Estudio de genómica comparativa de genes bacterianos relacionados con la ruta catabólica inferior del naftaleno: metabolismo del salicilato

Aitor Zarandona Garai
Master en Bioinformática y Bioestadística
Microbiología, biotecnología y biología molecular

Paloma Pizarro Tobías
David Merino Arranz

04/06/2019

Licencias alternativas (elegir alguna de las siguientes y sustituir la de la página anterior)

A) Creative Commons:

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-SinObraDerivada 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial-CompartirIgual 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-NoComercial 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-SinObraDerivada 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento-CompartirIgual 3.0 España de Creative Commons

Esta obra está sujeta a una licencia de Reconocimiento 3.0 España de Creative Commons

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Estudio de genómica comparativa de genes bacterianos relacionados con la ruta catabólica inferior del naftaleno: metabolismo del salicilato*
Nombre del autor:	*Aitor Zarandona Garai*
Nombre del consultor/a:	*Paloma Pizarro Tobías*
Nombre del PRA:	*David Merino Arranz*
Fecha de entrega (mm/aaaa):	06/2019
Titulación:	*Master en Bioinformática y bioestadística*
Área del Trabajo Final:	*Microbiología, biotecnología y biología molecular*
Idioma del trabajo:	*Castellano*
Palabras clave	*Genómica comparada, naftaleno, biodegradación*

Resumen del Trabajo (máximo 250 palabras):

El objetivo de este estudio es realizar un análisis de genómica comparativa de los genes bacterianos relacionados con la ruta catabólica inferior del naftaleno: la ruta del salicilato. El naftaleno es uno de los hidrocarburos policíclicos aromáticos (PAH) más contaminante del planeta. Este compuesto químico es muy complicado de degradar, sin embargo, existen bacterias, algas y hongos capaces de metabolizarlo. Este trabajo se focaliza en la enzima salicilato hidroxilasa (EC:1.14.13.1), la cual cataliza la conversión del salicilato a catecol. Su secuencia codificante se encuentra en el plásmido NAH7 del microorganismo *Pseudomonas putida* G7, una de las especies degradadoras de PAHs más estudiadas. Partiendo de la secuencia aminoacídica se realiza una búsqueda BLAST con el objetivo de hallar proteínas homólogas. La comparación de dichas secuencias se realiza mediante un alineamiento múltiple y la creación de árboles filogenéticos. El último paso se basa en conocer la localización de los genes y si se tratan de unidades transcripcionales o forman parte de un operón. Se comparan los operones con el mismo número de genes. BLAST, MEGAX, FGENESB, y SnapGene son algunas de las herramientas bioinformáticas utilizadas durante el trabajo. Los resultados obtenidos del estudio guían a una conclusión: el gen que codifica la enzima salicilato hidroxilasa no es parte de un operón altamente conservado.

Abstract (in English, 250 words or less):

The intent of this document is to make a genomic comparative analysis of bacterial genes involved in the lower naphthalene catalytic pathway, the salicylate pathway. Naphthalene is one of the most pollutant polycyclic aromatic hydrocarbon (PAH). These components are hard to degrade but some bacteria, fungi and algae are able to metabolize aromatic compounds. This study focuses on salicylate hydroxylase enzyme (EC:1.14.13.1), which catalyzes the reaction from salicylate to catechol. Its coding sequence is found in the NAH7 plasmid of the *Pseudomonas putida* G7 strain, one of the most studied PAH metabolizing microorganisms. Starting from the aminoacidic chain a blast is made in order to find homologous proteins. Phylogenetic trees and multiple sequence alignments are also done in order to compare the preselected sequences. The final step consists in finding its genomic localization, whether it is part of an operon or not and making a comparative of the operons that are similar. BLAST, MEGAX, FGenesB and SnapGene are some of the bioinformatics tools used to reach the objectives. The results obtained from the phylogenetic trees, database and sequence comparison lead to a conclusion, the gene that codes the salicylate hydroxylase enzyme is not part of a highly conserved operon.

Índice

1. Introducción 1
1.1 Contexto y justificación del Trabajo 1
1.2 Objetivos del Trabajo 4
1.3 Enfoque y método seguido 4
1.4 Planificación del Trabajo 5
1.5 Breve sumario de productos obtenidos 9
1.6 Breve descripción de los otros capítulos de la memoria 9
2. Resto de capítulos 11
3. Conclusiones 17
4. Glosario 23
5. Bibliografía 24
6. Anexos 27

Lista de figuras

Figura 1: Ruta catabólica del naftaleno y salicilato en el género *Pseudomonas*…….2

Figura 2: Plásmido NAH7 que contiene la información genética para la degradación del naftaleno……3

Figura 3: Reacción catabolizada por la enzima salicilato hidroxilasa…….3

Figura 4: Planificación del mes de marzo……6

Figura 5: Planificación del mes de abril ……..6

Figura 6: Planificación del mes de mayo……7

Figura 7: Planificación del mes de junio……..7

Figura 8: Diagrama de Gantt…….8

Figura 9. Ejemplo de resultado de FGENESB……

Figura 10. Imagen comparativa de los operones de 3 genes…….19

Figura 11. Imagen comparativa entre los operones de los organismos Acitenobacter baumanii y Pseudomonas oryzae…….19

Figura 12. Imagen comparativa de los operones de 2 genes……..20

1. Introducción

1.1 Contexto y justificación del Trabajo

Los compuestos aromáticos se encuentran entre los contaminantes más prevalentes y persistentes del medio ambiente. Los hidrocarburos policíclicos aromáticos (PAHs) constituyen una gran parte de estos compuestos orgánicos dañinos para la salud debido a sus efectos citotóxicos, mutagénicos y carcinógenos entre otros. Los PAHs están formados por 2 o más anillos de benceno colocados en línea, angularmente o ramificados [1].
El naftaleno, antraceno, fenantreno y pireno son algunos de los PAHs más conocidos. Estos compuestos son creados a partir de la combustión de combustibles fósiles o de procesos industriales, y son vertidos al medio ambiente por emisión directa al aire, fugas de efluentes industriales o aguas residuales entre otras vías de emisión. Los PAHs se encuentran presentes en suelos, masas de agua y aire. Son compuestos hidrófobos y su persistencia en el ecosistema se debe a su escasa solubilidad en el agua, perdurando en forma de sedimentos [2].

Estudios recientes demuestran que la degradación microbiológica de los PAHs es el proceso con mejores resultados en la descontaminación de suelos superficiales y sedimentos, pudiendo eliminar totalmente los compuestos (mineralización) o parcialmente transformados. La biorremediación se presenta como una alternativa importante para el tratamiento de estos compuestos aromáticos. Se trata de un método seguro y de bajo coste económico [3].

Con el propósito de establecer la biorremediación como un método viable para la descontaminación de zonas contaminadas por estos compuestos orgánicos, es necesario profundizar más acerca de los microorganismos capaces de utilizarlos como fuente de carbono, los procesos enzimáticos involucrados y las condiciones ambientales óptimas para el proceso.

Los PAHS son un grupo muy complejo, y las propiedades químicas y físicas varían con el peso molecular y numero de anillos aromáticos de los compuestos. Cuanto mayor es el peso molecular más complejo es el compuesto y mayores problemas presenta el estudio de su degradación [4].

De entre los PAH más conocidos y meritados anteriormente, es el naftaleno el objeto de estudio en el que se centra el presente trabajo. Este es el PAH más simple y soluble de todos y es por ello uno de los

PAH más estudiados con el fin de entender y predecir las rutas catabólicas de los PAH de mayor complejidad [5].
El naftaleno es un sólido blanco cristalino con un olor muy característico y está compuesto por 2 únicos anillos aromáticos. Y aunque sea el PAH más simple, no por ello deja de ser menos peligroso para la salud. Es uno de los 16 PAH clasificados como contaminantes prioritarios por la Agencia de Protección Medioambiental de Estados Unidos (USEPA) [6]. Estudios recientes demuestran que la lente ocular y los pulmones son especialmente sensibles al naftaleno y está clasificado como posible agente carcinógeno [7].

La ruta catabólica del naftaleno puede ser dividida en dos: la ruta catabólica superior y la ruta catabólica inferior [8]. La ruta catabólica superior reúne todas las reacciones químicas desde la utilización del naftaleno hasta la obtención del salicilato. La ruta catabólica inferior del naftaleno engloba las reacciones que suceden desde el salicilato hasta su posterior degradación total o parcial (Figura 1).

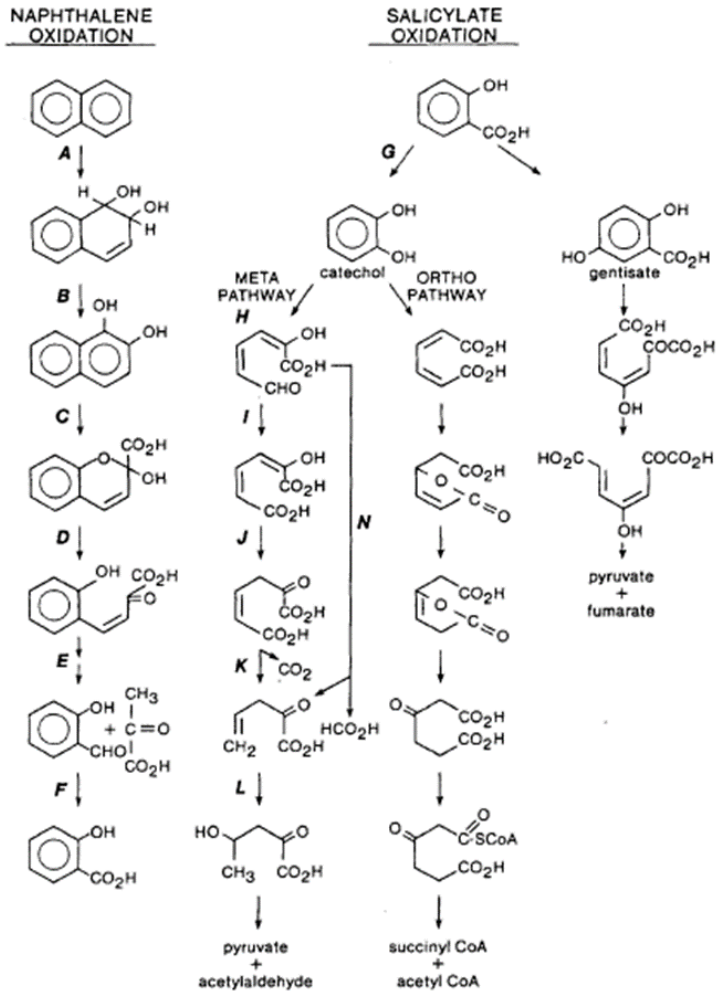


FIGURA 1. Ruta catabólica del naftaleno y salicilato en el género *Pseudomonas* [9]

Los genes asociados con la ruta catabólica del naftaleno se encuentran en el plásmido NAH7, un plásmido de 83kbps [9]. Estos genes se dividen en dos operones dentro del plásmido. En el primero se encuentran los genes *nahAaAbAcAdBCDEF*, los genes encargados de codificar la ruta catabólica superior del naftaleno, la conversión del naftaleno a salicilato [10]. En el segundo operón se encuentran los genes *nahGTHINLOMKJ.* Estos genes codifican las proteínas que toman parte en la ruta inferior del naftaleno incluyendo la conversión del salicilato en catecol [11]. El gen *nahG* es el que codifica la proteína de interés.

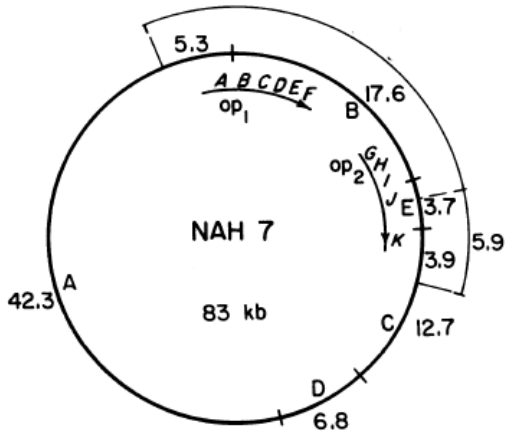


FIGURA 2. Plásmido NAH7 que contiene la información genética para la degradación del naftaleno [9].

La proteína que cataliza el proceso de transformación del salicilato a catecol es el salicilato hidroxilasa (Figura 2). Es una flavoproteína compuesta por 434 aminoácidos. Brevemente explicado, la enzima se une al salicilato y un reductor externo (NADH o NADPH) en un patrón de orden aleatorio, formando un complejo enzima-sustrato reducido, y luego un oxígeno molecular se une al complejo para la producción de catecol, CO2 y H2O [12].

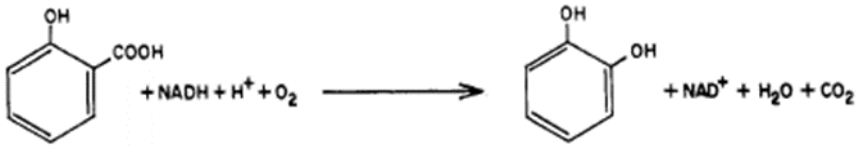


FIGURA 3. Reacción catabolizada por la enzima salicilato hidroxilasa [13].

En definitiva, el objetivo de este trabajo es analizar la ruta inferior del naftaleno, más concretamente la conversión de salicilato a catecol. Con este fin, se realizará un estudio de genómica comparativa entre microorganismos realizadores de esta ruta catabólica. Para ello se tomará como ejemplo la cepa G7 del microorganismo *Pseudomonas*

putida. Este microorganismo es especialmente capaz de degradar hidrocarburos aromáticos y ha sido muy estudiado [13].

1.2 Objetivos del Trabajo

A continuación, se exponen los objetivos que se quieren alcanzar:

1.2.1 Objetivos generales

Los objetivos generales del TFM son:

1. Analizar el metabolismo del salicilato
2. Realizar un estudio genómico comparativo de diferentes microorganismos

1.2.2 Objetivos específicos

1. Identificar las proteínas y genes que toman parte en el metabolismo del salicilato y los microorganismos que lo realizan.
2. Localizar los genes involucrados en la ruta y estudiarlos
3. Realizar un alineamiento múltiple.
4. Realizar un árbol filogenético.
5. Predecir operones y realizar comparaciones entre operones que contengan el mismo número de genes.

1.3 Enfoque y método seguido

Existen varios métodos o enfoques a seguir, pero para empezar es conveniente comenzar por una minuciosa labor de investigación respecto a la ruta catabólica inferior del naftaleno. Es imprescindible conocer qué compuestos químicos se degradan y cuáles se crean, qué proteínas toman parte y qué microorganismos la realizan.

Una vez obtenida esta información, se puede encaminar el trabajo de dos formas, mezclando las partes teóricas y prácticas del estudio o realizándolas separadamente.
La parte teórica consiste en la obtención de información acerca de la ruta, microorganismo, enzimas y genes. Pero esta parte también requiere de lectura e información acerca de las herramientas que vayan a ser utilizadas. La práctica se basa en los análisis comparativos, creaciones de base de datos y arboles filogenéticos.

En este trabajo se ha optado por el primer enfoque, ya que además de permitir controlar mejor los tiempos prefijados para la realización del TFM, se considera que dicho enfoque ameniza la lectura e invita a obtener una mayor comprensión de los conceptos objetos de estudio En este sentido, se realizará un pequeño estudio acerca de las herramientas bioinformáticas disponibles antes de realizar cada tarea. Esto es, en lugar de informarse acerca de todas las herramientas

óptimas para el desempeño del estudio desde el principio, se inclina por trabajar con cada herramienta tras obtener información suficiente y se está seguro de su adecuada adaptación al trabajo.

1.4 Planificación del Trabajo

En este bloque se explicarán las tareas a realizar en el Trabajo de Fin de Máster. Junto a ello se mostrará un calendario con las fechas de las diferentes entregas y el tiempo estimado para cada tarea. Por último, se analizarán los posibles riesgos que puedan afectar al TFM

1.4.1 Tareas

El proyecto se dividirá en las siguientes tareas:

- Identificar la ruta del salicilato y los microorganismos que la realizan.
- Identificar la proteína catalizadora y obtener su cadena proteica (enzima modelo).
- Crear una base de datos mediante la herramienta informática BLAST o similar.
- Construir el perfil de la proteína y realizar arboles filogenéticos.
    - Determinar la mejor herramienta informática para la realización de alineamientos entre proteínas.
    - Determinar la mejor herramienta informática para la creación de árboles filogenéticos.
- Localizar si los genes que transcriben la proteína de interés se encuentran en plásmido o cromosoma.
- Analizar las secuencias.
    - Predecir operones
    - Realizar comparaciones genoma-genoma entre operones similares

1.4.2 Calendario

A continuación, se muestra el calendario del Trabajo de Fin de Máster por meses. Se tiene en cuenta la disponibilidad del autor para trabajar durante las mañanas y los fines de semana y festivos, debido a que trabaja durante las tardes.

En el calendario aparecen reflejadas a partir del día 18 las tareas programadas anteriormente. La primera es obtener información acerca de la ruta y sus componentes, a pesar de que ya se haya trabajado en ello para poder realizar las primeras dos PECs. Durante el mes de marzo se dará por finalizada la obtención de información sobre la ruta y la enzima. Este será el fin para la búsqueda de información acerca de la proteína comenzada con anterioridad y se dará comienzo a la creación de la base de datos.

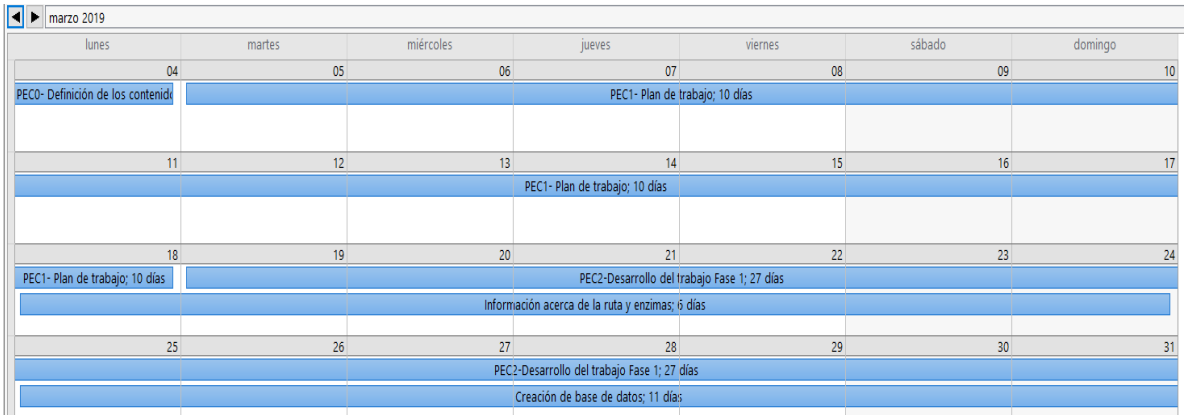


Figura 4. Planificación del mes de marzo.

En abril se comenzará creando la base de datos. Se continuará creando el perfil de la proteína y los alineamientos entre proteínas. Los días anteriores a las entregas de las PECs se dejarán 2-3 días de margen para la adecuada redacción de ellas.

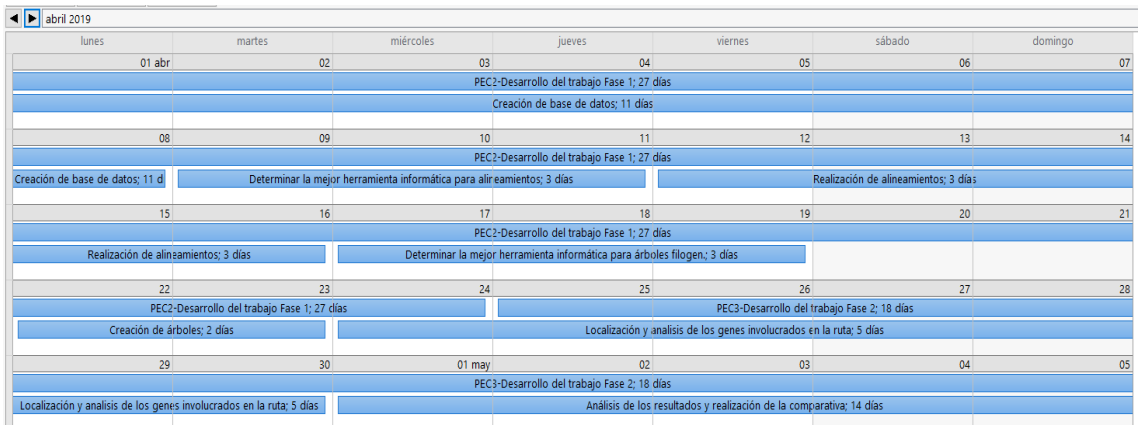


Figura 5. Planificación del mes de abril.

En mayo se encaminará el trabajo hacia su final, realizando los análisis de genómica comparativa, tales como la predicción de operones y su posterior comparación. En este mes también se comenzará con la redacción del producto final, tratando de no adelantar tareas con el fin de disponer tiempo para contrarrestar posibles desajustes en la temporalización del TFM.

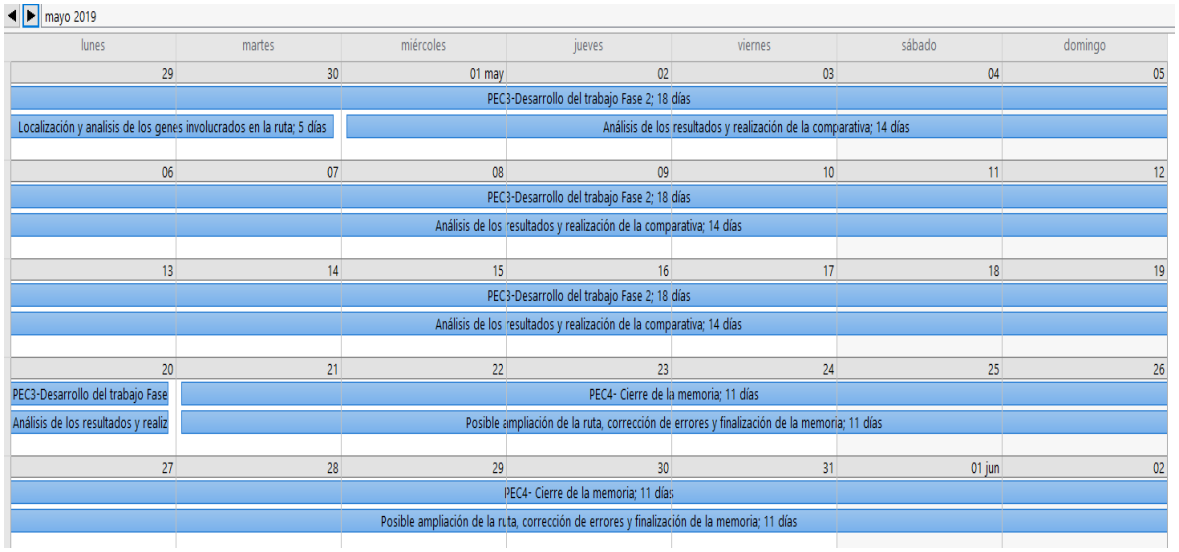


Figura 6. Planificación del mes de mayo

Para finalizar, en junio se procederá con las últimas fases del Trabajo de Fin de Master: redacción de la memoria, elaboración de la defensa y la defensa pública.

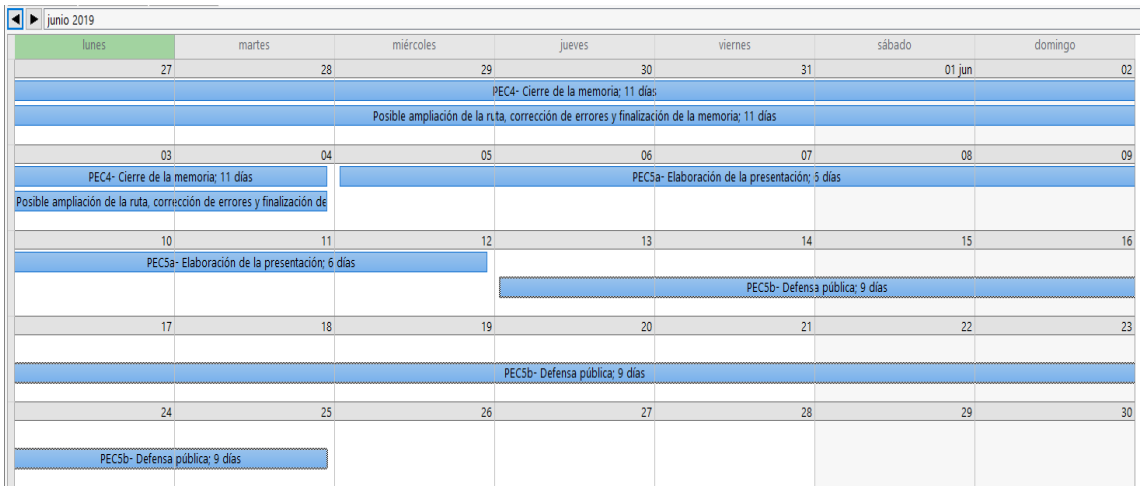


Figura 7. Planificación del mes de junio.

Por último, se muestra un diagrama de Gantt con la información completa acerca de la planificación del TFM.

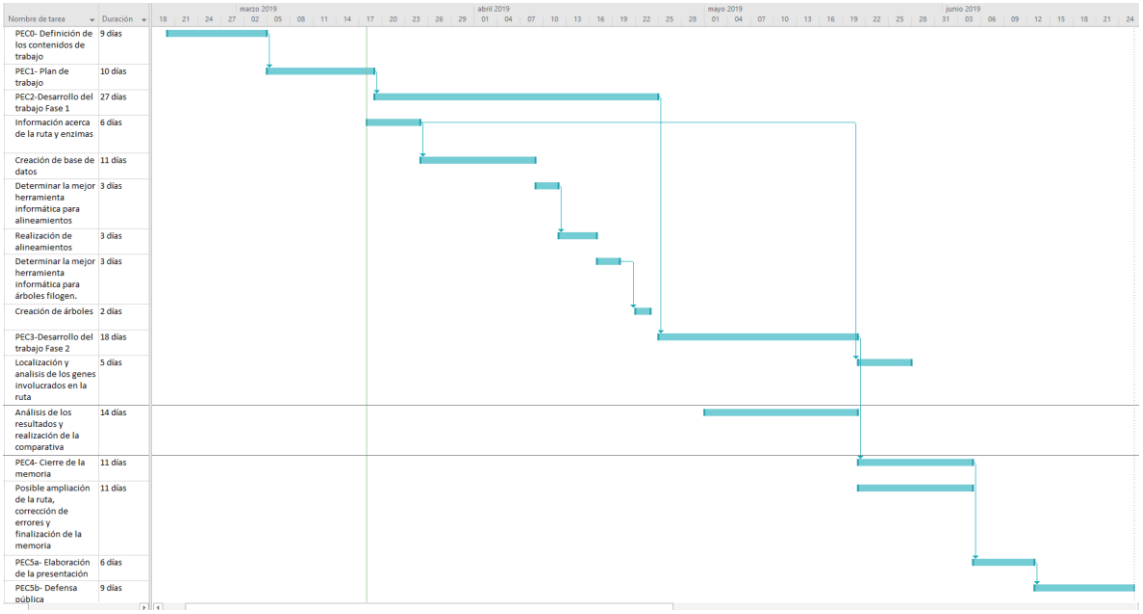


Figura 8. Diagrama de Gantt

1.4.3 Hitos

En la siguiente tabla se muestran los hitos, fechas clave para el trabajo y en las que es importante ser lo más preciso posible y no demorarse, puesto que repercutiría en las siguientes entregas o tareas.

Hito	PEC	Fecha de entrega
Plan de trabajo	PEC 1	18/03/2019
Creación base de datos	PEC2	15/04/2019
Alineamientos entre proteínas	PEC2	24/04/2019
Análisis comparativo	PEC3	20/05/2019
Entrega de memoria	PEC4	4/06/2019
Elaboración de la presentación	PEC5	12/06/2019
Defensa pública	PEC6	25/06/2019

Tabla 1. Hitos y sus fechas de entrega.

1.4.4 Análisis de riesgos

En este trabajo existen posibles factores que podrían afectar negativamente en la realización de las tareas o en su planificación. A continuación, se enumeran algunos de estos posibles factores:

1. El proyecto no dispone de un tiempo muy extenso.
2. La mala realización del plan de trabajo, no sabiendo administrar bien el tiempo.
3. Falta de conocimiento en las herramientas informáticas necesarias.
4. Mala elección de la enzima/microorganismo modelo.
5. Reconocimiento tardío de errores y tiempo limitado para su corrección.

1.5 Breve sumario de productos obtenidos

El trabajo se verá finalizado una vez estén realizados los siguientes entregables

1.5.1 Plan de trabajo

En él se redacta el trabajo que se va a realizar y el enfoque que va a tener. Se especifican los objetivos del TFM y se muestra la planificación de los hitos. También se analizan los posibles riesgos que conlleva la realización del estudio.

1.5.2 Memoria

En esta parte se documenta todo el trabajo realizado durante el Trabajo de Fin de Máster. Se explica el contexto en el que se realiza y la razón por la que es interesante para la sociedad. Se detallan los objetivos a alcanzar y los métodos a seguir, y, por último, los resultados obtenidos.

1.5.3 Presentación virtual

Se realizará una presentación oral que resuma el proyecto y sus resultados. Para ello precisará de un material de apoyo visual que favorezca la comprensión del trabajo expuesto.

1.5.4 Autoevaluación del proyecto

La autoevaluación será el reflejo de la consecución de los objetivos alcanzados. En ella se reflexionará sobre el motivo de los objetivos que han sido obtenidos con éxito, así como sobre los que han quedado pendientes de alcanzar. Y además, la autoevaluación incluirá unas conclusiones acerca del resultado final del trabajo.

1.6 Breve descripción de los otros capítulos de la memoria

La memoria está compuesta por 6 apartados sin considerar esta introducción:

- Materiales y métodos: En este capítulo se detallan todos los procesos que se llevan a cabo durante el estudio. De igual

manera se informa acerca de las herramientas bioinformáticas empleadas.

- Resultados: Se presentan los productos obtenidos tras el análisis realizado mediante las herramientas bioinformáticas.
- Conclusiones: En este apartado se valoran tanto los resultados obtenidos durante el estudio como la realización personal del trabajo.
- Glosario: Definición de conceptos y aclaración de siglas y acrónimos empleados durante la memoria.
- Bibliografía: Es en este apartado donde se muestran todas las fuentes de información mencionadas durante el trabajo. De ellas se ha extraído el conocimiento para la realización del estudio.
- Anexos: Último apartado en el cual se exponen los archivos demasiado grandes para ser mostrados durante la memoria. De este modo no se entorpece la lectura y se almacenan al final del documento.

2. Resto de capítulos

2.1 Materiales y métodos

Lo primero tras informarse acerca de la ruta del salicilato ha sido seleccionar el microorganismo y la proteína modelo. Tal y como se menciona en la introducción, el microorganismo seleccionado es *Pseudomonas putida G7.* Se ha escogido este microorganismo debido a que ha sido ampliamente estudiado y se conoce mucho acerca del propio microorganismo, proteína de interés y genes que lo codifican [13].

2.1.1 Realización del BLAST y de la base de datos

El primer paso es obtener la secuencia de aminoácidos que componen la proteína salicilato hidroxilasa (EC:1.14.13.1). Se parte de la secuencia de la proteína molde y así encontrar proteínas homologas con el fin de compararlas y poder alinearlas más adelante. Dos secuencias son homologas en caso de que compartan un ancestro evolutivo común, en otras palabras, ambas secuencias parten de una misma secuencia. Para ello, es necesario hallar secuencias con un alto porcentaje de identidad, una similitud muy elevada como para ser considerada producto del azar [14].
Para ello se utiliza el programa BLAST de NCBI: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome.

BLAST (Herramienta de búsqueda básica de alineación local) es un método de búsqueda de similitud de secuencia, en el que una secuencia de nucleótidos o una proteína de interés se compara con secuencias de nucleótidos o proteínas en una base de datos. Su finalidad es identificar regiones de alineación local e informar de aquellas alineaciones que obtienen una puntuación superior a una puntuación umbral [15].

El BLAST se realiza con el algoritmo PSI-BLAST (Position-Specific Iterated BLAST) en lugar de utilizar el algoritmo predeterminado pBLAST con el objetivo de obtener unos mejores resultados.
Este es un algoritmo más complejo que el pBLAST y se divide en varias fases.
El proceso comienza con una búsqueda mediante el algoritmo pBLAST. A continuación, se realiza un alineamiento múltiple y se genera una matriz de puntuación especifica de la posición (PSSM). Esta matriz captura el patrón de conservación en el alineamiento y lo almacena como una matriz de puntuaciones para cada posición en la misma. Este perfil se usa en lugar de la matriz de sustitución original para una búsqueda adicional de la base de datos con intención de detectar secuencias que coincidan con el patrón de conservación especificado por el PSSM. Las secuencias recién detectadas en esta segunda ronda

de búsqueda, que están por encima del umbral de puntuación (valor e) especificado, se agregan nuevamente a la alineación. Este proceso se continúa de forma iterativa hasta que no se detecten secuencias nuevas por encima del umbral establecido. De este modo, PSI-BLAST es mucho más capaz de detectar similitudes entre secuencias distantes que una sola consulta con el algoritmo pBLAST [16].

Como resultado del BLAST se obtiene una larga lista de cadenas proteicas similares. Se debe filtrar el resultado y tratar de trabajar con las proteínas potencialmente homólogas. No existe ninguna certeza para poder asegurar al 100% que dos proteínas sean homologas partiendo del porcentaje de similitud. Sin embargo, existe un mínimo arbitrariamente establecido para considerar dos proteínas homólogas. Seleccionaremos las proteínas con un porcentaje de identidad >40%, query cover >90% y valor E $<10^{-6}$ [14].

Una vez filtrado el resultado se genera una base de datos en una hoja de Excel. Esta se compone de alrededor de 50 especies de microorganismos. En ella se almacena información acerca de la cadena proteica, cepa que la produce, localización del gen, existencia de la anotación del genoma completo y cobertura del alineamiento como datos importantes.

2.1.2 Realización del alineamiento múltiple

El alineamiento múltiple se realiza con la información almacenada en la base de datos. Existen varias herramientas bioinformáticas capaces de realizar alineamientos entre proteínas y construir arboles filogenéticos. Cada una de ellas muestra tanto ventajas como desventajas.
Después de realizar un pequeño estudio acerca de estas herramientas se selecciona el software MEGAX [17]. Este software es de código libre y se descarga desde el siguiente link: https://www.megasoftware.net/. La elección de este programa bioinformático se debe a que ofrece gran variedad de algoritmos y métodos con los que trabajar y tiene una interfaz “user-friendly” [18]. Es una herramienta muy completa y adecuada para principiantes debido a su facilidad de manejo. Puede realizar alineamientos de secuencias múltiples (MSA) con varios algoritmos tales como MUSCLE o CLUSTALW. A su vez, también permite la realización de árboles filogenéticos como máximum-likelihood (ML), neighbor-joining (NJ) o máximum-parsimony (MP). Debido a su gran variedad y flexibilidad es uno de los programas más populares para el análisis filogenético, especialmente recomendado para trabajadores noveles en estudios filogenéticos. Por este motivo, en este trabajo se utiliza el mismo programa para realizar tanto el alineamiento como el árbol filogenético.

Para empezar a trabajar con el software MEGAX se deben transformar los datos al formato fasta. Se crea un archivo fasta que contenga las secuencias aminoacídicas de las proteínas previamente escogidas, y desde el programa se lee el fichero. Para la creación de este tipo de

archivos, primero se incluyen las secuencias aminoacídicas de todas las proteínas en un fichero de texto para después convertirlo en un archivo fasta mediante el programa MEGAX. Las secuencias dentro del fichero de texto deben estar encabezadas por el signo > antes del nombre que se desee dar a la secuencia.

Se cree que MUSCLE es el método más apropiado para el alineamiento múltiple de este estudio. CLUSTALW ha sido uno de los referentes y una de las herramientas más utilizadas a la hora de realizarlos. No obstante, se ha visto superada en potencia y precisión por nuevas herramientas como MUSCLE. Cabe destacar que no todas las secuencias poseen la misma longitud y MUSCLE es un método más apropiado el alineamiento de las secuencias que se disponen [19].

El alineamiento es de vital importancia a la hora de realizar análisis filogenéticos. De hecho, el alineamiento puede tener mayor influencia en el resultado que la selección del método del tipo de árbol filogenético. Sin embargo, muchos estudios dedican mayor atención a la selección del tipo de árbol filogenético en lugar de escoger el mejor método para la correcta realización del alineamiento [20].

Tal y como se ha mencionado anteriormente, MUSCLE necesita de mayor potencia de computación. A pesar de ello, en este estudio no se dispone de una base de datos muy pesada para analizar y se realiza el alineamiento sin ningún problema de este tipo.

2.1.3 Creación del árbol filogenético

Una vez realizado el alineamiento correctamente se procede a generar el árbol filogenético. Un árbol filogenético es un diagrama que representa las líneas de descendencia evolutiva de diferentes especies, genes u organismos de un ancestro común. Estos son útiles para organizar el conocimiento de la diversidad biológica, para estructurar clasificaciones y para proporcionar información sobre los eventos que ocurrieron durante la evolución [21].

En este caso se vuelve a disponer de una gran variedad de métodos a elegir. Los métodos más utilizados son el NJ, el ML y MP. Los algoritmos ML y MP están basados en métodos probabilísticos y de caracteres, y el algoritmo NJ en métodos de medición de distancias [22].
Los árboles formados mediante MP y ML son más robustos debido a que utilizan modelos evolutivos más complejos y requieren mayor poder de computación. Esto es, producen árboles más fidedignos y realistas.
Por otra parte, los árboles NJ son más básicos con lo que son más fáciles de analizar y observar.

En lo que al caso objeto de estudio respecta, no se dispone de una gran cantidad de secuencias para analizar, con lo que ningún método ocasiona problemas debido a la potencia computacional exigida. Al no saber decantarse por la utilización de un tipo de árbol en concreto, se

generan 3 árboles mediante los 3 métodos previamente mencionados con el fin de ser analizados. Todos ellos son generados a partir del alineamiento MUSCLE y con un valor Bootstrap de 500.

2.1.4 Localización y visualización de las secuencias.

Hasta este momento se ha estado trabajando con los aminoácidos que componen las proteínas de interés. A partir de ahora, sin embargo, se trabaja con los genes y las secuencias de nucleótidos que codifican estas proteínas.

Para conseguir la secuencia genómica de las proteínas, se accede mediante el buscador de PubMed a las proteínas de interés. En la base de datos anteriormente realizada se encuentran los números de acceso o accesion number de estas proteínas. Inmediatamente después de acceder a estos enlaces se muestra en pantalla la localización del gen (plásmido o cromosoma) y la secuencia de nucleótidos del genoma o plásmido completo en formato fasta. Esta información ha sido accesible en la mayoría de proteínas que se contemplan en la base de datos.

Una vez se han obtenido los genomas o plásmidos completos, el primer paso será visualizar las secuencias. Para lo cual se utilizará el programa Artemis de Sanger Institute (https://www.sanger.ac.uk/science/tools/artemis). Este programa ofrece la posibilidad de observar y visualizar las secuencias previamente seleccionadas.

2.1.5 Predicción y comparación de operones

Uno de los objetivos del trabajo es realizar comparaciones genoma-genoma de los operones que contengan el mismo número de genes.

Un operón es un sistema regulador genético que se encuentra en las bacterias y sus virus, en el que los genes que codifican proteínas relacionadas funcionalmente se agrupan a lo largo del ADN. Esta característica permite que la síntesis de proteínas se controle coordinadamente en respuesta a las necesidades de la célula. Al proporcionar los medios para producir proteínas solo cuando y donde se requieren, el operón permite a la célula conservar energía, siendo esto una parte importante de la estrategia de vida de un organismo. Los genomas de las bacterias y las arqueas generalmente contienen un número pequeño de operones altamente conservados y un número mucho mayor de únicos o raros [23].

Primero se debe tratar de predecir los operones presentes en las secuencias. La primera herramienta bioinformática que se utiliza para ello es FGENESB de la web Softberry:

http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb.
FGENESB es un programa de predicción de genes procariotas *ab initio* basado en los modelos de cadena de Markov de las regiones de codificación, la traducción y los sitios de terminación. Este programa también incluye una predicción de operones basada en las distancias entre los genes predichos [24]. Es una herramienta muy sencilla de manejar y una de las herramientas predictoras de genes y operones más utilizadas [25].
Gracias a esta herramienta se introduce el genoma o plásmido completo en el que está la proteína de interés, lo que permite obtener los posibles genes de dicho plásmido o genoma (Figura 9).

El gen puede ser clasificado como parte de un operón o como unidad transcripcional. Una unidad transcripcional es una secuencia de nucleótidos en el ADN que codifica una única molécula de ARN, junto con las secuencias necesarias para su transcripción; y normalmente, contiene un promotor, una secuencia de codificación de ARN y un terminador [26].

Partiendo de todo el genoma que se introduce se busca directamente el gen y se anota si la predicción ha confirmado la existencia de un operón y la cantidad de genes que lo componen.

```
Prediction of potential genes in microbial genomes
Time:   Tue Jan  1 00:00:00 2005
Seq name: NC_007926.1 Pseudomonas putida plasmid NAH7, complete sequence
Length of sequence - 82232 bp
Number of predicted genes - 113
Number of transcription units - 46, operons - 22
   N      Tu/Op   Conserved  S              Start          End    Score
                  pairs(N/Pv)
   1      1 Op  1      .       +     CDS          1 -        774    258
   2      1 Op  2      .       +     CDS        785 -       1711    644
   3      1 Op  3      .       +     CDS       1708 -       2055    206
   4      1 Op  4      .       +     CDS       2052 -       2468    260
   5      2 Tu  1      .       +     CDS       2736 -       3383    161
   6      3 Op  1      .       +     CDS       3775 -       4407    258
   7      3 Op  2      .       +     CDS       4447 -       5208    237
   8      4 Tu  1      .       -     CDS       5448 -       5639    152
   9      5 Op  1      .       +     CDS       5580 -       5849    129
  10      5 Op  2      .       +     CDS       5880 -       7127    410
  11      6 Tu  1      .       -     CDS       7382 -       7606     74
  12      7 Op  1      .       -     CDS       7780 -       8961    554
  13      7 Op  2      .       -     CDS       9053 -       9481    191
  14      8 Op  1      .       -     CDS      10076 -      10372    282
  15      8 Op  2      .       -     CDS      10369 -      10671    182
  16      9 Op  1      .       +     CDS      10883 -      11101    187
  17      9 Op  2      .       +     CDS      11165 -      11326    234
  18      9 Op  3      .       +     CDS      11330 -      11482    101
  19      9 Op  4      .       +     CDS      11507 -      11680     97
  20      9 Op  5      .       +     CDS      11692 -      12114    170
  21     10 Tu  1      .       +     CDS      12774 -      12968    118
  22     11 Tu  1      .       -     CDS      13236 -      13490    124
  23     12 Tu  1      .       +     CDS      13489 -      13938    244
  24     13 Tu  1      .       -     CDS      14199 -      14438    198
  25     14 Op  1      .       +     CDS      14501 -      15052    195
  26     14 Op  2      .       +     CDS      15107 -      15529    168
  27     14 Op  3      .       +     CDS      15553 -      15882    255
  28     14 Op  4      .       +     CDS      15875 -      16147    205
  29     14 Op  5      .       +     CDS      16161 -      16367    100
  30     14 Op  6      .       +     CDS      16388 -      16645    180
  31     15 Tu  1      .       -     CDS      17273 -      18058    341
  32     16 Op  1      .       -     CDS      18196 -      18582    347
  33     16 Op  2      .       -     CDS      18586 -      19305    511
  34     16 Op  3      .       -     CDS      19302 -      19757    246
  35     16 Op  4      .       -     CDS      19788 -      19952     78
  36     17 Op  1      .       +     CDS      20188 -      20580    140
  37     17 Op  2      .       +     CDS      20573 -      22132    692
  38     17 Op  3      .       +     CDS      22143 -      25079   1365
  39     17 Op  4      .       +     CDS      25155 -      25904    157
  40     18 Tu  1      .       -     CDS      25934 -      26467    240
  41     19 Tu  1      .       +     CDS      26992 -      27348    357
  42     20 Tu  1      .       +     CDS      27761 -      28549    386
```

Figura 9. Ejemplo de resultado de FGENESB.

El programa exige que se establezca un organismo cercano a la secuencia introducida con el fin de realizar la predicción más certera posible. Los siguientes organismos son los seleccionados como más cercanos:

- *Pseudomonas putida* KT2440 para las secuencias del género *Pseudomonas*.
- *Acitenobacter sp* ADP1 para las secuencias del género *Acitenobacter*
- *Xanthomonas campestris* para las secuencias del género *Xanthomonas*
- Bacterial generic para las secuencias en las que no se observa ningún organismo cercano seleccionable.

En un principio se intentó predecir el operón introduciendo solamente la secuencia del gen en lugar de la secuencia del genoma en su totalidad, pero la herramienta no es capaz de predecir los operones a no ser que se introduzca una secuencia más extensa. Después de anotar los operones de todas las secuencias se procede a la utilización de la siguiente herramienta bioinformática.

Inmediatamente después de obtener la información acerca de los operones, la siguiente tarea es la de visualizar y comparar los operones.

La herramienta SnapGene (GSL Biotech disponible en snapgene.com) se utiliza para poder visualizar la organización de los operones. En ella se introducen los genomas completos de los organismos guardados en la base de datos. SnapGene es una herramienta bioinformática que permite visualizar y realizar simulaciones de manipulación de ADN. Es especialmente sencilla de manejar y está dotada de múltiples “add-on” con variedad de utilidades. No es una herramienta de código libre, aun así, se ha podido emplear gracias a una demostración gratuita de 30 días [27].

La recién mencionada herramienta se utiliza con el objetivo de seleccionar la secuencia que forma parte del operón y guardarla para su posterior análisis. Para ello, se cargan los archivos fasta con los genomas o plásmidos completos desde el programa con el comando open file. Una vez la herramienta reproduce la secuencia, con el comando select range se eligen los nucleótidos que constituyen el operón de interés. Esta información se muestra en los resultados de la predicción de FGENESB, donde queda reflejada la posición de los nucleótidos que componen los genes.

Como último paso para la realización del análisis comparativo se realizan comparaciones genoma-genoma entre los operones que tengan el mismo número de genes.
A partir de este momento solo se utilizan las secuencias de los operones previamente seleccionadas con el programa SnapGene.

Easyfig (http://mjsull.github.io/Easyfig/) es una herramienta para trazar figuras de comparación de múltiples genomas o regiones genómicas de archivos de anotaciones (por ejemplo, GenBank y EMBL). El resultado es una imagen clara y de interpretación sencilla [28]. Por tanto, esta herramienta permite concluir si los operones comparten los mismos genes o solamente tienen la misma cantidad de genes sin que estos compartan similitud. El programa reproduce las secuencias de los operones mediante líneas horizontales, y las posiciona una encima de la otra. Para la creación de la imagen es necesario un archivo con los resultados del BLAST entre secuencias, archivo que la propia herramienta genera. El resultado es una imagen con tantas líneas horizontales se deseen comparar, donde se observa un sombreado débil, fuerte o inexistente dependiendo del porcentaje de similitud que muestren las secuencias como resultado del BLAST.

2.2 Resultados

En este apartado se procede a analizar los resultados obtenidos atendiendo a la metodología explicada anteriormente. Con el objetivo de simplificar la lectura, solo se escribirá la cepa de los organismos en caso de que en la base de datos existan 2 o más organismos de la misma especie.

En primer lugar, cabe destacar que los resultados del BLAST se consideran muy satisfactorios, ya que gracias a ellos se ha creado una base de datos sólida con alrededor de 50 secuencias o microorganismos. En un principio solo se anotaron las secuencias aminoacídicas, y se creía que podría haber problemas a la hora de buscar su localización genómica y de conseguir la información del genoma o plásmido completo. Sin embargo, la información acerca de estos microorganismos es muy amplia y se ha podido obtener toda la información necesaria para la realización del estudio en prácticamente la totalidad de las especies anotadas en la base de datos [Anexo I]

El alineamiento se realiza correctamente y es utilizado para generar los árboles filogenéticos. Se observa como la diferencia de longitud de las secuencias no ha supuesto ningún problema en la correcta realización de él [Anexo II].

Debido a la duda sobre cuál sería el árbol filogenético más conveniente para este trabajo, se crean 3 de los árboles más utilizados en los estudios filogenéticos [Anexo III]. Las 2 especies más alejadas son *Phialocephala scopiformis* y *Marssonina brunnea*, los dos únicos organismos eucariotas del árbol, puesto que son hongos y no bacterias.
En los árboles se puede observar cómo todos los organismos del género *Pseudomonas* comparten un ancestro cercano, ya que las secuencias poseen un alto porcentaje de identidad. Lo mismo ocurre con el otro género mayoritario de la base de datos, *Acitenobacter*. Todos los organismos de este género se encuentran cercanos, no obstante, son los

organismos más lejanos o diferenciados respecto al organismo molde *Pseudomonas putida* G7. *Nephila clavipes* y *Xanthomonas arborícola* son las especies más cercanas a *Pseudomonas putida* G7 sin considerar los organismos del género *Pseudomonas*.

Los resultados obtenidos de la predicción de operones son totalmente heterogéneos. Con el fin de compararlos, primero se separan las secuencias en 2 grandes grupos conforme a los resultados del programa FGENESB. Por un lado quedan las unidades transcripcionales (Tu), y por otro, las que forman parte de un operón (Op).
La mayoría de las secuencias del género *Pseudomonas* han sido clasificadas como unidades transcripcionales. La secuencia de la proteína utilizada como molde durante todo el estudio, *Pseudomonas putida* G7, también se encuentra entre las unidades transcripcionales. Este dato se contrapone con la información presente en la bibliografía, donde esta enzima se presenta como parte de un operón junto a otros 9 genes. Este problema se puede deber a que el organismo seleccionado como más cercano a la hora de realizar la predicción no tiene en cuenta su localización en un plásmido. A pesar de ello, en este estudio se cree que la selección de los organismos más cercanos ha sido la correcta tras la realización de varias pruebas.

A continuación, se clasifican los organismos teniendo en cuenta la cantidad de genes que forman el operón:

- 2 genes (6 organismos): *Pseudomonas putida ND6, Pseudomonas sagitaria, Pseudomonas delhiensis, Chromohalobacter salexigens, Halomonas titanicae, Acitenobacter johnsonii*.
- 3 genes (8 organismos): Pseudomonas *oryzae*, Pseudomonas *psychrotolerans*, Acitenobacter *baumanii*, Acitenobacter *pittii*, Xanthomonas *sacchari*, Acitenobacter *bereziniae*, Acitenobacter *lactucae*, Serratia *sp*.
- 4 genes: *Acitenobacter junii*
- 5 genes: *Xanthomonas campestris*
- 7 genes: *Halomonas anticariensis*
- 9 genes: *Kushneria phyllosphaerae*

La herramienta bioinformática FGENESB predice que solo 18 de las 44 secuencias forman parte de un operón. En otras palabras, el 41% de las secuencias de la base de datos son clasificadas como unidades transcripcionales. Los organismos *Phialocephala scopiformis* y *Marssonina brunnea* no se han tenido en consideración en esta tarea debido a que son hongos y el estudio se centra en bacterias.

En la imagen comparativa de los operones de 3 genes se observa que los operones del género *Acitenobacter* mantienen un elevado porcentaje de identidad en todos los genes que componen el operón. Lo mismo sucede con las 2 secuencias del género *Pseudomonas*, pero esta vez, solo en parte del operón (gen de interés). El resto de los genes que

forman los operones no alcanzan el porcentaje mínimo de identidad compartida.

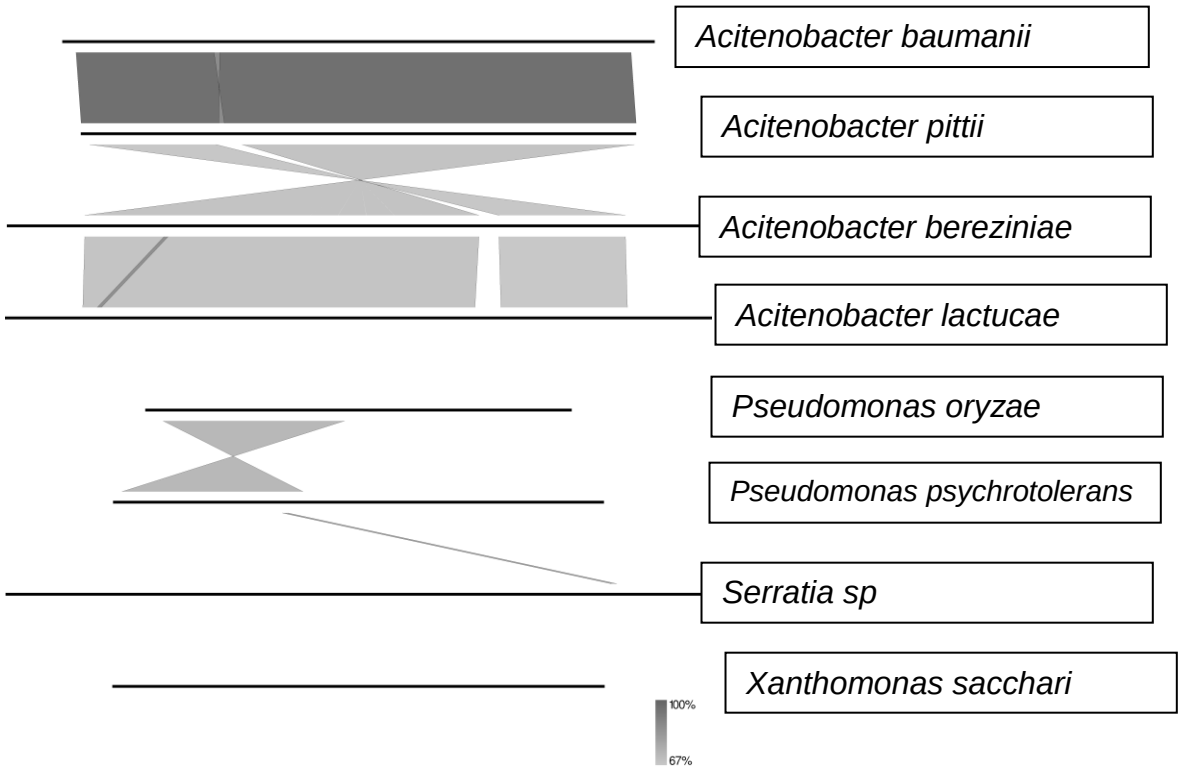


Figura 10. Imagen comparativa de los operones de 3 genes.

Las líneas reproducen la secuencia nucleotídica del operón. El sombreado entre líneas representa el porcentaje de similitud que existe entre estos operones. Cuanto más oscuro sea el sombreado mayor será éste. En caso de que el porcentaje de identidad este por debajo del mínimo establecido no habrá nada coloreado entre secuencias. Este sombreado se representa con forma de reloj de arena en caso de que las secuencias sean similares en sentidos diferentes de la hebra.

La comparación solo se realiza entre una secuencia y su inmediatamente anterior o posterior. Por ello, a continuación se realiza una comparación entre las secuencias de los 2 géneros más cercanos al organismo molde. En la figura 6 no se aprecia una similitud significativa entre los operones de diferentes géneros.

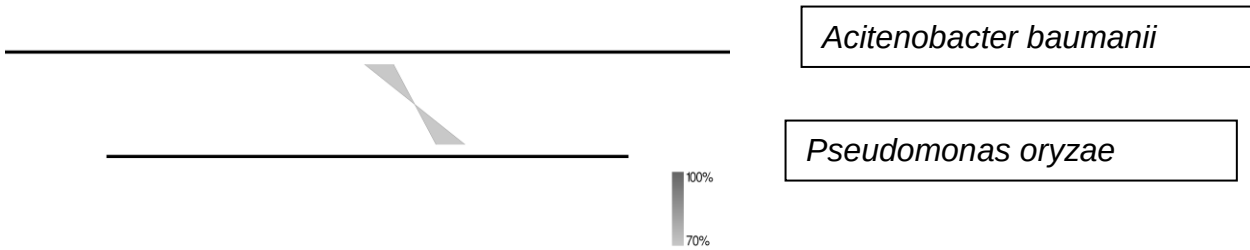


Figura 11. Imagen comparativa entre los operones de los organismos *Acitenobacter baumanii* y *Pseudomonas oryzae*

Dicho esto, se procede a analizar la imagen comparativa los operones formados por 2 genes:

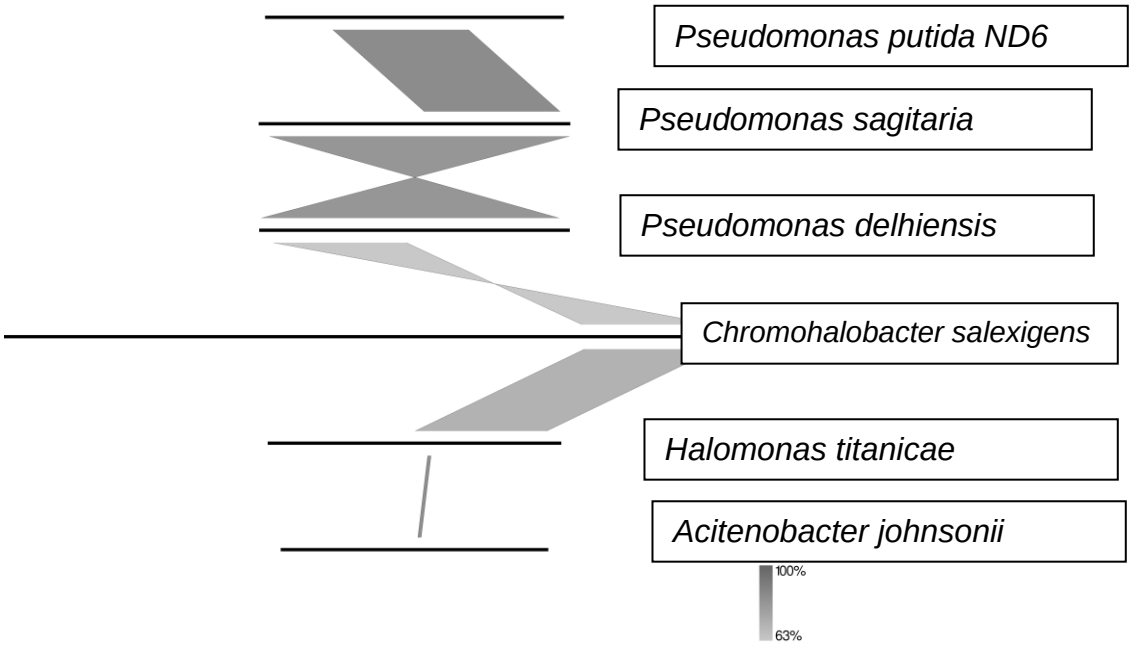


Figura 12. Imagen comparativa de los operones de 2 genes.

En este caso se obtiene información parecida a la de la imagen de los operones de 3 genes. Los operones de las secuencias del género *Pseudomonas* están altamente conservados, tal y como indica el color gris oscuro de la imagen. Atendiendo a la comparación entre secuencias de géneros diferentes, solamente parece haber coincidencia en parte del operón o en un gen, probablemente en el gen codificador de la enzima salicilato hidroxilasa.

Los resultados de las imágenes comparativas de operones indican similitud entre los operones del mismo género. En cambio, rechazan la posible homología entre los operones de las secuencias más alejadas en los árboles filogenéticos.

Existe la posibilidad de razonar esta afirmación mediante 2 argumentos.
Por un lado, las secuencias han sido halladas tras realizar un BLAST contra la enzima molde, es decir, éstas guardan un elevado porcentaje de identidad respecto a esta enzima, pero podría ser que 2 secuencias de la base de datos no lo posean entre ellas.
Por otro lado, el programa Easyfig reproduce líneas entre las secuencias que superen el porcentaje mínimo de alrededor del 65%. Es decir, indica similitud entre las secuencias que superen este porcentaje de igualdad de identidad. En la creación de la base de datos se ha creído conveniente establecer el grado de identidad mínimo para considerar dos genes potencialmente homologas del 40%.

3. Conclusiones

Teniendo en cuenta el estudio realizado, son varias las conclusiones alcanzadas. Los resultados de las imágenes comparativas coinciden con los árboles filogenéticos, existe mayor grado de conservación en los operones de las especies que comparten ancestros más cercanos.

Del trabajo se concluye que el gen codificador de la enzima encargada de catalizar la conversión del salicilato a catecol no se encuentra en un operón altamente conservado. Este dato puede explicar la razón por la que la especie *Pseudomonas putida* es un excepcional degradador del naftaleno y lo pueda utilizar como fuente de carbono. Las demás especies del estudio coinciden en ser capaces de convertir el salicilato en catecol, sin entrar en detalle en la posible eficacia/eficiencia de cada enzima. En cambio, estos microorganismos pueden carecer de los genes necesarios para codificar el resto de enzimas que toman parte en la ruta completa del naftaleno. Por otro lado, es posible que el resto de organismos realicen una ruta alternativa o incompleta, donde otros genes estén involucrados.

La bibliografía indica que el gen utilizado como molde durante todo el estudio se encuentra en un operón. No obstante, y a pesar de intentarlo con varias herramientas bioinformáticas distintas, no se ha podido elaborar una predicción que coincida con este dato. Aun considerando el gen molde parte de un operón, las comparaciones hubieran tenido el mismo resultado. Se baraja la hipótesis de fallo en la predicción debido a su localización en un plásmido. Tal y como se muestra en la base de datos, muy pocas son las secuencias que se encuentran en un plásmido, con lo que las predicciones hubieran resultado ser similares a las obtenidas.

Todo esto demostraría la importancia del organismo *Pseudomonas putida* y su gran utilización en estudios donde se clona este operón del plásmido NAH7 donde se almacenan todos los genes codificadores de la ruta del naftaleno.

Los resultados y el trabajo realizado durante el estudio se consideran satisfactorios. La metodología se considera la correcta una vez haber finalizado el estudio.

Una de las lecciones del trabajo ha sido que siempre se debe trabajar meditando y teniendo en cuenta la información adquirida, pues esta es la llave para avanzar y pensar en cuál es el siguiente paso que se debe realizar.

A principio del semestre se establecieron unos objetivos generales, los cuales fueron concretándose a medida que avanzaba el estudio. Desde el comienzo del proyecto se trabajó para intentar llegar a los objetivos establecidos, y estos han sido alcanzados. Se ha conseguido realizar un

estudio de genómica comparativa de la enzima catalizadora de la ruta del salicilato.

La planificación que se ha seguido también ha sido la establecida desde un inicio. Debido al desconocimiento de las herramientas útiles y necesarias para el desarrollo del trabajo se estableció un margen para la lectura de información y toma de decisiones. En algún caso se fijaron parámetros poco precisos, pero el error se corrigió pronto, se retrocedió y el incidente no supuso apenas demora. Tras realizar la planificación se propuso utilizar una herramienta bioinformática en concreto, Vector NTI. No obstante, al no conseguir poner en funcionamiento dicho software se escogió otra herramienta de características similares, más concretamente SnapGene.

Atendiendo a las líneas de trabajo futuro, sería de interés expandir el estudio al resto de proteínas y enzimas que toman parte en la ruta del naftaleno, finalizando así la investigación. En este documento no se han podido explorar el resto de enzimas debido a la falta de tiempo.

En suma, a continuación, se detallan las conclusiones brevemente resumidas:

- Los árboles filogenéticos coinciden con los resultados de las comparaciones entre operones. Los operones de las secuencias más cercanas evolutivamente son los que guardan mayor similitud.
- La proteína molde presenta una cantidad considerable de proteínas potencialmente homólogas. Esta escasa variación evolutiva demuestra la importancia de este gen para las bacterias.
- Atendiendo la localización del gen en diferentes organismos, existe una gran variabilidad.
- Los objetivos han sido alcanzados.
- La planificación ha sido correcta, potenciando la realización del trabajo.
- En un futuro seria de interés poder completar este estudio con el resto de enzimas presentes en la ruta del naftaleno.

4. Glosario

ML: Maximum-likelihood

MP: Maximum-parsimony

NJ: Neighbor-joining

MSA: Alineamiento de secuencias múltiples

5. Bibliografía

1. Cerniglia C, Biodegradation of polycyclic aromatic hydrocarbons. *Biodegradation* 3: 351–368, 1992.

2. 26. Kronenberg M, Trably E, Bernet N, Patureau D, Biodegradation of polycyclic aromatic hydrocarbons: Using microbial bioelectrochemical systems to overcome an impasse, Environ Pollut, 509-523, 231, 2017.

3. Doley R, Barthakur M, Goswami BS. Microbial Degradation of Aromatic hydrocarbon: Naphthalene through Nocardiopsis alba RD3. *Int. J. Curr. Microbiol. Appl. Sci.* **6**:1174–1181, 2017.

4. 27. Meckenstock RU, Boll M, Mouttaki H, Koelschbach JS, Cunha TP, Weyrauch P, Dong X, Himmelberg AM, Anaerobic Degradation of Benzene and Polycyclic Aromatic Hydrocarbons, J Mol Microbiol Biotechnol, 92-118, 26(1-3), 2016.

5. Seo J, Keum Y, Li Q, Bacterial Degradation of Aromatic Compounds, *Int. J. Environ. Res. Public Health*, 6, 278- 309, 2009.

6. Hassanshahian M, Abarian M, Cappello S, Biodegradation of Aromatic Compounds. Biodegrad. *Bioremediation Polluted Syst. - New Adv. Technol*, 2015.

7. Yen K, Serdar CM, Genetics of naphthalene in pseudomonads, 1988.

8. Stohs, S. J., Ohia, S. & Bagchi, D, Toxicology, 180, 97–105, 2002.

9. Yen, K.-M. & Gunsalus, IC,Proc. Natl Acad. Sci. USA, 79, 874–878, 1982.

10. Harayama, S., Rekik, M., Wasserfallen, A., & Baroch, A. Mol. Gen. Genet. 210, 241-247,1987

11. Sota M *et* al, Journal of Bacteriology, 188, 4057-4067, 2006.

12. Zhao H., Chen B., Li Y. Cai B, Microbiological Research 160, 307-313, 2005

13. You I. Ghosal D., Gunsalus I. C, Biochemistry, 30, 1635-1641, 1991.

14. 21. Pearson W. R., An introduction to sequence similarity ("homology") searching. Current protocols in bioinformatics, Chapter 3, Unit3.1, 2013.

15. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. "Basic local alignment search tool." J. Mol. Biol. 215:403-410, 1990.

16. https://www.ncbi.nlm.nih.gov/books/NBK2590/

17. Khumar S, Glen S, Michael L, Christina K, Koichiro, Molecular Biology and Evolution, Volume 35, Issue 6, June 2018, Pages 1547–1549, 2018.

18. Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ, Nature Methods Supplement, vol.7 no.3, pages 16-25, 2010.

19. Anisimova, M., Cannarozzi, G., & Liberles, D, Finding the balance between the mathematical and biological optima in multiple sequence alignment. Trends in Evolutionary Biology, 2, e7, 2010

20. Kemler, M., Goker, F. Oberwinkler and D. Begerow. Implications of molecular characters for the phylogeny of the Microbotryaceae (Basidiomycota: Urediniomycetes). BMC Evolutionary Biology 6, 35, 2006.

21. Baum, D. Reading a Phylogenetic Tree: The Meaning of Monophyletic Groups. Nature Education 1(1):190, 2008.

22. Weiß M, Göker M, Molecular phylogenetic reconstruction. In: Kurtzman CP, Fell JW, Boekhout T (Eds) The Yeasts – A Taxonomic Study. 5th edn. Elsevier, Amsterdam, pp 159-174, 2011.

23. Osbourn A.E., Field B. Operons. Cell Mol. Life Sci., 66:3755–3775, 2009.

24. Wang Q, Lei Y, Xu X, Wang G, Chen LL, Theoretical prediction and experimental verification of protein-coding genes in plant pathogen genome Agrobacterium tumefaciens strain C58. PLoS One. ;7(9): e43176, 2012.

25. Mavromatis K., Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC, Rigoutsos I, Salamov A, Korzeniewski F, Land M, Lapidus A, Grigoriev I, Richardson P, Hugenholtz P, Kyrpides NC, Use of simulated data sets to evaluate the fidelity of metagenomic processing methods, Nat Methods. 2007 Jun;4(6):495-50, 2007.

26. Pierce B. A., Genetics: A conceptual approach. 2nd Edition, 2005.

27. http://www.getsynbio.com/future-here-literally-making-comparison-snapgene-genome-compiler/. 20/05/2019.

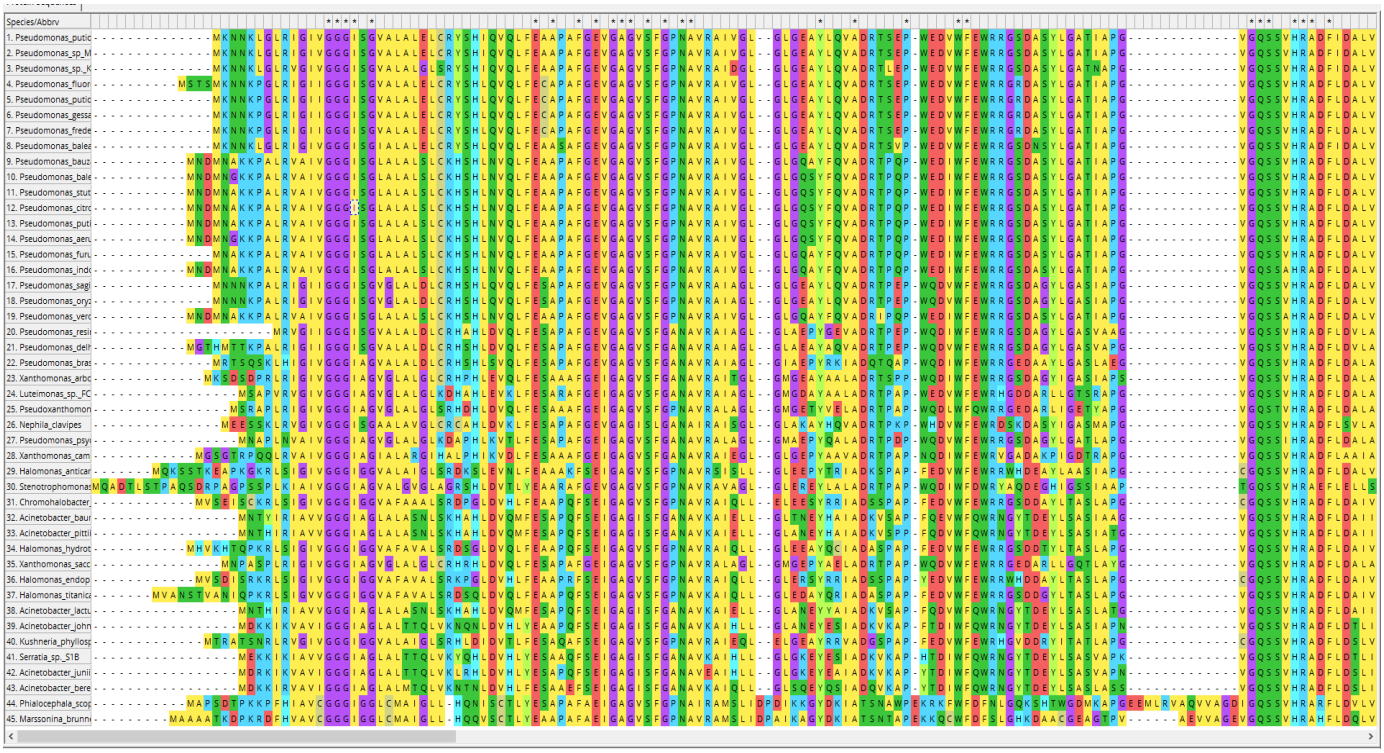
28. Sullivan MJ, Petty NK, Beatson SA, Easyfig: a genome comparison visualizer. Bioinformatics.;27(7):1009–1010, 2011,

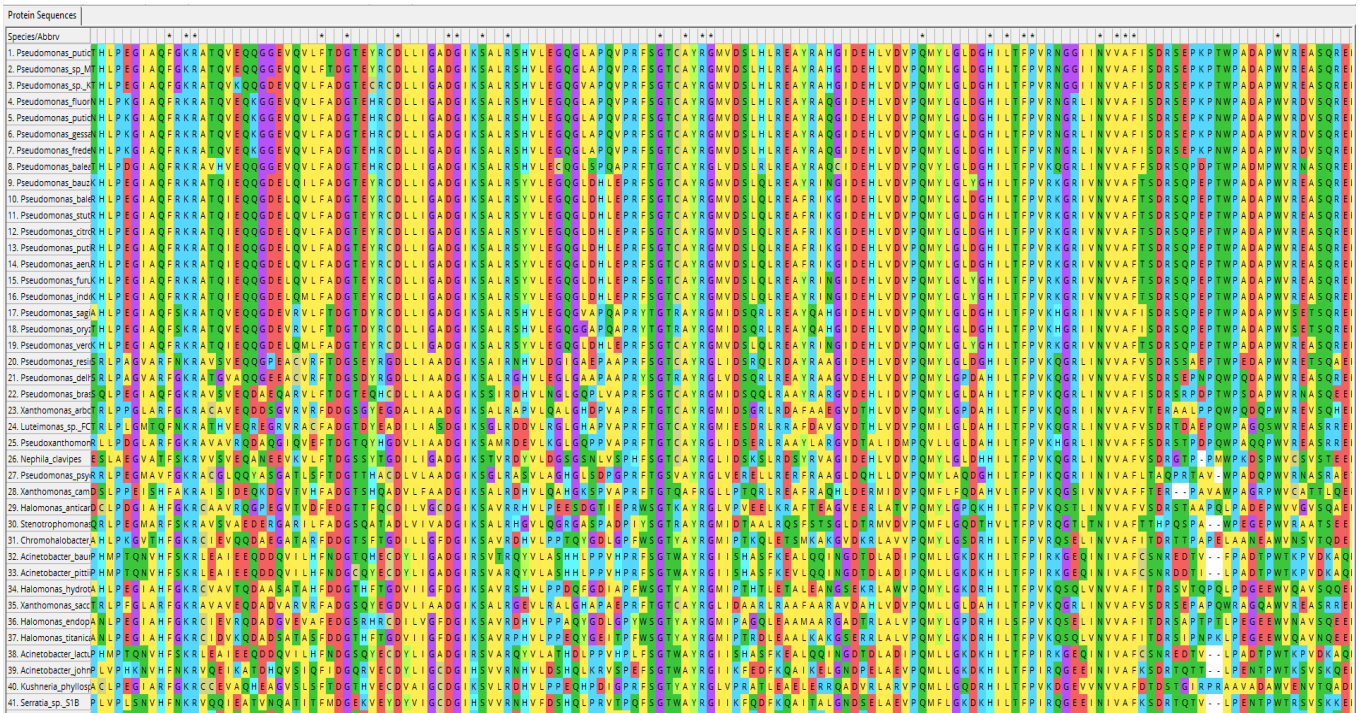
6. Anexos

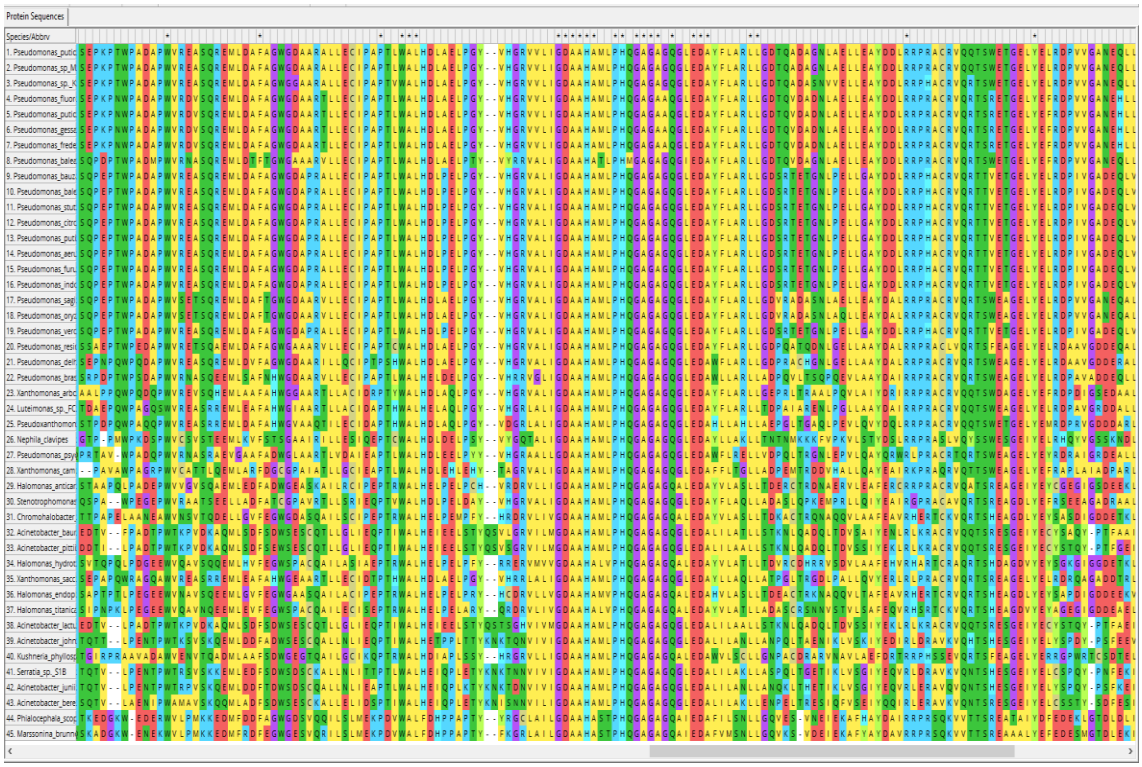
Anexo I: Base de datos del estudio

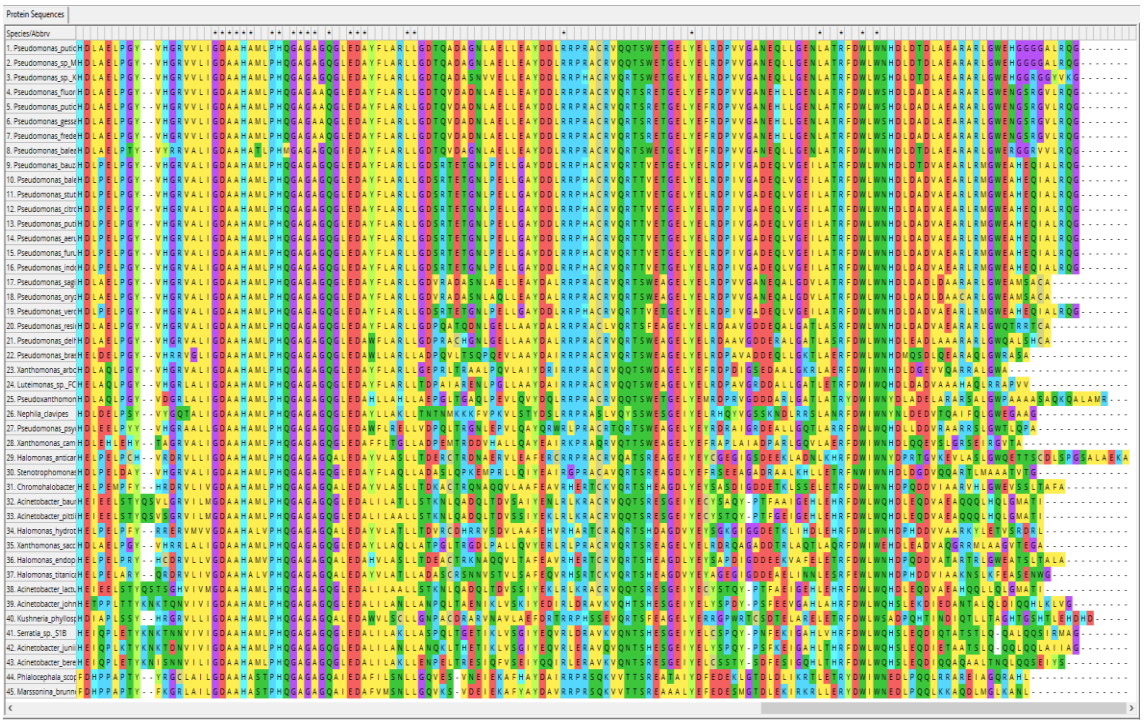
Microorganismo	Cepa	ACC num	Localización	¿Genoma cor	Identidad	Cobertura del alinean	Secuencia ADN		FgenesB	TU/OP	Genes
Proteina template											
Pseudomonas putida	G7	YP_53483	Plasmido	Si			Si		si	tu	
Resultados del BLAST											
Pseudomonas sp	MC1	WP_0114	Plasmido	Si	434(100%)	100%	Si		si	tu	
Pseudomonas sp	KB35B	ABB7220	Cromosoma	Si	410/434(94%	100%	Si		si	tu	
Pseudomonas fluorescens	PC20	YP_00288	plasmido	Si	402/434(93%	100	Si		si	tu	
Pseudomonas putida	ND6	NC_0052	plasmido	Si	402/434(93%	100	SI		si	op	2
Pseudomonas gessardii	LZ-E	KP997281	Cromosoma	Si	402/434(93%	100	No				
Pseudomonas frederiksbergensis	AS1	PRJNA34	Plasmido	Si	402/434(93%	100	Si		si	tu	
Pseudomonas balearica	LS401	LONE020	Cromosoma	Si	387/434(89%	100	No				
Pseudomonas bauzanensis	W13Z2	JFHS0100	Cromosoma	Si	367/434(85%	100	Si		si	tu	
Pseudomonas putida	CSV86	AMWJ010	Cromosoma?¿?¿	Si	366/434(84%	100	Si		si	tu	
Pseudomonas stutzeri	B1SMN1	AMWJ010	Cromosoma	Si	366/434(84%	100	No				
Pseudomonas citronellolis	SJTE-3	CP015878	Cromosoma	Si	366/434(84%	100	Si		si	tu	
Pseudomonas balearica	DSM 6083	CP007511	Cromosoma	Si	366/434(84%	100	Si		si	Tu	
Pseudomonas aeruginosa	CGMCC 1860	GQ39616	Plasmido	Si	366/434(84%	100	Si		si	tu	
Pseudomonas furukawaii	KF 707	AP014862	Cromosoma	Si	367/434(84%	100	Si		si	tu	
Pseudomonas indoloxydans	JCM 14246	QASO010	Cromosoma	si	365/434(84%	100	Si		si	tu	
Pseudomonas sagittaria	JCM 18195	FOXM010	Cromosoma	Si	373/424(88%	100	Si		si	op	2
Pseudomonas oryzae	KCTC 32247	LT629751	Cromosoma	Si	373/424(88%	100	Si		si	op	3
Pseudomonas veronii	1YB2	WP_0178	Cromosoma	Si	364/424(84%	100	Si		si	Tu	
Pseudomonas resinovorans	NBRC 106553	BAN4985	Cromosoma	Si	326/417 (78%	97	Si		No		
Pseudomonas delhiensis	CCM 7361	FNEC010	Cromosoma	Si	333/422 (79%	97	Si		si	op	2
Pseudomonas brassicacearum	38D4	RON4210	Cromosoma	Si	313/423 (74%	97	Si		si	tu	
Xanthomonas arboricola	CPBF 426	WP_1191	Cromosoma	Si	278/402 (69%	92	Si		si	op	7
Luteimonas oceanisediminis sp.	FCS-9	WP_0471	Cromosoma	Si	278/400 (70%	92	Si		si	Tu	
Pseudoxanthomonas spadix	BD-a59	WP_0141	Cromosoma	Si	271/402 (67%	92	?¿				
Nephila clavipes	Nep-004	PRD19853	Cromosoma	Si	258/408(63%	92	Si		si	Tu	
Pseudomonas psychrotolerans	SB8	KTT54448	Cromosoma	Si	258/403 (64%	92	Si		si	op	3
Xanthomonas campestris	17	AKC80633	Cromosoma	Si	241/414(58%	95	Si		si	op	5
Halomonas anticariensis	FP35	EPC01679	Cromosoma	Si	239/414 (58%	95	Si		si	Op	7
Stenotrophomonas sp	UBA9336	HBS57385	Cromosoma	Si	235/411 (57%	94	Si		si	tu	
Chromohalobacter salexigens	40a_TX	WP_1100	Cromosoma	Si	225/406 (56%	93	Si		si	Op	2
Acinetobacter baumannii	984213	WP_0338	Cromosoma	Si	218/419 (52%	95	Si		si	Op	3
Acinetobacter pittii	ARLG1942	OTU2198	Cromosoma	Si	218/419 (52%	95	Si		si	op	3
Halomonas hydrothermalis	Y2	ATH7735	Cromosoma	Si	223/401 (56%	92	Si		si	tu	
Xanthomonas sacchari	CFBP4641	PPU798	Cromosoma	Si	270/400(68%	92	Si		si	op	3
Halomonas endophytica	MC28 568	PMR7204	Cromosoma	Si	235/411(57%	94	Si		si	tu	
Halomonas titanicae	BH1	ELY22802	Cromosoma	Si	221/404(55%	94	Si		si	op	2
Acinetobacter lactucae	OTEC-02	WP_0811	Cromosoma	Si	212/416(51%	95	Si		si	op	3
Acinetobacter johnsonii	XBB1	ALV74875	Plasmido	Si	203/416(49	95	Si		si	op	2
Kushneria phyllosphaerae	EAod3	SPJ32208	Cromosoma	Si	220/411(54%	93	Si		si	op	9
Serratia sp	S1B	PVZ83448	Cromosoma	Si	200/416(48%	95	Si		si	op	3
Acinetobacter junii	AJ_351	RSE28865	Cromosoma	Si	196/412(48%	94	Si		si	op	4
Acinetobacter bereziniae	NIPH-3	WP_0048	Cromosoma	Si	196/416(47%	95	Si		si	op	3
Phialocephala scopiformis	CBS 120377	XP_0180	Cromosoma	Si	174/432(40%	95	Si		Hongo		
Marssonina brunnea	MB_m1	XP_00729	Cromosoma	Si	174/431(40%	97	Si		Hongo		

Anexo II: Alineamiento múltiple de las secuencias



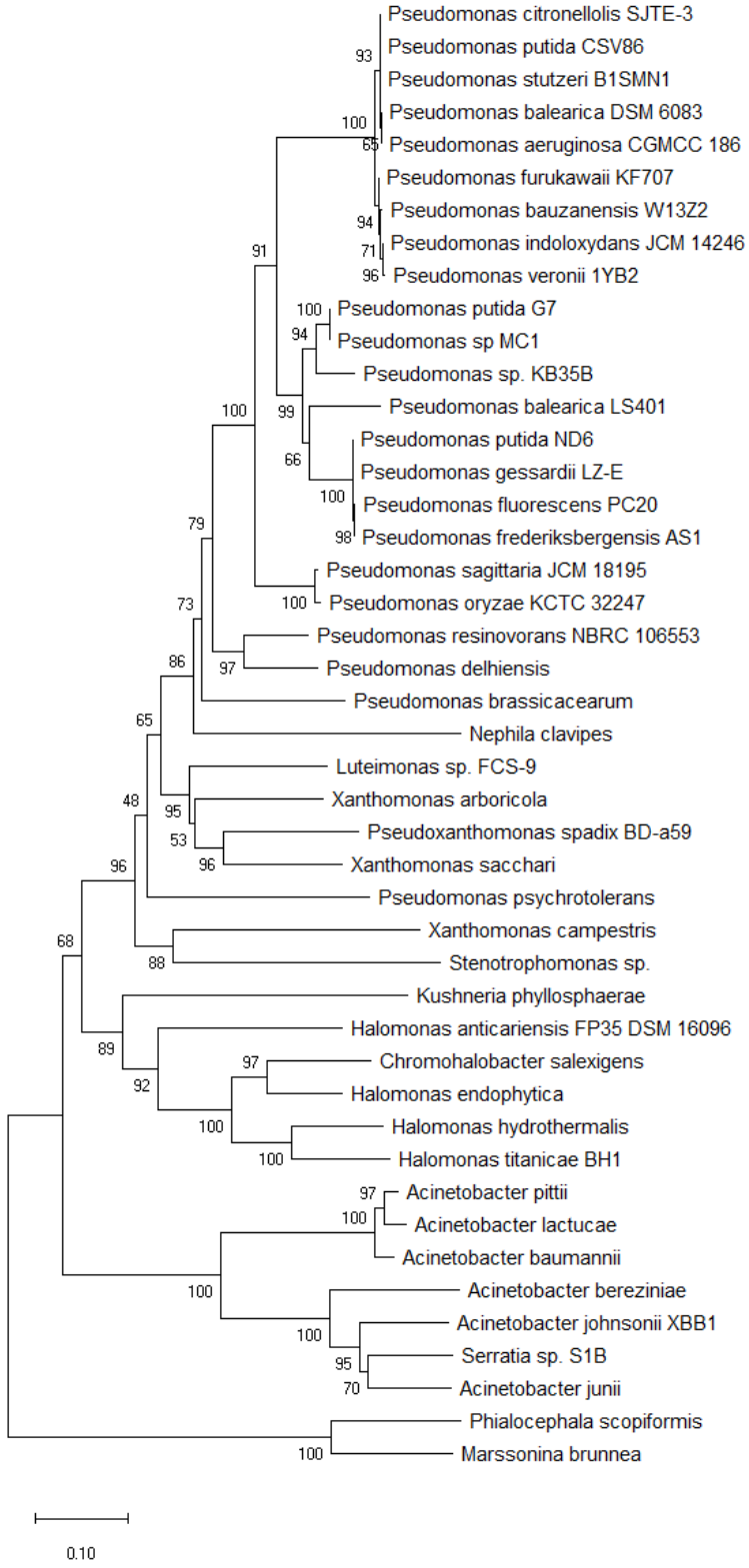



Protein Sequences
Species/Abbrv

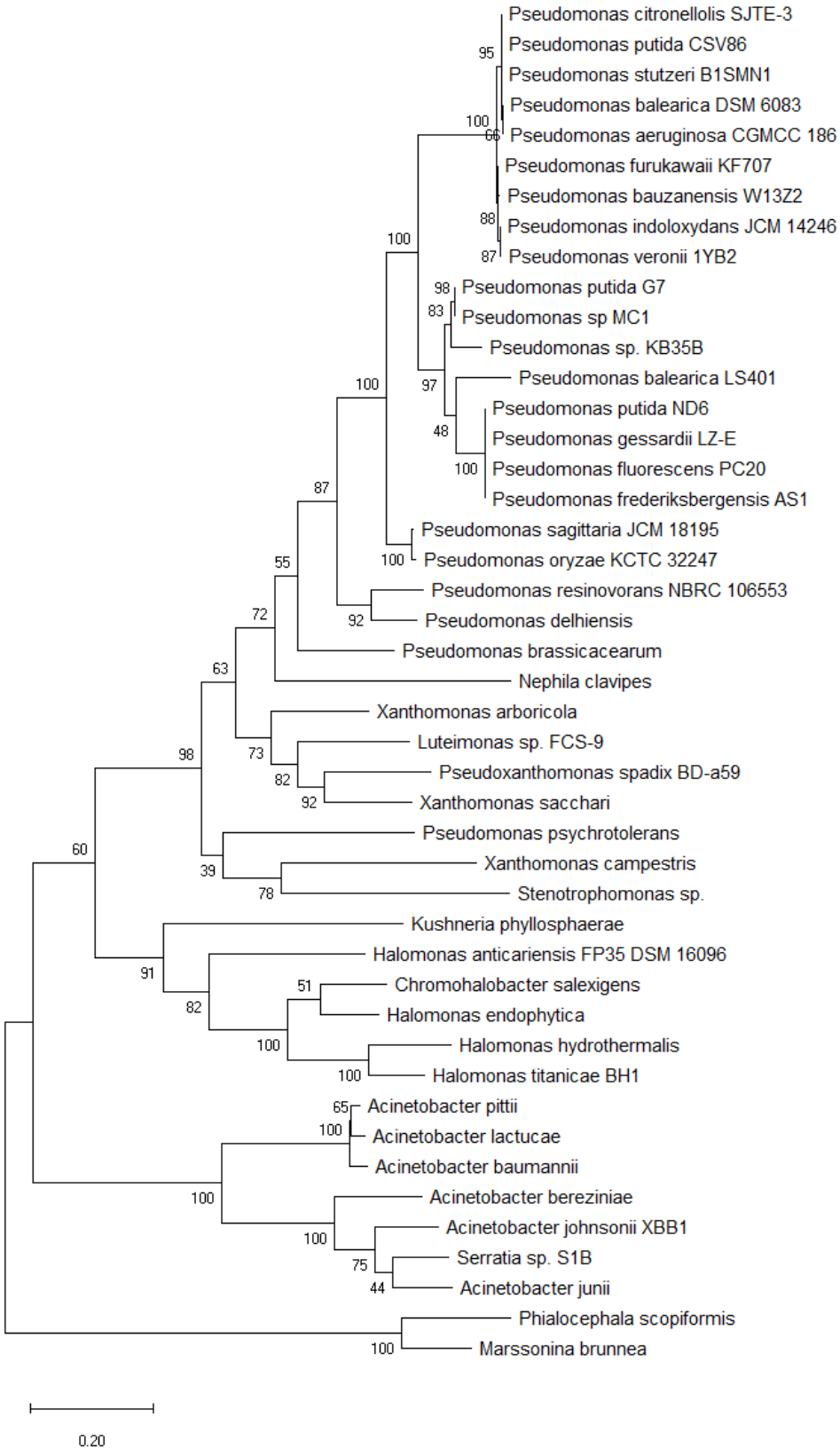

Protein Sequences
Species/Abbrv

Anexo III: Árboles filogenéticos

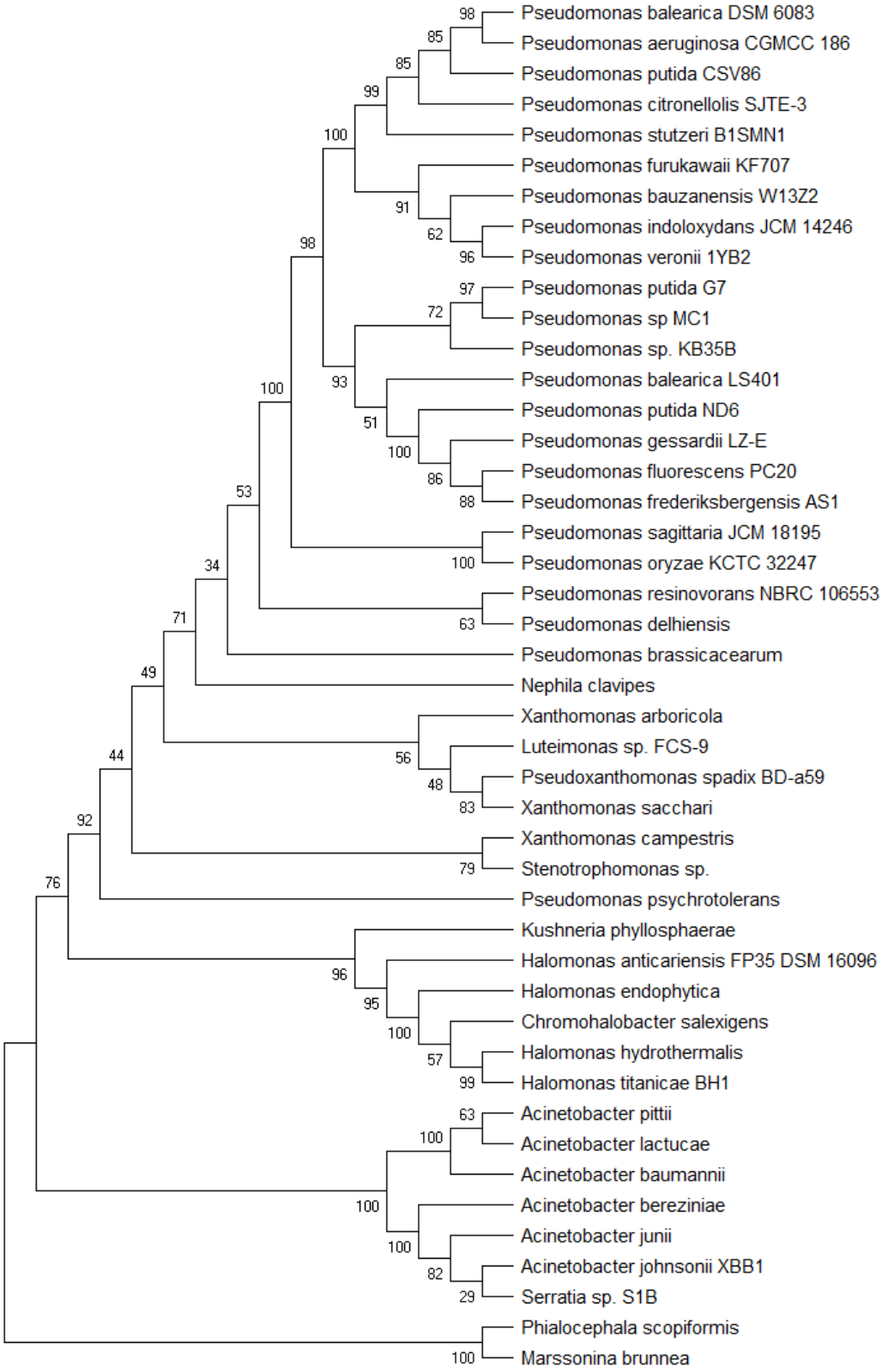
Árbol NJ



Árbol ML


Pseudomonas citronellolis SJTE-3
Pseudomonas putida CSV86
Pseudomonas stutzeri B1SMN1
Pseudomonas balearica DSM 6083
Pseudomonas aeruginosa CGMCC 186
Pseudomonas furukawaii KF707
Pseudomonas bauzanensis W13Z2
Pseudomonas indoloxydans JCM 14246
Pseudomonas veronii 1YB2
Pseudomonas putida G7
Pseudomonas sp MC1
Pseudomonas sp. KB35B
Pseudomonas balearica LS401
Pseudomonas putida ND6
Pseudomonas gessardii LZ-E
Pseudomonas fluorescens PC20
Pseudomonas frederiksbergensis AS1
Pseudomonas sagittaria JCM 18195
Pseudomonas oryzae KCTC 32247
Pseudomonas resinovorans NBRC 106553
Pseudomonas delhiensis
Pseudomonas brassicacearum
Nephila clavipes
Xanthomonas arboricola
Luteimonas sp. FCS-9
Pseudoxanthomonas spadix BD-a59
Xanthomonas sacchari
Pseudomonas psychrotolerans
Xanthomonas campestris
Stenotrophomonas sp.
Kushneria phyllosphaerae
Halomonas anticariensis FP35 DSM 16096
Chromohalobacter salexigens
Halomonas endophytica
Halomonas hydrothermalis
Halomonas titanicae BH1
Acinetobacter pittii
Acinetobacter lactucae
Acinetobacter baumannii
Acinetobacter bereziniae
Acinetobacter johnsonii XBB1
Serratia sp. S1B
Acinetobacter junii
Phialocephala scopiformis
Marssonina brunnea
0.20

Árbol MP


Pseudomonas balearica DSM 6083
Pseudomonas aeruginosa CGMCC 186
Pseudomonas putida CSV86
Pseudomonas citronellolis SJTE-3
Pseudomonas stutzeri B1SMN1
Pseudomonas furukawaii KF707
Pseudomonas bauzanensis W13Z2
Pseudomonas indoloxydans JCM 14246
Pseudomonas veronii 1YB2
Pseudomonas putida G7
Pseudomonas sp MC1
Pseudomonas sp. KB35B
Pseudomonas balearica LS401
Pseudomonas putida ND6
Pseudomonas gessardii LZ-E
Pseudomonas fluorescens PC20
Pseudomonas frederiksbergensis AS1
Pseudomonas sagittaria JCM 18195
Pseudomonas oryzae KCTC 32247
Pseudomonas resinovorans NBRC 106553
Pseudomonas delhiensis
Pseudomonas brassicacearum
Nephila clavipes
Xanthomonas arboricola
Luteimonas sp. FCS-9
Pseudoxanthomonas spadix BD-a59
Xanthomonas sacchari
Xanthomonas campestris
Stenotrophomonas sp.
Pseudomonas psychrotolerans
Kushneria phyllosphaerae
Halomonas anticariensis FP35 DSM 16096
Halomonas endophytica
Chromohalobacter salexigens
Halomonas hydrothermalis
Halomonas titanicae BH1
Acinetobacter pittii
Acinetobacter lactucae
Acinetobacter baumannii
Acinetobacter bereziniae
Acinetobacter junii
Acinetobacter johnsonii XBB1
Serratia sp. S1B
Phialocephala scopiformis
Marssonina brunnea