UNIVERSIDAD AUTÓNOMA METROPOLITANA
UNIDAD XOCHIMILCO
DIVISIÓN DE CIENCIAS BIOLÓGICAS Y DE LA SALUD
DEPARTAMENTO DE SISTEMAS BIOLÓGICOS

QUÍMICO FARMACÉUTICO BIÓLOGO

INFORME DE ACTIVIDADES DEL SERVICIO SOCIAL:

DESARROLLO DE UN ALGORITMO BIOINFORMÁTICO PARA IDENTIFICAR REGIONES ÚNICAS EN FAMILIAS DE GENES HOMÓLOGOS: UNA HERRAMIENTA CON APLICACIONES EN BIOMEDICINA Y BIOTECNOLOGÍA

PERTENECIENTE AL PROYECTO GENÉRICO:

"Obtención de materias primas, principios activos, medicamentos y productos biológicos"

ALUMNO: **Leticia Martinez Esquivel**

MATRÍCULA: 2173082723

ASESORES: **Dr. Jesús Eduardo Zúñiga León**
Dr. Juan Esteban Barranco Florido

LUGAR DE REALIZACIÓN: Universidad Autónoma Metropolitana Unidad Xochimilco, Laboratorio N-104, Edificio N, Departamento de Sistemas Biológicos.

FECHA DE INICIO Y TERMINACIÓN: Inicio 17 de noviembre del 2021. Terminación 17 de mayo del 2022.

Índice

1. Introducción

La homología es la base de la biología comparada, y el reconocimiento de conjuntos de genes o proteínas homólogos, incluida la anotación del genoma, la inferencia filogenética y los estudios de la estructura de las proteínas (13).
La clasificación de familias de genes, es decir, la agrupación de genes o proteínas en familias a menudo proporciona información importante sobre la evolución y la función de los genes. Además, es un primer paso crítico en muchas áreas de la biología comparativa y la bioinformática (4).
Por lo tanto, la clasificación automatizada de familias de genes es muy deseable, en los últimos 20 años se han desarrollado muchos métodos basados en secuencias para la clasificación automatizada de familias de genes (15). Estos métodos se pueden dividir en tres categorías principales: La primera categoría usa árboles filogenéticos para inferir familias de genes. La segunda categoría agrupa genes según similitudes con firmas de secuencias conocidas como motivos o dominios y la tercera categoría, se basa en comparaciones por pares de secuencias de proteínas de longitud completa y, por lo general, implican el uso de técnicas de agrupamiento (17).

2. Marco Teórico

Las estadísticas basadas en kmers se han convertido en una parte fundamental de las comparaciones de secuencias basadas en genes y del genoma completo sin alineamiento, primero cuentan la frecuencia de Kmers y luego correlacionan estos vectores de frecuencia de Kmers (o sus transformaciones) para evaluar la similitud u homología entre dos secuencias biológicas (14).

Los Kmers son subsecuencias cortas de longitud k contenido dentro de una secuencia de ADN (AGTC) o proteínas. Principalmente se utilizan en el contexto de la genómica y el análisis de secuencias. Sin embargo, también tienen otras aplicaciones como por ejemplo en: metagenómica, biotecnología y genética.

Metagenómica

La variación de frecuencia y espectro de Kmers se usan en metagenómica para el análisis y el agrupamiento de las lecturas de secuenciación en cada organismo, las cuales posteriormente se ensamblarán, así mismo, los Kmers se emplean en la recuperación de marcos de lectura a partir de lecturas sin procesar, estimación de la abundancia de especies en muestras metagenómicas, determinación de qué especies están presentes en las muestras y la identificación de biomarcadores presentes en las enfermedades (8).

Biotecnología

Los Kmers en las secuencias de ADN se han utilizado en aplicaciones biotecnológicas para controlar la eficiencia de la traducción. Específicamente, para regular el aumento y la disminución de producción de proteínas (23). Con respecto al aumento de la producción de proteínas, se ha utilizado la reducción de la frecuencia de dinucleótidos para producir tasas más altas de síntesis de proteínas. Además, el sesgo de uso de codones se ha modificado para crear secuencias sinónimas con mayores tasas de expresión de proteínas. La aplicación más estudiada de *Kmers* para disminuir la eficiencia de traducción es la manipulación de pares de codones para atenuar virus con el fin de crear vacunas (1).

Genética

En genética los Kmers son utilizados para la: cuantificación de isoformas de ARN a partir de datos de RNA-seq, clasificación del haplogrupo mitocondrial humano, detección de sitios de recombinación en genomas, estimación del tamaño del genoma utilizando la frecuencia de Kmers, detección de *novo* de secuencias repetidas como elementos transponibles (21).

Debido a que el número de Kmers crece exponencialmente para valores de *k*, contar *Kmers* para valores grandes de *k* (normalmente>10) es una tarea computacionalmente difícil. Para solucionar este problema se han desarrollado diversas herramientas:

- Jellyfish utiliza una tabla hash multiproceso sin bloqueos para el conteo de *Kmers* en los lenguajes de programación Python, Ruby y Perl (19).
- KMC es una herramienta para el conteo de Kmers que utiliza una arquitectura multidisco para una velocidad optimizada (7).
- Gerbil utiliza un enfoque de tabla hash pero con soporte adicional para la aceleración de GPU (10).
- K-mer Analysis Toolkit (KAT) utiliza una versión modificada de Jellyfish para analizar los recuentos de Kmers (18).

2.1 Uso de Python y Jupyter Notebook en el contexto biológico

La informática ha revolucionado las ciencias biológicas en las últimas décadas, de modo que las investigaciones en biología molecular, bioquímica y otras biociencias utilizan programas informáticos. Uno de ellos es Python el cual se ha convertido en un lenguaje de programación popular en las biociencias, por su semántica sencilla y su sintaxis limpia lo convierten en un lenguaje de fácil acceso, es expresivo y se adapta a la programación orientada a objetos, paradigmas modernos; y las bibliotecas disponibles y los conjuntos de herramientas de terceros los cuales amplían la funcionalidad del lenguaje central a prácticamente todos los dominios biológicos (análisis de secuencias y estructuras, filogenómica, sistemas de gestión de flujo de trabajo, etc.). Por otro lado, Jupyter es una herramienta web de código abierto, que los investigadores usan para combinar códigos de software, salida computacional, texto enriquecido y recursos multimedia, lo cual hace más fácil la interpretación de varios elementos en un solo documento (20). Los avances computacionales que más se han llevado a cabo han sido en hardware, software y algoritmos (9).

3. Justificación

Una familia de genes hace referencia a un grupo de genes homólogos funcional y estructuralmente relacionados (6). En otras palabras, estos genes que codifican para proteínas comparten similitud de secuencia, unidades funcionales e incluso comparten patrones de interacción (11). Además, se ha propuesto que estos grupos de genes se forman a partir de la duplicación de un gen original (3).

La clasificación de estos genes o proteínas en familias ha permitido comprender mejor su importancia biológica, y esto se debe en gran medida a la información sobre su actividad, estructura, función y papel metabólico (24). El estudio individual de genes que pertenecen a una familia puede resultar complejo ya que presentan regiones conservadas, sin embargo, la identificación de regiones cortas únicas dentro de estos genes puede resolver este problema. El uso de oligonucleótidos juega un papel fundamental en el estudio y caracterización de genes en general. Los oligonucleótidos, también conocidos como primers, cebadores o iniciadores, son secuencias cortas de ADN de cadena simple (diseñados artificialmente) que se utilizan en la reacción en cadena de la polimerasa ("Polymerase Chain Reaction: PCR"). El uso de estas moléculas sintéticas ofrece varias ventajas incluyendo la identificación de varios genes a la vez usando secuencias degeneradas o identificación de genes individuales usando secuencias altamente específicas. Estas secuencias permiten identificar genes y/o regiones de interés dentro de todo un genoma en una amplia variedad de organismos. Aunque el uso de los oligonucleótidos se encuentra ampliamente extendido, su diseño es una tarea tediosa, sobre todo cuando se intenta identificar un gen que pertenece a una familia de genes homólogos. El diseño y uso de oligonucleótidos degenerados es un enfoque desarrollado para identificar varios genes a la vez o genes homólogos (16), sin embargo, si se trata de identificar genes individuales a partir de una familia de genes esta no es la mejor opción. Por lo tanto, a pesar de las tecnologías disponibles para identificar genes a gran escala (por ejemplo, la secuenciación masiva) muchos laboratorios prefieren realizar estudios individuales de genes haciendo necesario el diseño y uso de oligonucleótidos. Por esta razón es necesario el desarrollo de algoritmos bioinformáticos que faciliten el diseño de oligonucleótidos para identificar y caracterizar genes individuales a partir de una familia de genes homólogos.

4. Objetivos

4.1 Objetivo general

Desarrollar un algoritmo bioinformático para identificar regiones únicas en familias de genes homólogos para su aplicación en biomedicina y biotecnología.

4.2 Objetivos específicos

1. Adquirir conocimientos básicos y avanzados del lenguaje de programación Python.
2. Desarrollar un algoritmo en lenguaje Python para identificar regiones de ADN en secuencias fasta.
3. Determinar secuencias únicas dentro de genes homólogos usando la familia del complejo Velvet de *Aspergillius nidulans.*
4. Identificar genes de la familia Egh16 en hongos de los órdenes hipocreales y eurotiales.
5. Validar la funcionalidad del algoritmo con la familia de genes Egh16.
6. Diseñar un par de oligonucleótidos a partir de las regiones únicas identificadas y validar su especificidad en la familia de genes Egh16.

5. Materiales y métodos

5.1 Microorganismo

Beauveria bassiana cepa CHE-CNRCB 546 proporcionada por SENASICA.

5.2 Medios de cultivo

- PDA: Extracto de papa 4 g, Dextrosa 20 g, Agar 15 g.
- SDAY: Dextrosa 40 g, Peptona 10 g, Extracto de levadura 10 g, Agar 15 g.

5.3 Cultivo Monospórico

1. Siembra.
Sembrar el hongo aislado en medio SDAY.
Incubar a 28°C durante 7 días (hasta alcanzar una esporulación óptima).

2. Preparar una solución de Tween 80 al 0.1%.
Esta solución se puede mantener en stock en refrigeración (4 °C).
Esterilizar la solución en la autoclave a 15 libras de presión durante 15 minutos.

3. Conteo de esporas.
En un tubo de 50 ml agregar 5 mL de la solución de Tween al 0.1%.
Colocar una pequeña porción del hongo y agitar ligeramente para que se separen todas las esporas. Agitar en un vórtex por espacio de 15 segundos. Cargar 100 µl de la suspensión de esporas en una cámara de Neubauer y contar el número con la ayuda de un microscopio óptico.
Si la suspensión de esporas está concentrada hacer una dilución (100 µl de la suspensión anterior y agregar 900 µl de agua destilada).
Hacer las diluciones que sean necesarias para tener una suspensión a una concentración de 50 a 100 esporas por mililitro.

$$\frac{Esporas}{Cuadro}\ x\ 25\ x\ 100\ x\ FD = \frac{Esporas}{mL}$$

Donde Esporas/Cuadro corresponde al promedio de conidios contadas en cinco cuadros; 25 corresponde al total de cuadros centrales (equivalente a 0.1 mm3); 1000 corresponde a una equivalencia (1000 µL/mL); FD corresponde al factor de dilución de la muestra.

4. Siembra de esporas.
Sembrar 100 µl de la concentración deseada en una placa con medio PDA y distribuirla sobre la superficie de la placa dentro de la cámara de flujo laminar.
Realizar la siembra por cuadruplicado e incubar a 28°C durante 7 días.
Monitorear el cultivo a partir de las 24 h de incubación con el objetivo de 10x de un microscopio óptico.

5. Selección de colonia
Cortar con una hoja de bisturí estéril una colonia en formación (2 mm² aprox.) y transferirla a otra placa con PDA (la idea es que la colonia seleccionada provenga de una sola espora).
Todo lo anterior se debe realizar en una campana de flujo laminar, en condiciones de esterilidad.

5.4 Reacción en cadena de la polimerasa (PCR)

Esta técnica se utilizó para amplificar una región del gen BBA_03121 con las condiciones que se mencionan en la tabla 1.
La PCR se basa en una reacción enzimática *in vitro* que amplifica millones de veces una secuencia específica de ADN durante varios ciclos. Requiere de elementos específicos como el templado o molde (ADN o ADNc), una enzima, oligonucleótidos o primers, desoxirribonucleótidos trifosfatados (dNTP´s) y una ADN polimerasa termoestable (22). La reacción de síntesis de ADN se consigue mediante tres pasos de repetición (Fig. 1): 1) desnaturalización del molde en cadenas sencillas a altas temperaturas (94°C); 2) hibridación (50°C) de los cebadores con cada hebra original para la síntesis de una nueva hebra; y 3) extensión (72°C) de las nuevas hebras de ADN de los cebadores (5).

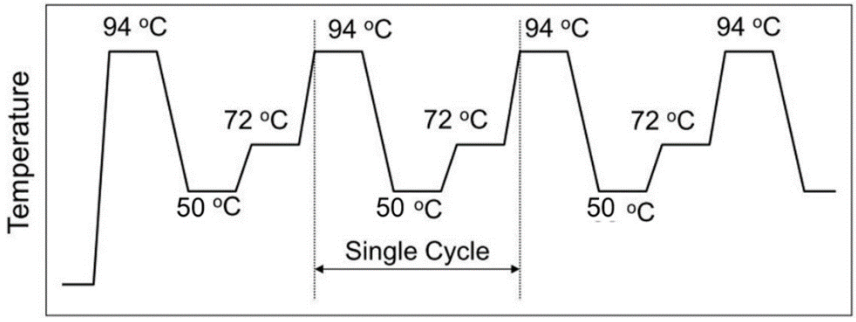

Fig. 1. Técnica de PCR. Condiciones utilizadas para la amplificación del gen BBA_03121 de la subfamilia de la familia Egh16 mediante múltiples ciclos de polimerización.

Tabla 1. Condiciones de PCR para la amplificación de una región del gen BBA_03121.

Componente	50-µl rxn	Concentración final
10x PCR Buffer - Mg	5 µl	1x
50 mM *MgCl2*	1.5 µl	1.5 mM
10 mM dNTP Mix	2 µl	0.2 mM
10 µM Forward primer	2.5 µl	0.5 µM
10 µM Reverse primer	2.5 µl	0.5 µM
Templado DNA	1 µl	1-500 ng
Taq DNA polimerasa (5 U/ µl)	1 µl	1.0 -2.5 U/rxn *
Agua estéril	34.5 µl	-

*Utilizar hasta 2.5 U para objetivos más largos

5.4.1 Diseño de oligonucleótidos

Para la identificación del gen BBA_03121 perteneciente a la subfamilia de la familia Egh16, se diseñó un par de oligonucleótidos con las siguientes especificaciones: longitud 20 - 21, temperatura de fusión (Tm 54 ± 1 °C), porcentaje de GC menor a 60 %, tamaño del amplicón (300 - 500 pb) y estabilidad del extremo 3' esto con el fin de evitar hairpins y otras estructuras secundarias como homodímeros y heterodímeros, y así obtener una buena amplificación del fragmento deseado.
El par de primers diseñados fueron Bb_RT1F (CATGGTATACCGCCAGATCAA) de 21 nucleótidos y Bb_RT1R (CAATCTCCATGCGCTTCTTG) de 20 nucleótidos.

5.5 Purificación de ADN

5.5.1 Extracción de ADN genómico a partir de *B. bassiana*

1. El micelio se molió en un mortero utilizando N2 líquido y se hicieron alícuotas de 0.5 ml del volumen del tubo eppendorf previamente enfriados con N2 líquido.
2. A cada tubo, se le adicionaron 800 µl de BE*, se incubó a 68°C durante 30 min.
3. Se centrifugó a 14,000 rpm durante 5 min.
4. Al sobrenadante se le agregó 100 µl de acetato de sodio 3 M pH 5.2, se mezcló y se colocó en hielo durante 1 hora (puede ser solo 10 minutos).
5. Se centrifugó 5 min a 14,000 rpm del sobrenadante se tomaron 700 µl y se le adicionaron 700 µl de isopropanol se deja así toda la noche.
6. Se centrifugó 5 min a 14,000 rpm, se desechó el sobrenadante y se obtiene una pastilla.
7. Se dejó secar el pellet y posteriormente el ADN se resuspendió en 50 µl de agua estéril.

5.5.2 Purificación de ADNg usando columnas de sílice

1. Se inició con 3 muestras 1,2,3, el primer paso fue juntar las muestras 1 y 2 en un solo tubo para incrementar el volumen de muestra dejando la muestra 3 como muestra patrón.
2. Posteriormente se añadieron al tubo 3µl de RNAsa y se mantuvo a temperatura ambiente durante 30 min.
3. Se agregaron 100 µl de tampón B3, y la muestra fue sometida a vortex brevemente de 15 a 30 seg.
(Si se observan partículas insolubles, centrifugar durante 5 min a alta velocidad aproximadamente a 11.000 rpm) y transferir el sobrenadante a un nuevo tubo de microcentrífuga.
4. Se agregaron 210 µl de etanol (96–100 %) a la muestra y se dio vórtice vigorosamente durante aproximadamente 1 min.
5. Para la muestra, se utilizó un NucleoSpin® Tissue Columna acoplado a un tubo de recolección.
6. Se descargo la muestra del paso anterior a la columna teniendo cuidado de asegurarse de cargar todo el precipitado en la columna.
7. Centrifugar durante 1 min a 10.000 rpm y desechar el exceso de alcohol conservando la columna de sílice para el siguiente paso.

8. Acoplar a la columna un nuevo tubo de recolección (incluido) con paso de flujo.
(Si la muestra no se extrae completamente a través de la matriz, repita el paso de centrifugación a 10.000 rpm)
9. Se agregaron 500 µl de tampón BW y se procedió a centrifugar durante 1 min a 10,000 rpm se deseche el flujo continuo y coloque la columna de vuelta al tubo de recolección.
10. Se agregaron 600 µl de tampón B5 a la columna y se procedió a centrifugar durante 1 min a 10.000 rpm se desechó el flujo continuo, se colocó la columna de regreso al tubo de recolección.
11. Centrifugar la columna durante 1 min a 10,000 rpm, dejar secar a temperatura ambiente durante 2 minutos para eliminar el alcohol residual.
12. Eluir con 30 µl de agua estéril primero con 15 µl y luego con otros 15 µl.
13. Finalmente revelar con gel de electroforesis.
*BE: Buffer de extracción: EDTA 50 mM y SDS 0.2%.

5.5.3 Secuencias (genomas y proteomas)

Los archivos con la etiqueta "*translated*" hacen referencia a secuencias de aminoácidos, mientras que los de la etiqueta "*rna_from_genomic*" hacen referencia a secuencias de nucleótidos. Estas secuencias fueron descargadas a partir del portal Centro Nacional para la Información Biotecnológica (NCBI).

GCA_010099065.1_ASM1009906v1_translated_cds
GCA_002871155.1_ASM287115v1_translated_cds
GCA_001682635.1_ASM168263v1_translated_cds
GCA_000770705.1_BBA1.0_translated_cds
GCA_000280675.1_ASM28067v1_translated_cds
GCA_010099065.1_ASM1009906v1_rna_from_genomic
GCA_002871155.1_ASM287115v1_rna_from_genomic
GCA_001682635.1_ASM168263v1_rna_from_genomic
GCA_000770705.1_BBA1.0_rna_from_genomic
GCA_000280675.1_ASM28067v1_rna_from_genomic

5.5.4 Bases de datos

NCBI: esta base de datos, entre otras cosas almacena secuencias de miles de organismos, y los formatos comunes son genbank y fasta.
UniProtKB: es un repositorio universas gratuito de secuencias e información de proteínas.

5.5.5 Programas

Python versión 3.6.7
Jupyter versión 6.4.6
Blastp versión 2.8.1+
Clustalo versión 1.2.4

6. Resultados y discusión

Para este trabajo se usaron los genes de la familia Egh16 de 5 cepas de *B. bassiana* secuenciadas y depositadas en la base de datos NCBI. Los miembros de la familia Egh16/Egh16H están implicados en la formación de apresorios y en la patogenicidad de los hongos filamentosos fitopatógenos y entomopatógenos, principalmente (2). Egh16H tiene una alta homología con las secuencias de *Magnaporthe grisea* y otros hongos patógenos de plantas, así como secuencias tanto del patógeno de insectos *Metarhizium anisopliae* como del patógeno humano *Aspergillus fumigatus*. No se encontraron homólogos cercanos de Egh16H en los hongos no patógenos *Neurospora crassa* y *Aspergillus nidulans* (12).

6.1 Descarga de secuencias, procesamiento, y obtención de subfamilias

Se descargaron 5 genomas y 5 proteomas en formatos FNA y FAA, respectivamente, estas se obtuvieron de la base de datos NCBI. Las secuencias de proteomas se convirtieron a base de datos BLAST con la aplicación makeblastdb v2.8.1+. Para las secuencias de genomas se extrajeron los nucleótidos (de forma iterada utilizando los leguajes de programación Jupyter v6.4.6 y Python v3.6.7) dejando únicamente el Locus_tag y la secuencia de nucleótidos en formato fasta.

6.2 Árbol filogenético

En este trabajo se estudió una subfamilia de la familia Egh16 en *B.bassiana*. Se realizó un blastp, y después de procesar la información se obtuvieron 13 secuencias pertenecientes a la familia Egh16. Posteriormente se llevó a cabo un alineamiento de secuencias con las 13 secuencias. A partir de estas 13 secuencias de nucleótidos recabadas de 5 cepas de *B. bassiana* se obtuvo el árbol filogenético de la (Fig. 3.A), la subfamilia que se encuentra en la figura 3B (recuadro azul) es con la que se realizaron los estudios del presente trabajo. Con este árbol filogenético se agruparon las subfamilias de la familia Egh16.

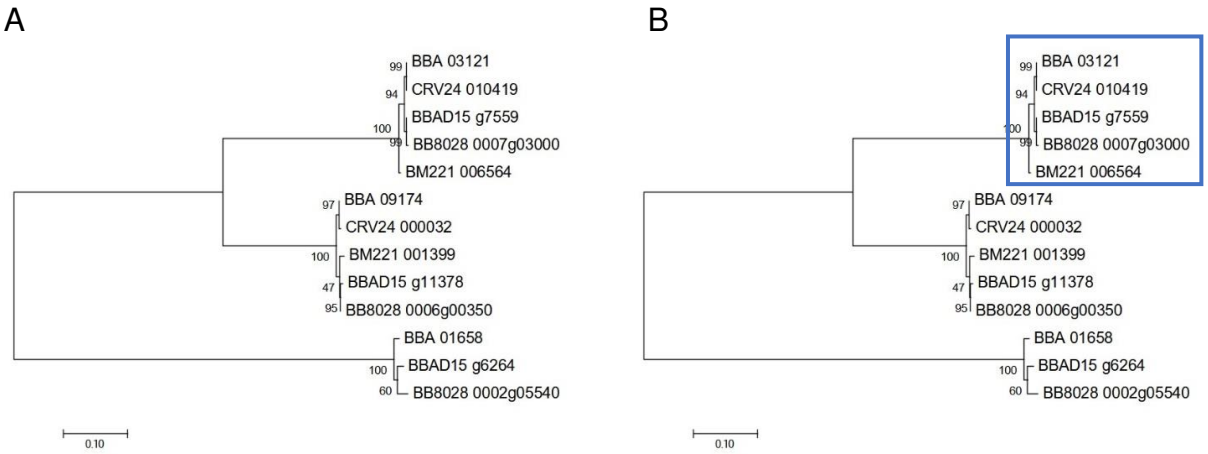

Fig. 3. Agrupación de genes. A) Árbol filogenético mostrando las subfamilias a partir de la familia Egh16. B) Árbol filogenético mostrando la subfamilia estudiada en este proyecto (Recuadro azul).

6.3 Desarrollo del Algoritmo Bioinformático

Para la identificación de Kmers compartidos se hizo un análisis seleccionando una longitud de 21 nucleótidos, para enumerar e identificar posiciones de inicio y término, este proceso se iteró con cada una de las secuencias de nucleótidos. Los Kmers obtenidos fueron generados a partir de los grupos identificados en el árbol filogenético.
Después de haber realizado el análisis filogenético, el algoritmo toma como input los grupos de genes, que se obtuvieron de dicho análisis.
El algoritmo consta de 4 partes:

1. La identificación de oligos forward.

En donde se utilizaron los parámetros descritos en el apartado 5.4.1 de no cumplir con ellos los oligos se descartan.

2. Identificación de oligos reverse.

Se obtuvieron con los parámetros descritos en el apartado 5.4.1 y con los oligos forward obtenidos anteriores a este paso.

3. Formación de estructuras secundarias

Con las posibles predicciones de los oligos forward y reverse, se buscó si en las últimas 7 bases complementarias se formaban estructuras secundarias (Hairpin, homodímeros y heterodímeros) que puedan afectar la PCR. Lo anterior dicho se ilustra en la (Fig. 2).

4. Predicción de los mejores pares de oligos compatibles.

Se eligió el par de oligos que no formaran estructuras secundarias, esto con el fin de tener una amplificación adecuada.

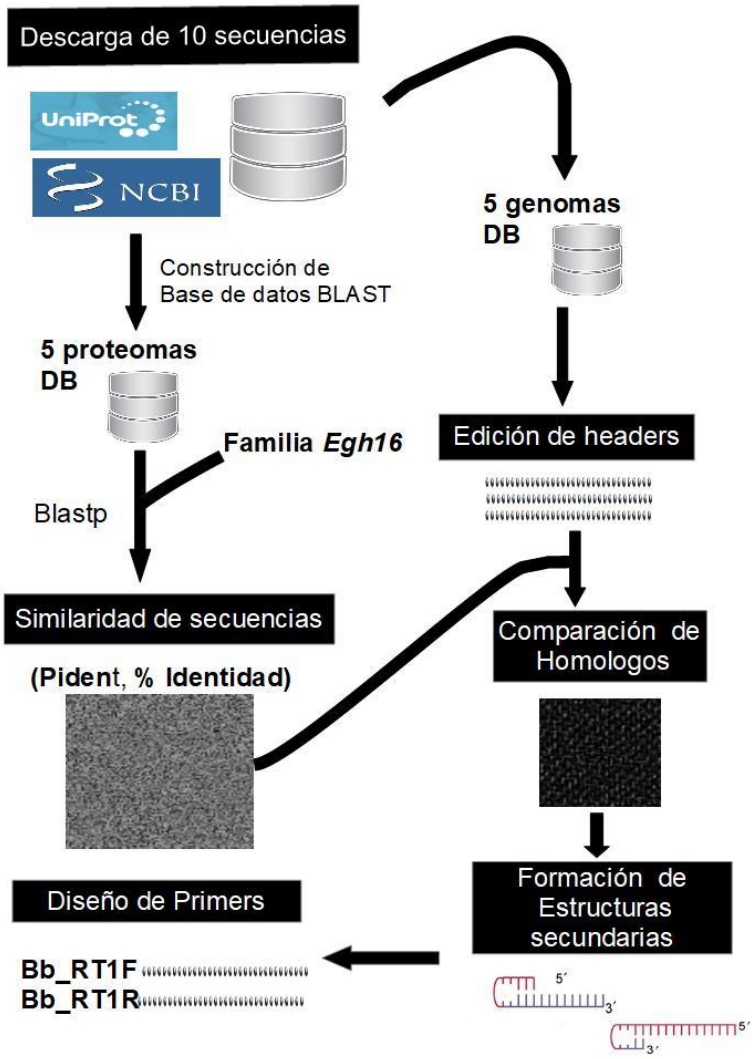

Fig 2. Estrategia del Análisis Bioinformático. A partir de secuencias se crea una base de datos (DB), con la cual se busca encontrar una similitud con el grupo de secuencias de la familia Egh16, posteriormente se hace una comparación de genes homólogos, formación de estructuras secundarias y una predicción de los posibles oligos.

6.4 Identificación *in silico* de oligos

La identificación y diseño de los oligos se realizaron usando los parámetros descritos en el apartado 5.4.1. Posterior a su diseño se hizo una validación de secuencias usando el programa Blastn contra los 5 genomas de *B. bassiana* (TaxId:176275). Después de realizar el Blastn se encontró que los oligos alinean en el gen BBA_03121 y a sus homólogos en otras cepas. En la (Fig. 4) se puede observar los posibles oligos a utilizar para la ampliación de una región del gen BBA_03121 y posibles homólogos en cepas de *B. bassiana*.

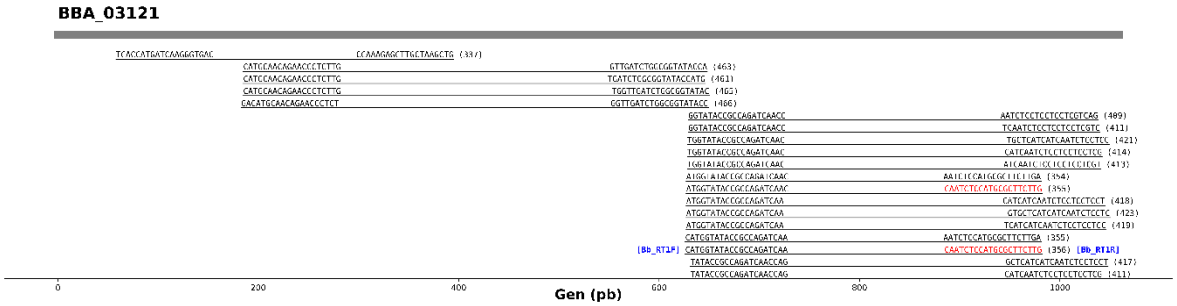

Fig. 4. Predicción de oligos. Los oligos que se muestran en esta figura están ordenados en el sentido 5' a 3', el nombre del gen mostrado (BBA _03121) se usó como molde para la predicción del par de oligos, entre paréntesis se muestran los tamaños de los amplicones generados y en color azul se indican los nombres de los oligos (Bb_RT1F y Bb_RT1R) utilizados en este trabajo para la identificación de una de las subfamilias de la familia de genes Egh16 en la cepa de *B. bassiana* CHE-CNRCB 546, la cual no se encuentra secuenciada.

6.5 Validación por PCR

La amplificación de la región del gen BBA_03121 se realizó con las especificaciones descritas en la tabla 1. Con este ensayo se pudo observar que hubo una amplificación eficiente usando la concentración 1.5 mM de MgCl2 descrita (Fig. 5), ya que se observan bandas específicas.

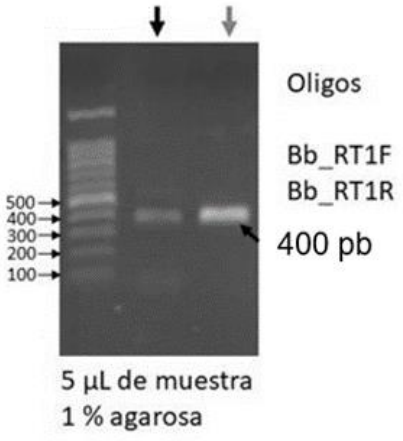

Fig.5 Análisis electroforético de la PCR. Se utilizaron 5 µl de muestra en cada carril, las flechas que se encuentran arriba, corresponden a los amplicones obtenidos a partir de los oligos Bb_RT1F y Bb_RT1R, el gel de agarosa que se utilizó fue al 1 %.

6.6 Secuenciación de ADN

Un segmento del gen homólogo de BBA_03121 identificado en la cepa de *B. bassiana* CHE-CNRCB 546 fue secuenciado para corroborar la similaridad entre las secuencias. La secuenciación del ADN se realizó por la empresa Macrogen (South Korea) y se utilizaron los oligonucleótidos (Bb_RT1F y Bb_RT1R). En la (Fig. 6) se muestra el electroferograma mostrando una buena resolución de las bases identificadas en la cepa de *B. bassiana* CHE-CNRCB 546 son similares.

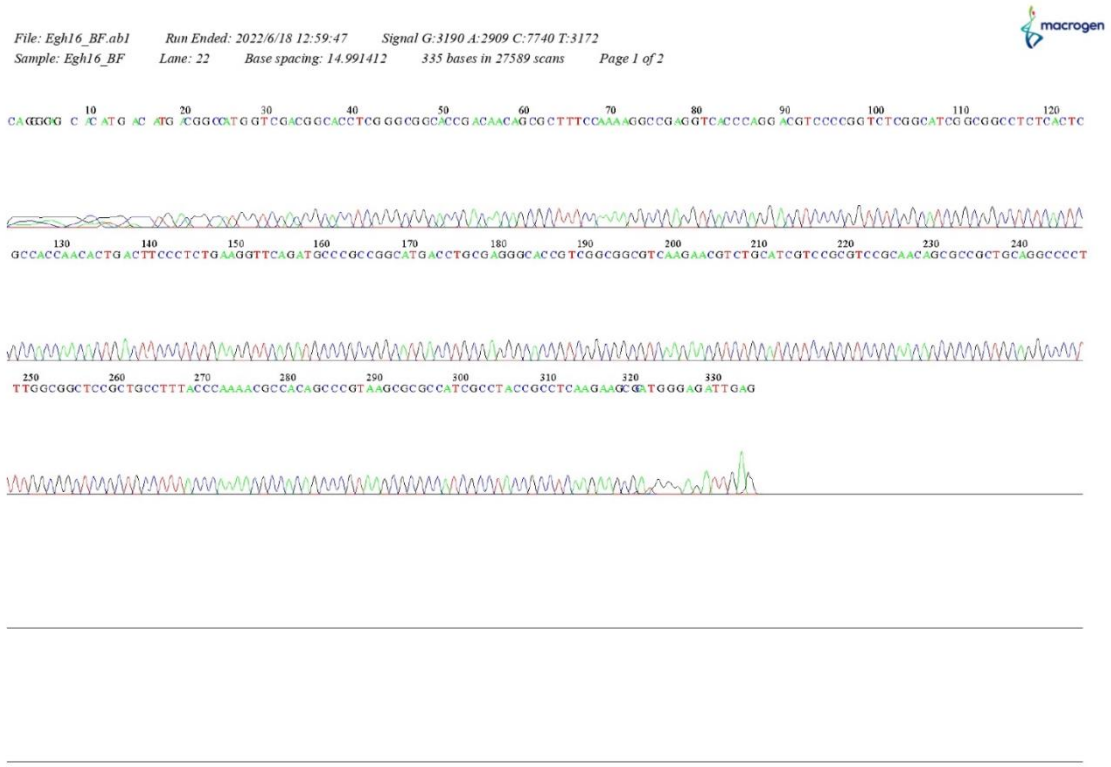

Fig.6. Electroferograma mostrando la secuencia de un gen homólogo de BBA_03121 identificado en la cepa CHE-CNRCB 546 de *B.bassiana.*

7. Conclusiones

- Los conocimientos adquiridos del lenguaje de programación Python permitieron desarrollar un algoritmo para identificar regiones específicas dentro de genes homólogos pertenecientes a una familia.
- Se identificó la familia Egh16 (constituida por 3 subfamilias) en cinco cepas de *B. bassiana* secuenciadas depositadas en la base de datos NCBI, las cuales pertenecen al orden de los Hypocreales.
- Se construyeron los oligonucleótidos Bb_RT1F y Bb_RT1R con el algoritmo desarrollado en este trabajo, con los cuales se identificó el gen BBA_03121 perteneciente a la familia de genes Egh16. BBA_03121 es homólogo de XP_007807765 localizado en el hongo *Metarhizium acridum*.
- La especificidad de los primers permitió realizar la PCR de forma eficiente, lo cual representa un resultado importante ya que el algoritmo predice y diseña primers altamente específicos para identificar genes de familias con un grado de homología.
- Se realizó un análisis *in silico* para validar la funcionalidad del algoritmo usando la familia de genes del complejo Velvet de *Aspergillus nidulans* para diseñar oligonucleótidos específicos para cada gen de la familia.

8. Referencias

1. Al-Saif, M., & Khabar, K. S. (2012). UU/UA dinucleotide frequency reduction in coding regions results in increased mRNA stability and protein expression. Molecular therapy: the journal of the American Society of Gene Therapy, 20(5), 954–959.
2. Cao, Y., Zhu, X., Jiao, R., & Xia, Y. (2012). The Magas1 gene is involved in pathogenesis by affecting penetration in Metarhizium acridum. Journal of microbiology and biotechnology, 22(7), 889–893.
3. Carmi., & Bolshoy, A. (2016). Gene-Family Extension Measures and Correlations. Life (Basel, Switzerland, 6(3), 30.
4. Chan, C. X., Mahbob, M., & Ragan, M. A. (2013). Clustering evolving proteins into homologous families. BMC bioinformatics, 14, 120.
5. Delidow, B. C., Lynch, J. P., Peluso, J. J., & White, B. A. (1993). Polymerase chain reaction : basic protocols. Methods in molecular biology (Clifton, N.J.), 15, 1–29.
6. Demuth, J. P., De Bie, T., Stajich, J. E., Cristianini, N., & Hahn, M. W. (2006). The evolution of mammalian gene families. PloS one, 1(1), e85.
7. Deorowicz, S., Kokot, M., Grabowski, S., & Debudaj-Grabysz, A. (2015). KMC 2: Fast and resource-frugal k-mer counting. Bioinformatics, 31(10), 1569–1576.
8. Dubinkina, V. B., Ischenko, D. S., Ulyantsev, V. I., Tyakht, A. V., & Alexeev, D. G. (2016). Assessment of k-mer spectrum applicability for metagenomic dissimilarity analysis. BMC bioinformatics, 17, 38.
9. Ekmekci, B., McAnany, C. E., & Mura, C. (2016). An Introduction to Programming for Bioscientists: A Python-Based Primer. PLoS computational biology, 12(6), e1004867.
10. Erbert, M., Rechner, S., & Müller-Hannemann, M. (2017). Gerbil: a fast and memory-efficient k-mer counter with GPU-support. Algorithms for molecular biology : AMB, 12, 9.
11. Frech, C., & Chen, N.(2010). Genome-wide comparative gene family classification. PloS one, 5(10), e13409.
12. Grell, M. N., Mouritzen, P., & Giese, H. (2003). A Blumeria graminis gene family encoding proteins with a C-terminal variable region with homologues in pathogenic fungi. Gene, 311, 181–192.
13. Hall BK, editor (1994) Homology. The Hierarchical Basis of Comparative Biology. San Diego: Academic Press pp. 339-358.
14. Huang, G. Da, Liu, X. M., Huang, T. L., & Xia, L. C. (2019). The statistical power of k-mer based aggregative statistics for alignment-free detection of horizontal gene transfer. Synthetic and Systems Biotechnology, 4(3), 150–156.
15. Kriventseva EV. (2005). Protein classification by clustering techniques. Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics. 6: Article 92
16. Lang, M. & Orgogozo, V. (2011). Identification of homologous gene sequences by PCR with degenerate primers. Methods in molecular biology (Clifton, N.J.), 772, 245-256.
17. Leva-Ulitsky, B., Diemer, K., & Thomas, P. D. (2005). On the quality of tree-based protein classification. Bioinformatics, 21(9), 1876–1890.
18. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017). KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics (Oxford, England), 33(4), 574–576.

19. Marçais, G., & Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of Kmers. Bioinformatics (Oxford, England), 27(6), 764–770.
20. Perkel J. M. (2018). Why Jupyter is data scientists' computational notebook of choice. Nature, 563(7729), 145–146.
21. Prabina Kumar Meher, Tanmaya Kumar Sahu, A.R. Rao (2016), Identification of species based on DNA barcode using k-mer feature vector and Random forest classifier, Gene, Volume 592, Issue 2, pp 316-324.
22. Tamay de Dios L, Ibarra C, Velasquillo C. (2013). Fundamentos de la reacción en cadena de la polimerasa (PCR) y de la PCR en tiempo real, Medigraphic, Vol. 2, Núm. 2, pp. 70-78.
23. Wang, Y., Fu, L., Ren, J., Yu, Z., Chen, T., & Sun, F. (2018). Identifying Group-Specific Sequences for Microbial Communities Using Long k-mer Sequence Signatures. Frontiers in microbiology, 9, 872.
24. Wu, C. H., Huang, H., Yeh, L. S., & Barker, W. C. (2003). Protein family classification and fuctional annotation. Computational biology and chemistry, 27(1), 37-47.

Dr. Jesús Eduardo Zúñiga León
Universidad Autónoma Metropolitana-Xoc

Dr. Juan Esteban Barranco Florido
Jefe del Departamento de Sistemas Biológicos
Universidad Autónoma Metropolitana-Xoc