J. Biochem. 592–600, 242, 1996.
9. Butland, G., Spiro, S., Watmough, N.J., Richardson, R.J. Two conserved glutamates in the bacterial nitric oxide reductase are essential for activity but not assembly of the enzyme. J. Bacteriol. 189–199, 183, 2001.
10. Denman, K.L. et al. Couplings between changes in the climate system and biogeochemistry. En Climate change 2007: the physical basis (IPCC Fourth Assessment Report), pp. 499–588. Cambridge, UK: Cambridge University Press. 2007.
11. Torres, M.J., Simon, J., Rowley, G., Bedmar, E.J., Richardson, D.J., Gates, A.J., Delgado, M.J. Nitrous Oxide Metabolism in Nitrate-Reducing Bacteria: Physiology and Regulatory Mechanisms. Adv. Microbiol. Physiol. 353-433, 68, 2016.
12. Gaimster, H., Alston, M., Richardson, D.J., Gates, A.J., Rowley, G. Transcriptional and environmental control of bacterial denitrification and N2O emissions. FEMS Microbiol. Letters fnx277, 365, 2018.
13. Udaondo, Z., Duque E., Daddaoua, A., Caselles, C., Roca, A., Pizarro-Tobías, P., Ramos, J.L. Developing robust protein analysis profiles to identify

bacterial acid phosphatases in genomes and metagenomic libraries. Environ. Microbiol. 3561-3571, 22, 2020.

14. Casciotti, K.L. y Ward, B.B. Phylogenetic analysis of nitric oxide reductase gene homologues from aerobic ammonia-oxidizing bacteria. FEMS Microbiol. Ecol. 197-205, 52, 2005.
15. Gribskov, M., Lüthy R., Eisenberg, D. Profile analysis. Meth. Enzymol. 183, 146-159, 1990.
16. Luethy R., Xenarios I., Bucher P. Improving the sensitivity of the sequence profile method. Prot. Sci. 3, 139-146, 1994.
17. Thompson, J.D., Higgins, D.G., Gibson, T. J. Improved sensitivity of profile searches through the use of sequence weights and gap excision. Comput. Appl. Biosci. 19–29, 10, 1994.
18. Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., Bucher, P. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. 265-274, 3, 2002.
19. Tavares, P., Pereira, A.S., Moura, J.J., Moura, I. Metalloenzymes of the denitrification pathway, J. Inorg. Biochem. 2087–2100, 100, 2006.
20. Zumft, W.G. Nitric oxide reductases of prokaryotes with emphasis on the respiratory, heme-copper oxidase type, J. Inorg. Biochem. 194–215, 99, 2005.
21. Field, S.J., Thorndycroft, F.H., Matorin, A.D., Richardson, D.J., Watmough, N.J. The respiratory nitric oxide reductase (NorBC) from Paracoccus denitrificans, Methods Enzymol. 79–101, 437, 2008.
22. Gonska, N., Young, D., Yuki, R. et al. Characterization of the quinol-dependent nitric oxide reductase from the pathogen Neisseria meningitidis, an electrogenic enzyme. Sci. Rep. 3637, 8, 2018.
23. Hoglen, J., Hollocher, T.C. Purification and some characteristics of nitric oxide reductase-containing vesicles from Paracoccus denitrificans, J. Biol. Chem. 7556–7563, 264, 1989.
24. Carr, G.J., Ferguson, S.J. The nitric oxide reductase of Paracoccus denitrificans, Biochem. J. 423–429, 269, 1990.
25. Sakurai,T., Nakashima, S., Kataoka, K., Seo, D., Sakurai, N. Diverse NO reduction by Halomonas halodenitrificans nitric oxide reductase, Biochem. Biophys. Res. Commun. 483–487, 333, 2005.
26. Sakurai, T., Sakurai, N., Matsumoto, H., Hirota, S., Yamauchi, O. Roles of four iron centers in Paracoccus halodenitrificans nitric oxide reductase, Biochem. Biophys. Res. Commun. 248–251, 251, 1998.

27. Timoteo, C.G., Pereira, A.S., Martins, C.E., Naik, S.G., Duarte, A.G., Moura, J.J., Tavares, P., Huynh, B.H., Moura, I. Low-spin heme b3 in the catalytic center of nitric oxide reductase from Pseudomonas nautica, Biochemistry 4251–4262, 50, 2011.

28. Heiss, B., Frunzke, K., Zumft, W.G. Formation of the N\N bond from nitric oxide by a membrane-bound cytochrome bc complex of nitrate-respiring (denitrifying) Pseudomonas stutzeri, J. Bacteriol. 3288–3297, 171, 1989.

29. Kastrau, D.H., Heiss, B., Kroneck, P.M., Zumft, W.G. Nitric oxide reductase from Pseudomonas stutzeri, a novel cytochrome bc complex. Phospholipid requirement, electron paramagnetic resonance and redox properties, Eur. J. Biochem. 293–303, 222, 1994.

30. Kumita, H., Matsuura, K., Hino, T., Takahashi, S., Hori, H., Fukumori, Y., Morishima, I., Shiro Y. NO reduction by nitric-oxide reductase from denitrifying bacterium Pseudomonas aeruginosa: characterization of reaction intermediates that appear in the single turnover cycle, J. Biol. Chem. 55247–55254, 279, 2004.

31. Cramm, R., Pohlmann, A., Friedrich, B. Purification and characterization of the single-component nitric oxide reductase from Ralstonia eutropha H16. FEBS Lett. 6–10, 460, 1999.

32. de Vries, S., Strampraad, M.J., Lu, S., Moenne-Loccoz, P., Schroder, I. Purification and characterization of the MQH2:NO oxidoreductase from the hyperthermophilic archaeon Pyrobaculum aerophilum. J. Biol. Chem. 35861–35868, 278, 2003.

33. Suharti, R.C., Strampraad, M.J., Schroder, I., de Vries, S. A novel copper A containing menaquinol NO reductase from Bacillus azotoformans. Biochem. 2632–2639, 40, 2001.

34. Wasserfallen, A., Ragettli, S., Jouanneau, Y., Leisinger, T. A family of flavoproteins in the domains Archaea and Bacteria. Eur. J. Biochem. 325–332, 254, 1998.

35. Gomes, C.M., Teixeira, M., Wasserfallen, A. Flavins and Flavoproteins (Ghisla, S., Krömeck, P., Macheroux, P., Sund, H., eds) pp. 219–222, Rudolf Weber-Agency for Scientific Publications, Berlín, 1999.

36. Frazão, C., Silva, G., Gomes, C.M., Matias, P., Coelho, R., Sieker, L., Macedo, S., Liu, M. Y., Oliveira, S., Teixeira, M., Xavier, A.V., Rodrigues-Pousada, C., Carrondo, M.A., Le Gall, J. 2000. Structure of a dioxygen reduction enzyme from Desulfovibrio gigas. Nat. Struct. Biol. 1041–1045, 7, 2000.

37. Gomes, C.M., Giuffrè, A., Forte, E., Vicente, J.B., Saraiva, L.M., Brunori, M., Teixera, M. A Novel Type of Nitric-oxide Reductase. J. Biol. Chem. 277, 25273-25276, 277, 2002.

38. Poole, R.K., Hughes, M.N. New functions for the ancient globin family: bacterial responses to nitric oxide and nitrosative stress. Mol. Microbiol. 775–783, 36, 2000.

39. Hausladen, A., Gow, A., Stamler, J.S. Flavohemoglobin denitrosylase catalyzes the reaction of a nitroxyl equivalent with molecular oxygen. P. Natl. Acad. Sci. USA 10108–10112, 98, 2001.

40. Gardner, A.M., Helmick, R.A., Gardner, P.R. 2002. Flavorubredoxin, an inducible catalyst for nitric oxide reduction and detoxification Escherichia coli. J. Biol. Chem. 8172–8177, 277, 2002.

41. Silaghi-Dumitrescu, R., Coulter, E.D., Das, A. et al. A flavodiiron protein and high molecular weight rubredoxin from Moorella thermoacetica with nitric oxide reductase activity. Biochemistry-US 2806–2815, 42, 2003.

42. Sarti, P., Fiori, P.L., Forte, E. et al. Trichomonas vaginalis degrades nitric oxide and expresses a flavorubredoxin-like protein: a new pathogenic mechanism? Cell Mol. Life Sci. 618–623, 61, 2004.

43. Chen, L., Liu, M.Y., Legall, J. et al. Rubredoxin oxidase, a new flavohemo-protein, is the site of oxygen reduction to water by the strict anaerobe Desulfovibrio gigas. Biochem. Bioph. Res. Co. 100–105, 193, 1993.

44. Silaghi-Dumitrescu, R., Kurtz, D.M., Jr. Ljungdahl, L.G. et al. X-ray crystal structures of Moorella thermoacetica FprA. Novel diiron site structure and mechanistic insights into a scavenging nitric oxide reductase. Biochemistry-US 6492–6501, 44, 2005.

45. Seedorf, H., Hagemeier, C.H., Shima, S. et al. 2007. Structure of coenzyme F420H2 oxidase (FprA), a di-iron flavoprotein from methanogenic Archaea catalyzing the reduction of O2 toH2O. FEBS J. 1588–1599, 274, 2007.

46. Romao, C.V., Vicente, J.B., Borges, P.T. et al. The dual function of flavodiiron proteins: oxygen and/or nitric oxide reductases. J. Biol. Inorg. Chem. 21, 39–52, 21, 2016.

47. Larkin, M.A., Blackshields, G., Brown, N.P., et al. Clustal Wand Clustal X version 2.0. Bioinformatics 2947–2948, 23, 2017.

48. Gomes, C.M., Vicente, J.B., Wasserfallen, A. et al. Spectroscopic studies and characterization of a novel electron-transfer chain from Escherichia coli involving a flavorubredoxin and its flavoprotein reductase partner. Biochemistry-US 16230–16237, 39, 2000.

49. Justino, M.C., Vicente, J.B., Teixeira, M. et al. New genes implicated in the protection of anaerobically grown Escherichia coli against nitric oxide. J. Biol. Chem. 2636–2643, 280, 2005.

50. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B.A., de Castro, E., Lachaize, C., Langendijk-Genevaux, P.S., Sigrist, C.J. The 20 years of PROSITE. Nucleic Acids Res. 245-249, 36, 2008.

51. Bairoch, A. PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res. 2241–2245, 19, 1991.

52. Bairoch,A., Bucher, P. PROSITE: recent developments. Nucleic Acids Res. 3583–3589, 22, 1994.

53. Bucher, P., Karplus, K., Moeri, N., Hofmann, K. A flexible motif search technique based on generalized profiles. Comput. Chem. 3 – 23, 20, 1996

54. Gribskov, M., McLachlan, A.D., Eisenberg, D. Profile analysis: detection of distantly related proteins', Proc. Natl Acad. Sci. USA 4355 – 4358, 84, 1987.

55. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 1792-1797, 32, 2004.

56. Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol. Biol. Evol. 1547-1549, 35, 2018.

57. Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol. Biol. Evol. 268-274, 32, 2015.

58. Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., Jermiin, L.S. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods 587-589, 14, 2017.

59. Hoang, D.T., Chernomor, O., von Haeseler, A., Minh, B.Q., Vinh, L.S. UFBoot2: Improving the ultrafast bootstrap approximation. Mol. Biol. Evol. 518–522, 35, 2018.

60. Letunic, I., Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nuc. Acids Res. 293-296, 29, 2021.

61. Sibbald, P.R., Argos, P. Weighting aligned protein or nucleic acid sequences to correct for unequal representation. J. Mol. Biol. 216, 813-818, 1990.

62. Schuepbach, T., Pagni, M., Bridge, A., Bougueleret, L., Xenarios, I., Cerutti, L. pfsearchV3: a code acceleration and heuristic to search PROSITE profiles. Bioinformatics 29, 1215-1217, 2013.

63. Pina-Martins, F., Paulo, O.S. NCBI Mass Sequence Downloader–Large dataset downloading made easy. Software X 5, 80-83, 2016.

64. Burnham, K.R., Anderson, D.R. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, Nueva York, 2002.

65. Müller, T., Vingron, M. Modeling Amino Acid Replacement. J. Comp. Biol. 761-776, 7, 2000.

66. Hendriks, J., Oubrie, A., Castresana, J., Urbani, A., Gemeinhardt, S., Saraste, M. Nitric oxide reductases in bacteria, Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1459, 2000.

67. Henikoff, S. Amino Acid Substitution Matrices from Protein Blocks. PNAS 89, 10915-10919, 1992.

68. Sean, R.E. Where did the BLOSUM62 alignment score matrix come from? Nature Biotech. 22, 1035-1036, 2004.

69. Albert, Y.Z. Handbook of Nature-Inspired And Innovative Computing. Página 673, 2006.

10 Anexos

10.1 Código Python para la conversión a formato FASTA

```
import csv
import math
file = open("XX/2208.csv")

rows = []
linies = file.readlines()
for row in linies:
    rows.append(row)
del rows[0]
with open(XX/Preparado_2208.fasta', 'w') as f:
    for elements in rows:
        x=elements.split(";")
        organism = x[5]
        organism _dos = organism.split(" ")
        nombre = organism _dos[0]+"_"+ organism _dos[1]
        genus = organism _dos[0]
        f.write(">"+nombre+"_"+x[0]+"["+genus+"]")
        f.write('\n')
        secuencia = x[7]
        longitud = int(x[6])
        i = math.ceil(longitud/60)
        l= 0
        s = 60
        for x in range(i):
            f.write(secuencia[l:s])
            f.write('\n')
            l=s
            s=s+60
```

10.2 Código Python para la recuperación de secuencias y creación de archivo FASTA

```
import csv
import math
file_bd = open("XX/2208.csv")
file = open("XX/verde.txt")

rows = []
linies = file_bd.readlines()
for row in linies:
    rows.append(row)
del rows[0]

rows2 = []
linies2 = file.readlines()
for row in linies2:
    rows2.append(row)
```

```
with open("XX/verde.csv", 'w') as f:
    for elementos in rows2:
        x=elementos.split(" ")
        ref_ID = x[2]
        for completos in rows:
            y=completos.split(";")
            ID = y[0]
            if ref_ID[:-1]==ID:
                organism = y[5]
                secuencia = y[7]
                f.write(organism+";"+secuencia+"\n")

with open('XX/verde.fasta', 'w') as f:
    for elementos in rows2:
        x=elementos.split(" ")
        ref_ID = x[2]
        for completos in rows:
            y=completos.split(";")
            ID = y[0]
            if ref_ID[:-1]==ID:
                organism = y[5]
                organism _dos = organism.split(" ")
                nombre = organism _dos[0]+"_"+ organism _dos[1]
                genus = organism _dos[0]
                f.write(">"+nombre+"_"+y[0]+"["+genus+"]")
                f.write('\n')
                secuencia = y[7]
                longitud = int(y[6])
                i = math.ceil(longitud/60)
                l= 0
                s = 60
                for x in range(i):
                    f.write(secuencia[l:s])
                    f.write('\n')
                    l=s

                s=s+60
```

10.3 Código Python para la conversión a formato FASTA (modificado)

```
import csv
import math
file_bd = open("XX")
file = open("XX")

rows = []
linies = file_bd.readlines()
for row in linies:
    rows.append(row)
```

```
del rows[0]

rows2 = []
linies2 = file.readlines()
for row in linies2:
    rows2.append(row)

with open("XX", 'w') as f:
    for elementos in rows2:
        x=elementos.split(" ")
        ref_ID = x[2]
        for completos in rows:
            y=completos.split(";")
            ID = y[0]
            if ref_ID[:-1]==ID:
                bicho = y[5]
                secuencia = y[7]
                f.write(bicho+";"+secuencia+"\n")

with open('XX', 'w') as f:
    for elementos in rows2:
        x=elementos.split(" ")
        ref_ID = x[2]
        for completos in rows:
            y=completos.split(";")
            ID = y[0]
            if ref_ID[:-1]==ID:
                bicho = y[5]
                bichos_dos = bicho.split(" ")
                nombre = bichos_dos[0]+"_"+bichos_dos[1]
                genus = bichos_dos[0]
                f.write(">"+y[0])
                f.write('\n')
                secuencia = y[7]
                longitud = int(y[6])
                i = math.ceil(longitud/60)
                l= 0
                s = 60
                for x in range(i):
                    f.write(secuencia[l:s])
                    f.write('\n')
                    l=s
                    s=s+60
```

10.4 Código Python para el cuenteo

```
import csv

file = open("XX")

rows = []
```

```
linies = file.readlines()
for row in linies:
    rows.append(row)

nombres = []
for elements in rows:
    start = elements.find("[") + len("[")
    end = elements.find("]")
    substring = elements[start:end]
    nombres.append(substring)

nombres_reducido = dict()
for names in nombres:
    if (names in nombres_reducido.keys()) == False:
        nombres_reducido[names] = 1
    else:
        nombres_reducido[names] += 1

with open (‘XX', 'w') as fw:
    for atom, num in nombres_reducido.items():
        fw.write(atom+";"+str(num)+"\n")
```

10.5 Código Python para determinar las intersecciones

```
import csv

file = open("XX")
file2 = open("XX")
file3 = open("XX")

rows = []
linies = file.readlines()
for row in linies:
    rows.append(row)

rows2 = []
linies = file2.readlines()
for row in linies:
    rows2.append(row)

rows3 = []
linies = file3.readlines()
for row in linies:
    rows3.append(row)

nombres = []
for elements in rows:
    start = elements.find("protein ") + len("protein ")
    end = elements.find(" [")
    substring = elements[start:end]
    nombres.append(substring)
```

```
nombres2 = []
for elements in rows2:
    start = elements.find("protein ") + len("protein ")
    end = elements.find(" [")
    substring = elements[start:end]
    nombres2.append(substring)
nombres3 = []
for elements in rows3:
    start = elements.find("protein ") + len("protein ")
    end = elements.find(" [")
    substring = elements[start:end]
    nombres3.append(substring)

nombres=set(nombres)
nombres2=set(nombres2)
nombres3=set(nombres3)

'''
azul
rojo
verde
'''

intersection1 = nombres.intersection(nombres2)
intersection2 = nombres.intersection(nombres3)
intersection3 = nombres2.intersection(nombres3)
intersectiont = nombres.intersection(nombres2.intersection(nombres3))

print(intersection1)
print(len(intersection1))
print(intersection2)
print(len(intersection2))
print(intersection3)
print(len(intersection3))
print(intersectiont)
print(len(intersectiont))
```