BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

INSTITUTO DE CIENCIAS

CENTRO DE INVESTIGACIONES EN CIENCIAS MICROBIOLÓGICAS

- ANÁLISIS BIOINFORMÁTICO DE BACTERIOCINAS EN EL GÉNERO *BURKHOLDERIA* Y DESARROLLO DE UNA HERRAMIENTA DE ANÁLISIS MASIVO DE DATOS DE SECUENCIAS -

TESIS
QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS (MICROBIOLOGÍA)

PRESENTA:

MCs. YAGUL PEDRAZA PÉREZ

DIRECTOR DE TESIS:

DR. LUIS ERNESTO FUENTES RAMÍREZ

PUEBLA, PUE. AGOSTO 2018

BUAP

Puebla, Pue. a 6 de julio 2018.

A LA ACADEMIA DEL POSGRADO
EN MICROBIOLOGÍA
CICM-ICUAP
P R E S E N T E.

Por este conducto los abajo firmantes, integrantes del Comité revisor de Tesis de Doctorado del alumno **Yagul Pedraza Pérez,** les informamos que hemos revisado el escrito titulado:

"Análisis bioinformático de bacteriocinas en el género *Burkholderia* y desarrollo de una herramienta de análisis masivo de datos de secuencias"

A nuestro juicio, el alumno **Yagul Pedraza Pérez** puede proceder a la impresión de la tesis y a la presentación del examen de grado correspondiente.

Sin más que agregar, nos despedimos afectuosamente de ustedes.

A t e n t a m e n t e.
"Pensar Bien Para Vivir Mejor"

D.C. Miguel Castañeda Lucio

D.C. Jesús Muñoz Rojas

D.C. Ismael Hernández Lucas

D.C. Rosa María Gutiérrez Ríos

D.C. Vianey Marín Cevada

Posgrado en Ciencias Microbiológicas
Instituto de Ciencias (ICUAP)

Edificio IC11,
Ciudad Universitaria
Col. San Manuel, Puebla, Pue. C.P. 72570
01 (222) 229 55 00 Ext. 2522
posgrado.microbiologia@correo.buap.mx

AGRADECIMIENTOS

A Dios por absolutamente todo.

A mi familia por todo su apoyo.

A mis amigos por todas las experiencias que hemos compartido.

Al Instituto de Ciencias de la BUAP por su programa de posgrados.

A la Vicerrectoría de Investigación y Estudios de Posgrado (VIEP) por el apoyo de beca otorgado durante una etapa de la realización de este proyecto.

Al Dr. Luis Ernesto Fuentes Ramírez por su dirección y su apoyo durante la realización de este proyecto.

Al Comité Tutoral conformado por la Dra. Rosa María Gutierrez Ríos, los Dres. Ismael Hernández Lucas, Miguel Castañeda Lucio, Jesús Muñoz Rojas y Luis Ernesto Fuentes Ramírez.

Al personal que labora en el posgrado en ciencias.

CONTENIDO

Índice de figuras

Índice de cuadros

1 RESUMEN

Una de las herramientas más usadas para comparar secuencias de proteínas o de ADN con las bases de datos es BLAST. Actualmente, BLAST está disponible a través de internet en varios servidores y pemite hacer búsquedas con una o unas pocas secuencias. Sin embargo, el análisis de la creciente información genética requiere herramientas que permitan manejar una gran cantidad de información, obtenida a partir de la comparación de secuencias, de manera rápida e intuitiva. A pesar de que es posible configurar BLAST en una computadora personal para realizar múltiples búsquedas, el manejo y análisis masivo de los datos obtenidos se convierte en un proceso que consume mucho tiempo.

En este trabajo se presenta una nueva estrategia de visualización que permite desplegar, de una manera intuitiva, una gran cantidad de resultados de BLAST en genomas bacterianos completamente secuenciados. Los replicones bacterianos circulares son proyectados como líneas horizontales con longitud fija de 360, que son los grados que tiene un círculo, y un sistema de coordenadas x, y fue creado, en donde el eje-x representa la longitud del replicon y el eje-y la cantidad de replicones usados. Cuando una secuencia de búsqueda coincide con un gen/proteína de un replicón particular, el resultado de BLAST es representado como un vector con origen en la posición x, y, magnitud proporcional a su longitud real y una dirección acorde al sentido de transcripción. De esta manera es posible representar decenas de miles de resultados de BLAST de manera simultanea y analizarla desde su totalidad hasta un resultado en particular en tiempo real, usando pocos recursos computacionales. Con esta herramienta se realizó una búsqueda bioinformática de bacteriocinas en el género *Burkholderia* usando 3780 secuencias de bacteriocinas y genes relacionados como secuencias de entrada. Se encontraron 175 clusters de genes y 45 bacteriocinas de tipo lectina, entre los que se incluyen péptidos procesados por peptidasas de la familia C39, sistemas de inhibición dependientes de contacto (CDI), bacteriocinas tipo *phage*, microcinas y péptidos *lasso*.

2 ABSTRACT

One of the most used tool to compare protein or DNA sequences against databases is BLAST. To date, BLAST is available through internet on several web servers and allows to perform searches with one or several sequences. Nevertheless, the analysis of growing genetic information needs tools that allow to handle, in a rapid and intuitive way, high amount of information obtained from sequence comparison. In spite of it is possible to configure BLAST in a personal computer to perform multiple searches, handling and analysis of massive data obtained becomes a time-consuming process.

In this work, we introduce a new visualization strategy that allows to display a great amount of BLAST-results in whole-sequenced bacterial genomes. The circular bacterial replicons are projected as horizontal lines with a fixed length of 360 that are the degrees in a circle, and an x, y coordinate system was created, where the x-axis represents the length of the replicon number and the y-axis the quantity of replicons used as database. When a query sequence matches with a particular gen/protein from a replicon, the BLAST-result is represented as a vector with origin at x, y position, magnitude proportional to their real length and direction according to the transcription sense. In this way, it is possible to represent several dozen of thousands of BLAST-results simultaneously and analyze them from the whole data and focusing on a particular result in real time, using low computational resources. With this tool we performed a bioinformatic search of bacteriocins in the *Burkholderia* genus using 3780 query sequences of bacteriocins and related genes. We found 175 gene clusters and 45 lectin-like bacteriocins, including peptides processed by C39 peptisases, contact-dependent inhibition systems, microcins and *lasso* peptides.

3 INTRODUCCIÓN

3.1. Bacteriocinas.

Las bacteriocinas son sustancias de naturaleza proteica que tienen la capacidad de inhibir el crecimiento de microorganismos, generalmente aquellos que están más relacionados filogenéticamente con la especie productora, aunque existen excepciones de bacteriocinas con amplio espectro de actividad. Los primeros reportes de estas sustancias datan de los años veinte cuando se estudiaron cepas de *Escherichia coli* en las que se presentaba un efecto antagonista, revelando un nuevo y complejo mecanismo de interacción bacteriana (1).

Estas proteínas se han encontrado en prácticamente todos los microorganismos en que se han buscado, y en algunas especies bacterianas se han descubierto más de una bacteriocina. A pesar la creciente información acumulada sobre los genes que las producen, sus mecanismos de regulación, modificaciones postraduccionales, estrategias de liberación, mecanismos de acción y de autoinmunidad, su papel dentro de las complejas comunidades microbianas aún plantea preguntas sobre los roles específicos que desempeñan estas sustancias en las interacciones ecológicas de los microorganismos (2). Las bacteriocinas, por un lado, permiten a ciertas cepas productoras colonizar nuevos ambientes en donde ya existen comunidades bacterianas establecidas. Al mismo tiempo, participan en la defensa de sitios al impedir el establecimiento de microorganismos ajenos. También son relevantes en la lucha por la supervivencia en ambientes con escasez de nutrientes y se han visto involucradas en la activación de mecanismos que dependen de *quorum sensing*. La producción de péptidos antimicrobianos no se limita a bacterias, en la década de los sesentas se comenzaron a descubrir este tipo de sustancias en protozoarios, insectos, anfibios, aves, mamíferos y plantas y posteriormente también se detectaron en algunos organismos extremófilos del dominio *Archaea* (2; 3).

Debido a la capacidad que tienen para inhibir el crecimiento de microorganismos, las bacteriocinas han captado la atención por su potencial aplicación como antibióticos de nueva generación, preservadores de alimentos, probióticos, control biológico e inclusive por su potencial uso como tratamiento contra el cáncer (4; 5; 6). El mayor éxito de la aplicación de bacteriocinas se encuentra en la preservación de alimentos, en especial de las bacteriocinas producidas por bacterias Gram-positivas(7; 8). Esto se debe a la premisa de que las bacterias ácido lácticas, que son Gram-positivas, se han usado en la fermentación de alimentos por más de cuatro milenios y no han causado daño al ser humano, por lo tanto se consideran seguras y las sustancias producidas por ellas también.

La búsqueda de nuevas bacteriocinas de manera experimental representa un gran esfuerzo debido al enorme espacio de estudio que implica la cantidad de potenciales bacteriocinas, multiplicado por el número de cepas productoras, multiplicado por el número de potenciales condiciones de expresión de estas sustancias. Por otro lado, la creciente disponibilidad de genomas bacterianos hace posible una nueva estrategia de búsqueda de bacteriocinas *in silico*. Para ello es importante tomar en cuenta el contexto genómico en el que se encuentran, pues esto facilita su identificación y caracterización bioinformática. Eso debido a que los genes que participan en la regulación de su expresión, modificaciones postraduccionales, procesamiento de péptido señal, transporte y/o mecanismos de autoinmunidad generalmente son más conservados que los genes que codifican para la bacteriocina (3).

Las bacteriocinas son extremadamente diversas, lo que las hace difícil de caracterizar y de clasificar. Actualmente no existe una clasificación que sea universalmente aceptada ya que su agrupación con base en alguna de sus características no se refleja satisfactoriamente con su agrupación basada en otra característica. Sin embargo, se pueden reconocer cuatro grandes grupos de estos péptidos antimicrobianos, que a su vez se pueden agrupar en diversas subcategorías dependiendo de sus características más importantes. Estos cuatro grupos principales son las bacteriocinas producidas por las bacterias Gram-positivas (9), las de Gram-negativas (10), las de *Archea* (11) y el cuarto grupo, conocido como eukariocinas (12), producidos por organismos eucariotas.

3.2. Bacteriocinas en bacterias Gram-Positivas.

Se ha hipotetizado que todas las bacterias producen al menos una bacteriocina y que la razón por la cual no se han detectado todas es por que no se han encontrado las condiciones adecuadas de cultivo que favorezcan su expresión. Aunque esto no sea necesariamente

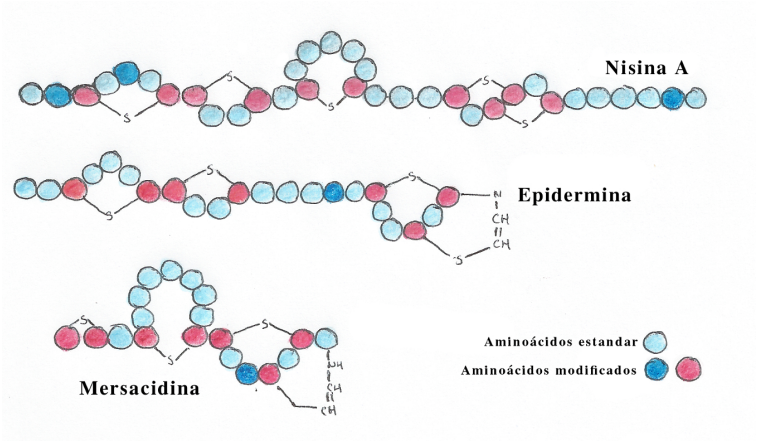

Figura 3.1: Esquema de algunos Lantibioticos representativos. Los residuos involucrados en la formación de las estructuras de (*beta*-Metil) Lantoninas se muestran en rojo. Otros residuos modificados se muestran en azul intenso.

cierto, sí resulta aparente que las bacteriocinas se encuentran ampliamente distribuidas en la naturaleza y que su presencia juega un papel importante en las interacciones microbianas.

Entre las bacteriocinas más conocidas en bacterias Gram-positivas están las de bacterias ácido lácticas (13). La aplicación potencial de éstas ha sido aceptada más fácilmente debido a que estos microorganismos han estado presentes en alimentos de consumo humano desde hace milenios sin causarnos daño y por lo tanto se considera que sus productos son inócuos. Las bacteriocinas de Gram-positivas se clasifican en cuatro grupos: 1) lantibióticos; 2) péptidos resistentes al calor; 3) bacteriocinas grandes, sensibles al calor; y se ha propuesto un cuarto grupo con proteínas complejas con lípidos o carbohidratos que aún no ha sido ampliamente aceptado (3; 9).

3.2.1. Bacteriocinas del grupo I, Lantibióticos.

Estas bacteriocinas son de tamaño menor a 5 kDa, se caracterizan por tener modificaciones postraduccionales drásticas de algunos de sus aminoácidos. A su vez son subdivididos en tres tipos: Lantibióticos, Laberintopeptinas y Sactibioticos (3).

El nombre de **Lantibióticos** deriva de *lantionina* y *antibiótico*, y se refiere a péptidos que sufren modificaciones postraduccionales generando una enorme cantidad de los aminoácidos lantionina (Lan), metil-lantonina, dehidro-alanina o ácido dehidro-aminobutírico (Figura 3.1). Ejemplos de este tipo de bacteriocinas incluyen a la Nisina, la Subtilina, Pep5 y Epidermina cuyo mecanismo de acción se da por la formación de poros en la membrana de las células sensibles, lo que conlleva a la disipación del potencial de la membrana y al eflujo de

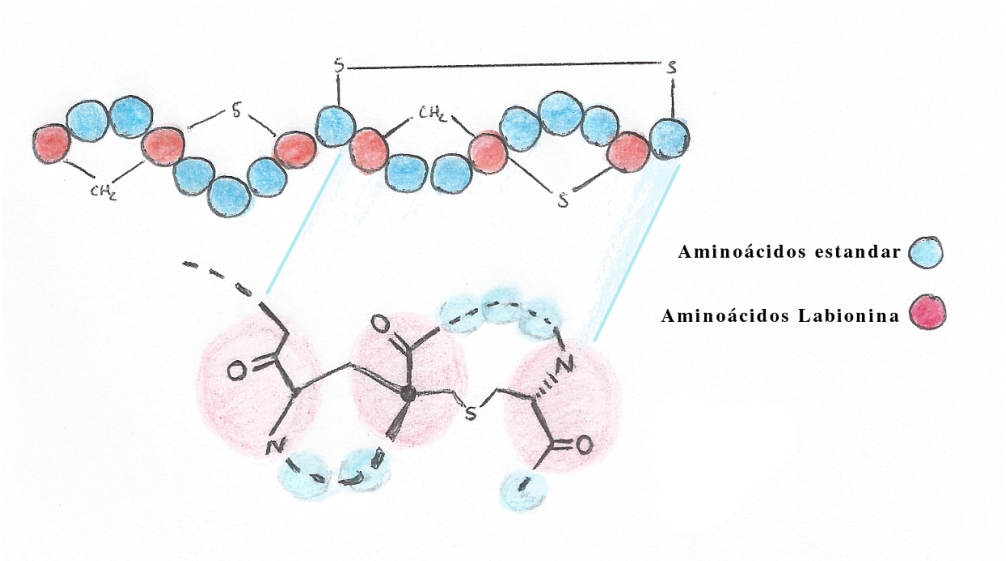

Figura 3.2: Estructura de Laberintopeptinas. A) Esquema de las laberintopeptinas que muestra la ubicación de los aminoácidos modificados postraduccionalmente (Lab). B) Comparación entre la Lantionina y la Labionina. La presencia de la labionina genera una estructura con dos anillos en forma de "8"que comparten un carbono *alpha*.

pequeñas moléculas. Otros ejemplos de este tipo incluyen a la Mercsacidina, la Cinnamicina, la Mutacina II y la Lacticina 481, que son lantibioticos globulares que tienen por mecanismo de acción la inhibición enzimática de la biosíntesis de la pared celular.

Algunos autores proponen subsiguientes agrupaciones con base a ciertas características. Se propone una subclase I para aquellos péptidos que son lineales y cuyos precursores son modificados por dos enzimas llamadas LanB, que deshidrata residuos de treonina y serina, y LanC, encargada del proceso de ciclización de residuos. En la subclase II se ubican a los lantibioticos que son de tipo globular y cuya modificación postraduccional de deshidratación y ciclización es llevada a cabo por una sola enzima llamada LanM. Una tercer subclase agrupa aquellos péptidos encontrados en los genomas cerca de enzimas de modificación, sin embargo, no se ha detectado capacidad antimicrobiana. La subclase IV incluye péptidos que son procesados por una enzima descrita recientemente llamada LanL que genera aminoácidos deshidratados por un mecanismo que implica una fosforilación como compuesto intermediario. Estas últimas dos subclases han originado el término de *Lantipéptido* que se refiere a compuestos que por estructura y mecanismos de síntesis están claramente relacionados con lantibióticos pero que aún no se les ha detectado actividad antagónica (3).

Las **Laberintopeptinas** fueron descubiertas en *Actinomadura namibiensis* cepa DSM613 y se han denominado de esta forma debido a que tienen una estructura comple-

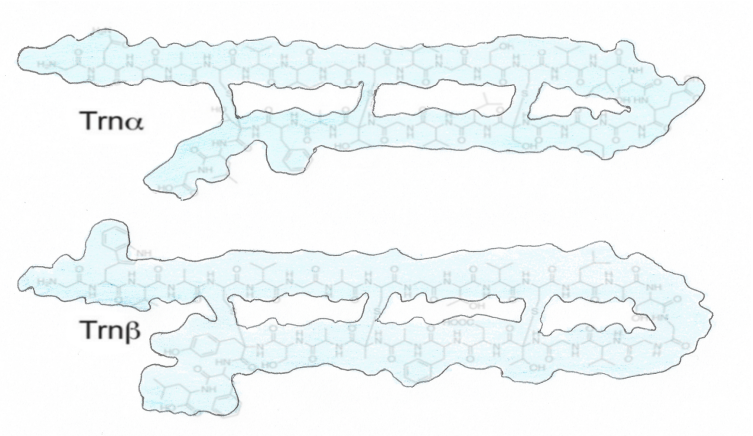

Figura 3.3: Conectividad estructural en la Turicina CD. La modificación postraduccional que sufre esta bacteriocina implica la unión de un azufre con el carbono *alpha* de la cadena principal del péptido.

ja como "laberinto". Contienen un inusual aminoácido modificado llamado labionina que tiene características similares a la lantonina (Figura 3.2). Cuando la labionina está presente se forman dos anillos (uno con 4 aminoácidos y otro con 5) en forma de "8"que comparten un carbono *alpha* en la unión entre ellos, confiriéndole al péptido gran estabilidad (14).

El cluster de genes que codifica para las laberintoninas en *A. namibiensis* consta de 5 genes *labKC, labA1/A3, labA2, labT1 y labT2*. Los genes *labA1/A3, labA2,* codifican para los precursores de las bactericocinas, comprenden un péptido señal de unos 20 aminoácidos y un propéptido de tamaño similar que será modificado postraduccionalmente para formar la laberintonina. El gen *labKC* codifica para una proteína bifuncional que tiene en su extremo N-terminal a una Ser/Thr cinasa y en su extremo C-terminal una ciclasa. Los genes *labT1* y *labT2* codifican para transportadores de tipo ABC implicados en la exportación de las laberintoninas (14).

Los **Sactibioticos** reciben este nombre porque contienen un enlace inusual entre un azufre y un carbono *alpha* (***s***ulphur to ***a***lpha ***c***arbon linkage-containing anti***biotics***). La Turicina CD es un ejemplo de este tipo de bactriocinas, está compuesto por dos péptidos llamados *alpha* y *beta* (Trnα y Trnβ), es producida por *Bacillus thuringensis* cepa 6431 y es activa contra *Clostridium*. Tienen enlaces covalentes entre el azufre de las cisteínas y el carbono *alpha* de residuos de serina, theonina, alanina o tirosina (Figura 3.3).

El cluster de genes que codifica para esta bacteriocina esta compuesto por los genes *trnFG*$\beta\alpha$*CDE* en donde TrnF y TrnG funcionan como transportadores del tipo ABC, Trnβ y Trnα son los propéptidos que, a diferencia de otros péptidos con firma GG, la ruptura se

da entre las dos glicinas y no después de la firma (15). TrnC y TrnD pertenecen a una familia de proteínas llamada Radical SAM que pueden romper la S-Adenosil-Metionina para formar un radical de 5'-desoxiadenosil que es necesario para llevar a cabo su catálisis (15; 16). TrnE es una peptidasa que probablemente desempeña un papel citoplasmático pues no presenta péptido señal para exportación (15).

3.2.2. Bacteriocinas del grupo II, péptidos resistentes al calor.

Esta clase de bacteriocinas está compuesta por un grupo heterogéneo de péptidos con tamaño menor a 10 kDa y con un contenido estándar de aminoácidos, cuyas únicas modificaciones se limitan a puentes disulfuro o ciclización entre el N- y el C-terminal. La subclase IIa incluye péptidos de tamaño entre 27 y 55 aminoácidos. La única modificación que sufren es la remoción del péptido líder que necesitan para su exportación, que frecuentemente incluye una firma de doble glicina. La región C-Terminal es hidrofóbica y/o anfipática, es menos conservada a nivel de secuencia y es la responsable de la actividad bactericida, incorporándose a la membrana formando un poro. La subclase IIb está formada por bacteriocinas con más de 50 aminoácidos y se caracterizan porque su actividad citotóxica está dada por un dímero, con poca o nula actividad como monómeros, que genera un poro en la membrana. En la subclase IIc se ubican las bacteriocinas que se encuentran circularizadas covalentemente entre su extremo N- y su C-terminal. Finalmente, la subclase IId incluye aquellos péptidos que no entran en ninguna de estas categorías por ser lineales y monoméricos pero sin similitud con las bacteriocinas de clase IIa (3).

3.2.3. Bacteriocinas del grupo III, bacteriocinas grandes sensibles al calor.

Denominadas también como Bacteriolisinas, son bacteriocinas grandes y sensibles al calor. Están constituidas por varios dominios con funciones particulares de reconocimiento de la célula sensible, translocación y actividad citotóxica. Ejemplos de bacteriocinas de este grupo son la Helveticina J producida por *Lactobacillus helveticus*, Zoocina A producida por *Streptococcus zooepidermicus*, Enterolisina A producida por *Enterococcus faecalis* y Millericina B por *Streptococcus milleri* (3). Cabe mencionar que en este grupo se suelen incluir proteínas líticas no relacionas con bacteriocinas que han sido más caracterizadas.

Los esquemas de clasificación de bacteriocinas han cambiado en los últimos 25 años, consecuencia de la identificación de nuevas bacteriocinas con características diferentes. La disponibilidad de secuencias y estructuras tridimensionales ha permitido tener una clasificación más refinada. Sin embargo, se preveen cambios en esta clasificación en un futuro, ya que

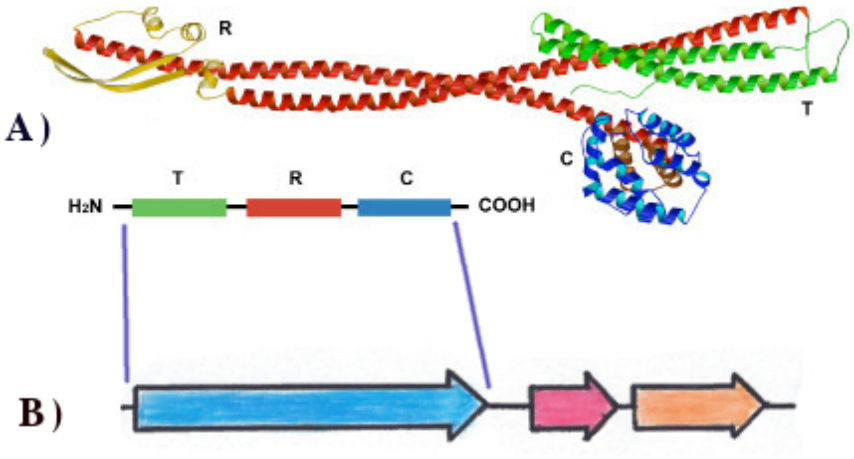

Figura 3.4: Estructura de Colicinas. A) La estructura conservada de las colicinas consta de 3 dominios, el dominio T necesario para la translocación al interior de la célula, el dominio R de reconocimiento y el dominio C con actividad citotóxica para la bacteria blanco. **B)** Organización genética de las bacteriocinas de tipo Colicina.

una notable característica que está emergiendo es la identificación de péptidos parecidos a lantibióticos en los que no se ha detectado actividad antimicrobiana.

3.3. Bacteriocinas en bacterias Gram-Negativas.

3.3.1. Colicinas.

Las bacteriocinas más ampliamente estudiadas son las producidas por *Escherichia coli*, llamadas **Colicinas**, y que predominantemente se encuentran en la familia *Enterobacteriaceae*. Su principal característica es que contienen tres dominios: el dominio N-terminal que participa en la translocación de la bacteriocina al interior de la bacteria sensible, un dominio central involucrado en el reconocimiento de la bacteria blanco y el dominio C-terminal que tiene la actividad citotóxica (Figura 3.4A). Su peso molecular oscila entre los 40 y los 80 kDa, generalmente son codificadas en plásmidos y su expresión es inducida por el sistema SOS. Algunas de estas colicinas pueden formar poros en la membrana interna de la célula sensible modificando su potencial de membrana y permitiendo la salida de iones y moléculas pequeñas, otras colicinas tienen actividad enzimática de tipo nucleasa o pueden detener la síntesis de proteínas (17).

Las colicinas están codificadas en plásmidos que se llaman colicinogénicos o pCol. Las cepas productoras, también llamadas colicinogénicas, son abundantes en intestinos de animales. En la organización genética de las colicinas, que pueden tener de uno a tres genes (Figura 3.4B), el primer gen del operon codifica para la colicina, se denomina *cxa* por ***c****olicin X* ***a****ctivity*, el siguiente gen es llamado *cxi* por *colicin X immunity*, el último gen es denominado *cxl* por ***c****olicin X* ***l****ysis*. La producción de la proteína codificada por este gen permite que la colicina pueda ser liberada al medio debido a la ruptura y muerte de la célula productora (17).

El operón de colicinas contiene dos sitios de unión a la proteína LexA, cada sitio une a un dímero de LexA, la unión de estos dímeros provoca una curvatura en el DNA que bloquea la transcripción. Cuando existe un daño en el DNA ocasionado por algún agente mutagénico como la radiación UV, químicos como la mitomicina C o condiciones de estrés, se activa la proteína RecA que induce la ruptura del represor LexA, permitiendo la transcripción de los genes del operón para colicina (17; 18). El espectro reducido de su capacidad antagónica se debe principalmente a que utilizan un receptor específico ubicado en la superficie de la célula sensible para llevar a cabo su actividad citotóxica, lo que permite su clasificación en dos grandes grupos, A y B.

Colicinas del grupo A.

Están codificadas en plásmidos multicopia pequeños de alrededor de 10 kb, se caracterizan por utilizar el transportador de cobalamina (Vitamina B12) BtuB como receptor de reconocimiento de la célula sensible y el mecanismo Tol para translocación a través de la membrana externa. BtuB es una proteína de membrana externa en forma de barril *beta* usado por las bacterias para transportar cobalamina del medio hacia el espacio periplásmico. El transporte de esta vitamina es asistido por el mecanismo compuesto por el sistema Ton. Una vez en el espacio periplásmico, la cobalamina es dirigida por BtuF hacia el transportador BtuCD que la translada al interior de la célula usando energía proporcionada por ATP (Figura 3.5). El sistema translocador Ton también participa en el transporte de sideróforos usando otras proteínas de membrana externa como FepA, FecA o FhuA (3).

Una vez que la colicina se ha unido al receptor BtuB a través de su dominio central de reconocimiento, utiliza a la proteína de membrana externa OmpF para llevar a cabo su translocación al espacio periplásmico asistido por el sistema Tol, durante este proceso la colicina se disocia de su proteína de inmunidad (3). OmpF es una proteína de membrana externa que facilita el transporte pasivo de moléculas hidrofóbicas pequeñas. El sistema Tol está constituido por cinco proteínas. TolA es una proteína de membrana interna con un solo

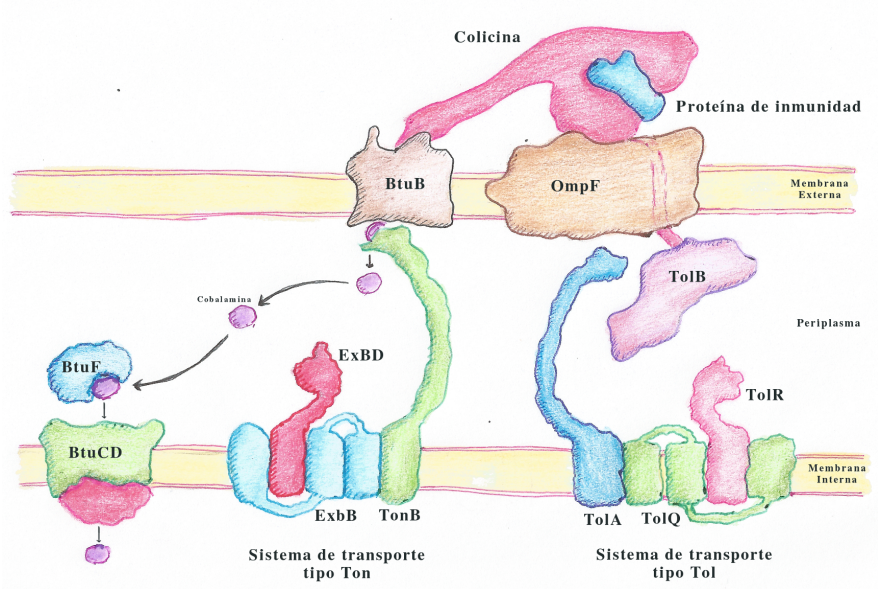

Figura 3.5: Sistemas trasportadores Tol y Ton. Este transportador es usado como proteína de reconocimiento por las colicinas del grupo A.

dominio transmembranal, está ubicada en el espacio periplásmico y se extiende con un dominio alargado en forma de tallo y termina con un dominio globular (Figura 3.5). TolQ y TolR se encuentran asociados a TolA y proveen energía por medio de una fuerza protón motriz, de manera similar al sistema ExbB-ExbD del sistema Ton. TolB es el encargado de asistir el transporte a través de OmpF en complejo con el dominio globular de TolA, por último Pal es una proteína de membrana externa ubicada en el lado periplásmico que interactúa con TolB induciendo un cambio conformacional que le permite unirse a TolA (19).

El grupo A comprende a las colicinas A, E1 a E9, K, N, U, S4, y Y. Dentro de este grupo existen dos mecanismos de acción citotóxica, las colicinas que matan a las células sensibles mediante la formación de poros en su membrana interna (Colicinas A, E1, K, N, U, S4 y Y) y aquellas que tienen actividad de nucleasas (E2, E7, E8, E9:DNasa, E3, E4, E5, E6:16S-RNAsa, E6:tRNAsa). Una vez que las colicinas formadoras de poro han sido translocadas al periplasma celular se dirigen hacia la membrana interna siguiendo un mecanismo Browniano en el que participan interacciones cada vez mas afines con diferentes componentes del sistema TolBAQR y, al entrar en contacto con la membrana interna, se insertan formando poros.

La proteína de inmunidad de estas colicinas es un péptido pequeño de alrededor de 11 a 18 kDa que es expresado constitutivamente a bajos niveles por las células productoras de colicina. Para proteger a la célula, esta proteína de inmunidad se localiza en la membrana interna del lado periplásmico e interactúa con la colicina evitando que se inserte en la mem-

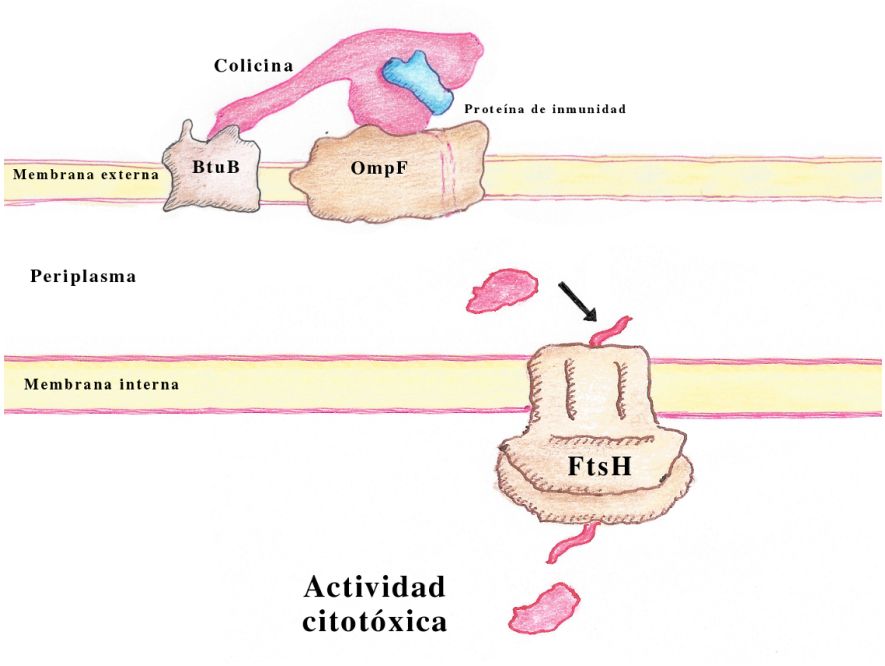

Figura 3.6: La proteína FtsH. Esta proteína se encuentra inmersa en la membrana interna y se encarga de transportar a la bacteriocina al citoplasma.

brana. Sin embargo, sólo puede proteger de la colicina expresada por el mismo operón y no protege de otras colicinas formadoras de poro (17).

Las colicinas que tienen actividad de nucleasas deben atravesar la membrana interna antes de llevar a cabo su actividad citotóxica. Este paso puede ocurrir de dos maneras, algunas colicinas utilizan un mecanismo de "auto-transporte" formando un canal en la membrana interna que no mata a la célula por cambios de voltaje en la membrana pero sí permite el tránsito del dominio citotóxico a través de la membrana interna. Otro mecanismo de translocación usado por algunas colicinas es asistido por una proteasa de membrana llamada FtsH (Figura 3.6) que separa el dominio citotóxico del resto de la colicina durante su translocación al interior de la célula (20). Una vez dentro, las DNAsas cortan indiscriminadamente el DNA de la bacteria, sin embargo, las 16S-RNAsas cortan de manera especifica el RNA de la subunidad menor del ribosoma en la posición 1493 ubicada cerca sitio de interacción del mRNA con el tRNA y que juega un papel importante en la traducción de la información genética. Por otro lado, las colicinas tRNAsas cortan a tRNA específicos en el sitio del anticodón. Las proteínas de inmunidad de las colicinas nucleasas son de alrededor de 10 KDa, se unen cerca del sitio activo bloqueando una porción de la zona de reconocimiento del sustrato impidiendo su catálisis (17).

Colicinas del grupo B.

Estas colicinas se encuentran codificadas en plásmidos monocopia de mayor tamaño, alrededor de 40 kb, que pueden codificar para más de una colicina. Se caracterizan por usar las proteínas de membrana externa FepA, FhuA o Cir para el reconocimiento de la célula blanco y el sistema tipo Ton de transporte para su translocación al periplasma celular (Figura 3.6). El sistema Ton es estructural y funcionalmente similar al sistema Tol; está constituido por TonB, una proteína de membrana interna ubicada en el espacio periplásmico que se puede unir a una proteína canal de membrana externa como BtuB, y por ExbB-ExbD que forman un complejo proteico, ubicado en la membrana interna, que provee de energía a TonB usando un gradiente de protones (Figura 3.5). Las colicinas descritas de este grupo también pueden tener tres tipos de mecanismos de citotoxicidad, las formadoras de poro (colicinas B, Ia, Ib, H, 5 y 10), al menos una colicina nucleasa (Colicina D) y la colicna M cuyo mecanismo de toxicidad es mediante la degradación de peptidoglucanos de la pared bacteriana (3; 21; 22). Los mecanismos de acción de estas colicinas, ya sea las formadoras de poro o las nucleasas son similares a los de las colicinas del grupo A (17).

3.3.2. Microcinas.

Las microcinas son bacteriocinas de bajo peso molecular desde 1 kDa hasta alrededor de 10 kDa, generalmente son resistentes a la ruptura por proteasas, rangos amplios de pH y temperaturas extremas. Son producidas principalmente por enterobacterias durante la fase estacionaria del crecimiento bacteriano en la que se presenta una baja cantidad de nutrientes. Pueden estar codificadas en plásmidos o en cromosoma y su mecanismo de liberación está mediado por transportadores dedicados a las microcinas y no implica la lisis de la célula productora como en el caso de las colicinas. Esto les permite producir la toxina continuamente sin perder viabilidad. Se aceptan dos grandes grupos de microcinas de acuerdo a su modificación postraduccional, su organización genética y la secuencia de su péptido señal (3).

Microcinas de Clase I.

Estas bacteriocinas son péptidos menores a 5 kDa, presentan modificaciones postraduccionales drásticas y su mecanismo de acción tiene que ver con la inhibición de enzimas que son vitales para la célula. Esta clase de microcinas tienen tres representantes que no están relacionados estructuralmente: la Microcina J25, Microcina B17 y Microcina C7 (Figura 3.7).

Microcina J25. La Microcina J25 (MccJ25) es altamente resistente a ruptura enzimática y a condiciones de autoclave (120 grados centígrados y 20 libras de presión por

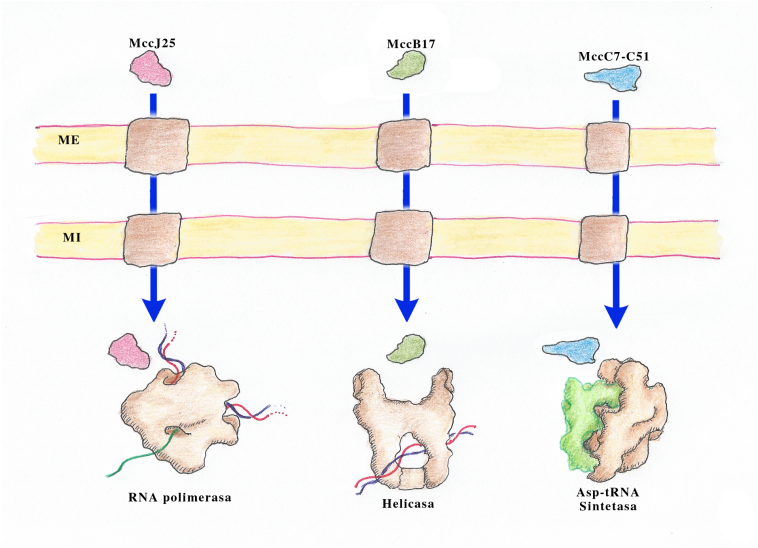

Figura 3.7: Actividad citotóxica de las Microcinas Clase I. Estas microcinas pueden entrar a la célula usando sistemas de transporte insertados en la membrana y, una vez en el citoplasma pueden tener diferentes efectos citotóxicos en la célula blanco.

20 minutos) gracias a su forma ciclizada en forma de nudo. La forma activa consta de 21 aminoácidos que proceden de un péptido precursor de 58 aminoácidos codificado por el gen *mcjA*. La producción de la MccJ25 depende de un cluster de los genes *mcjABCD*, el gen *mcjB* codifica para una proteasa implicada en la ruptura del péptido lider, *mcjC* codifica para una enzima similar a la asparagina-sintetasa que convierte al ácido aspártico en asparagina y es el responsable del proceso de ciclización propio de esta bacteriocina. Por último, *mjcD* codifica para un transportador de membrana de tipo ABC que está implicado en la exportación de la microcina al exterior de la célula y por tanto funciona también como proteína de inmunidad. Es probable que McjB y McjC se encuentren formando un complejo binario que a su vez se asocia con McjD por su extremo citoplasmático (23), optimizando así la producción de la microcina e impidiendo su acumulación dentro de la célula (Figura 3.8).

La estructura de esta microcina, conocida como *péptido lasso*, consta de un anillo formado por la Gly1 y el Glu8 en el cual se inserta el extremo C-terminal del péptido formando un nudo. Los voluminosos aminoácidos aromáticos Phe19 y Tyr20 impiden que el extremo C-terminal salga del nudo, dando como resultado una estructura bastante estable protegida contra lisis enzimática y altas temperaturas.

La incorporación de la MccJ25 por parte de la célula sensible es por medio de la proteína FhuA, una proteína de membrana externa implicada en la absorción de hierro, y del

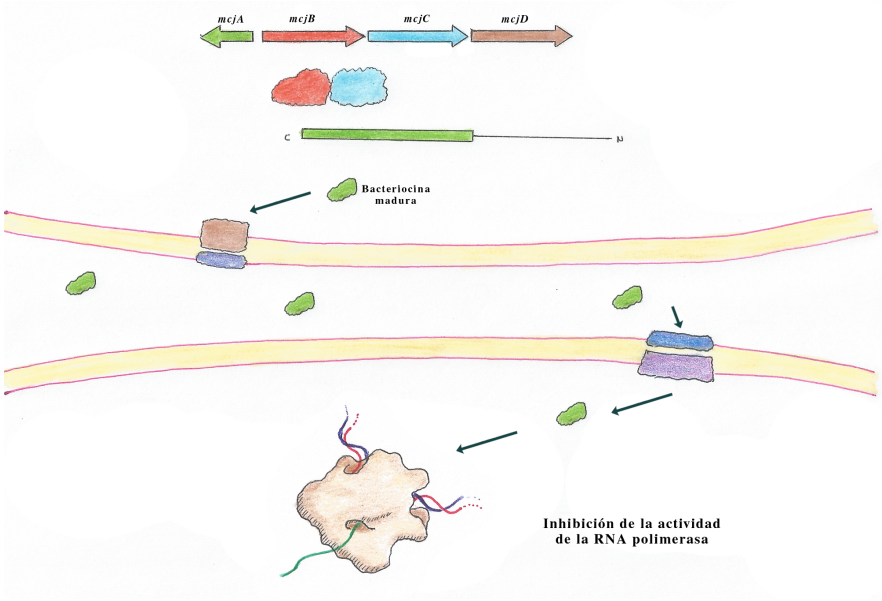

Figura 3.8: Microcina MccJ25, síntesis y mecanismos de acción. La Microcina MccJ25 es sintetizada por 4 genes. Su mecanismo de acción implica la inhibición de la RNA polimerasa de la célula sensible.

sistema de importación TonB. La MccJ25 puede inhibir la actividad de la RNA polimerasa a concentraciones micromolares, uniéndose al canal de incorporación de nucleótidos (3). Otro *péptido lasso* semejante a la MccJ25 es la Capistruina encontrada en *Burkholderia thailandensis* E264 (24).

Se ha observado la presencia de genes homólogos al cluster de *mcjABCD* en diferentes organismos y, a pesar de que no se han hecho estudios detallados sobre ellos, este tipo de bacteriocinas tiene aplicaciones biotecnológicas prometedoras (3).

Microcina B17. La Microcina B17 (MccB17) inhibe el proceso de replicación del ADN uniéndose a la subunidad *beta* de la girasa en la célula sensible. La MccB17 contiene varios anillos de oxazol y tiazol producto de las modificaciones postraduccionales que sufre. El *cluster* de genes encargado de la producción de la MccB17 se localiza en plásmido y está constituido por los genes *mcbABCDEFG*, de los cuales *mcbA* codifica para el péptido precursor de 69 aminoácidos de largo (Figura 3.9). Los genes *mcbBCD* codifican para tres proteínas que forman un complejo enzimático que se encarga de las modificaciones postraduccionales en tres pasos sucesivos; una ciclización, una deshidratación, y una deshidrogenación de los dipéptidos Gly-Ser para formar los anillos de oxazol y de los dipéptidos Gly-Cys para formar los anillos de tiazol (25).

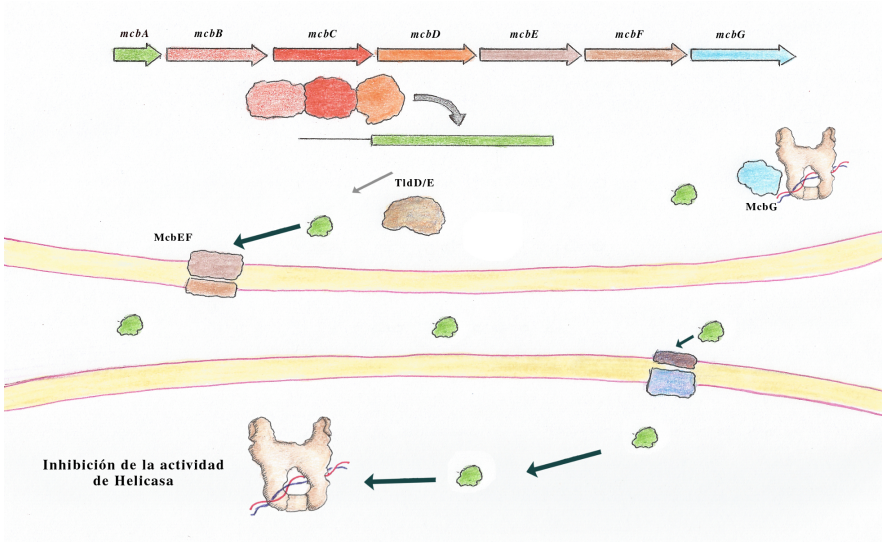

Figura 3.9: Microcina MccB17, síntesis y mecanismos de acción. La síntesis de la Microcina MccB17 requiere la participación de 7 genes. Es capaz de inhibir a la helicasa de la célula sensible.

A este péptido con anillos de tiazol y oxazol le es removido un péptido líder del amino terminal de 26 aminoácidos por alguna de dos proteasas codificadas en cromosoma TldD o TldE (26). Los genes *mcbEF* codifican para proteínas de membrana tipo ABC dedicadas a exportar la Microcina y finalmente el gen *mcbG* codifica para una proteína que confiere inmunidad. Ejemplos de bacteriocinas similares a la MccB17 son la Trifolitoxina producida por *Rhizobium leguminosarum bv. Trifolii* y la Streptolisina S producida por *Streptococcus pyogenes*.

Microcina C7-C51. Esta microcina es un heptapéptido que contiene en su extremo C-terminal a una adenosina monofosfato modificada. El cluster implicado en su producción consta de los genes *mccABCDEF*, de los cuales *mccA* codifica para el precursor de la microcina y es uno de los genes más pequeños conocidos, de tan solo 21 pares de bases. El primer aminoácido de la MccC7-C51 es una metionina formilada y el último es un ácido aspártico al que se une la adenosina monofosfato modificada. Los genes *mccBDE* están implicados en la maduración de la microcina mientras que *mccC* codifica para un transportador y *mccF* para una proteína que le confiere inmunidad a la célula productora (27).

Una vez sintetizada, la MccC7-C51 entra en la célula sensible a través de la membrana externa usando a la proteína OmpF y a los transportadores ABC YejABEF para su paso a través de la membrana interna (Figura 3.10). Dentro de la célula sensible, el grupo formil del amino terminal de la MccC7-C51 es removido y la microcina es procesada por una de tres posibles enzimas. La Peptidasa A, B o N corta entre los residuos Ala6 y Asp7 para separarla

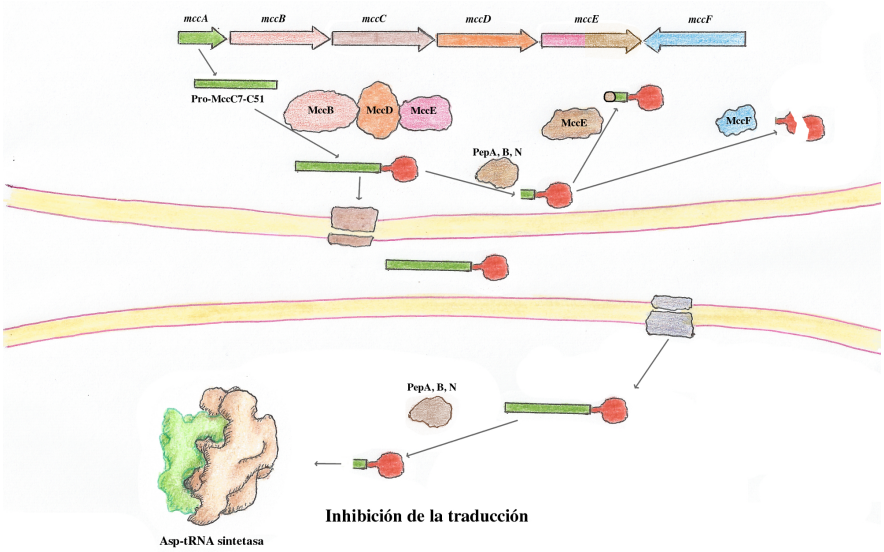

Figura 3.10: Microcina C7-C51, síntesis y mecanismos de acción. La Microcina MccJ25 es sintetizada por 6 genes y su mecanismo de acción comprende la inhibición de la Asp-tRNA Sintetasa durante la transcripción.

en un hexapéptido y en un análogo no hidrolizable del aspartil-adenilato, que es el sustrato de la Aspartil-tRNA sintetasa necesario para "cargar" al tRNA-Asp usado en la síntesis de proteínas (28; 29).

Microcinas de Clase II.

Este tipo de microcinas tienen un peso molecular entre 5 y 10 kDa, no sufren modificaciones postraduccionales grandes y son exportadas a través de la membrana interna por transportadores ABC. Se subdividen a su vez en dos categorías: Microcinas clase IIa y Clase IIb. Estas bacteriocinas causan despolarización de la membrana interna de las células sensibles.

Microcinas Clase IIa. Estas microcinas no sufren modificaciones postraduccionales, excepto por la formación de puentes disulfuro, si están presentes. Están codificadas en plásmidos grandes de bajo número de copias y su producción depende de cuatro genes, un precursor, una proteína de inmunidad y dos proteínas encargadas de exportar a la bacteriocina (3). Existen tres microcinas representativas de este grupo, la microcina Mcc24 (también llamada Colicina 24) es un péptido linear no modificado. Es secretada por *E. coli* uropatogénica y está codificada por un *loci* ubicado en un plásmido conjugativo de 43.5 kb. La microcina MccV y la MccL contienen puentes disulfuro y están codificados en dos unidades

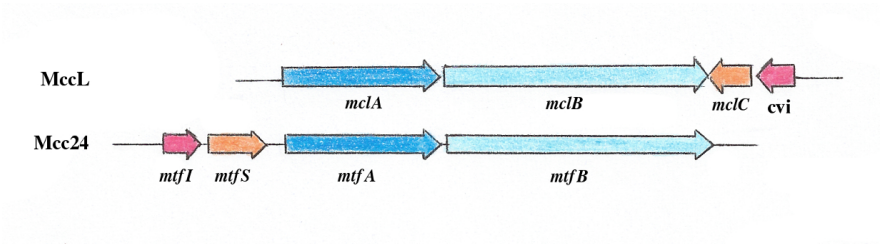

Figura 3.11: Microcinas de Clase II. Estas microcinas están codificadas por un *loci* de cuatro genes que se encuentra en plásmidos.

de transcripción convergentes (Figura 3.11).

El *loci* que codifica para la microcina Mcc24 contiene cuatro genes *mtfI, mtfS, mtfA* y *mtfB*. El primero codifica para la proteína de inmunidad, el segundo para el precursor de la bacteriocina y los otros dos para transportadores tipo ABC. La nomenclatura de los genes no se mantiene para las microcinas MccV (*loci cvi, cvaC, cvaA y cvaB*) y MccL (*loci cvi, mclC, mclA y mclB*), sin embargo la función de los genes es equivalente. El precursor de la bacteriocina es de alrededor de 103 aminoácidos, pero el péptido maduro tiene alrededor de 88 aminoácidos y en el caso de la Mcc24 es de 73 aminoácidos (3). Las microcinas usan un transportador del tipo ABC para atravesar la membrana interna que está compuesto por un homodímero de MtfB (CvaB o MclB) que tiene tres dominios: un dominio N-terminal localizado en el extremo citoplasmático que tiene una actividad de proteasa que probablemente está involucrado en la remoción del péptido líder, un dominio central poco conservado y un dominio C-terminal que contiene un sitio de unión a ATP. El segundo componente de la maquinaria transportadora (MtfA/CvaA/MclA) es una proteína accesoria localizada en el peripasma y ayuda a conectar al transportador ABC con la proteína de membrana externa TolC para ser exportada de la célula. El gen que codifica para esta proteína TolC se halla en el cromosoma y sirve para exportar otras sustancias además de la microcina (27).

Los mecanismos de acción de estas microcinas no están del todo claros, sin embargo, se sabe que la entrada de estas bacteriocinas a la célula sensible depende de las proteínas de membrana externa FepA, Cir, Fiu cuya función original es la de incorporar sideróforos a la célula, o SdaC involucrada en la incorporación de serina. Usan el sistema TonB para su translocación (3).

Microcinas Clase IIb. Estas microcinas son polipéptidos lineales que contienen, como modificación postraduccional, a un sideróforo en su extremo C-terminal que además contiene una secuencia altamente conservada de 10 aminoácidos. Bacteriocinas representan-

tes de este grupo son la MccE492, MccH74, MccM, MccG492 y MccI47. Están codificadas en cromosoma bajo una compleja estructura de genes ya que, además de los cuatro genes necesarios para las microcinas de clase IIa, se encuentran los genes que codifican para las enzimas necesarias para las modificaciones postraduccionales.

La microcina mejor caracterizada de esta clase es la MccE492 producida por *Klebsiella pneumoniae* RYC492. Su síntesis requiere de la participación de 10 genes (*mceABCDEFGHIJ*) que se encuentran organizados en al menos cinco unidades transcripcionales: el gene *mceA* codifica para el peptido precursor de 99 aminoácidos de largo y el *mceB* para la proteína de inmunidad; Las modificaciones son llevadas a cabo por los productos de los genes *mceC, mceD* y *mceI* que codifican para una glucosiltransferasa, una esterasa y una aciltransferasa respectivamente, también participa el gen *mceJ* que codifica para una proteína que no tiene homólogos con función conocida; los genes *mceG* y *mceH* codifican para un transportador ABC y su proteína accesoria, también *mceF* esta involucrado en este proceso; finalmente, se desconoce la función del gen *mceE* (30). La elaboración de las bacteriocinas de esta clase involucra la síntesis de los sideróforos Enterobactina y de Salmoquelina que luego son adaptados a la microcina (3).

3.3.3. Pyocinas y bacteriocinas *phage-like*.

Las Pyocinas son bacteriocinas producidas por *Pseudomonas*, se encuentran codificadas en cromosoma. Su producción basal es muy baja, sin embargo, se puede incrementar por la presencia de agentes mutagénicos como los rayos UV o la Mitomicina C. Se conocen tres grupos de Pyocinas: las llamadas Pyocinas tipo R (*R-type*) que son similares a la cola contráctil del bacteriofago P2, despolarizan la membrana citoplasmática interna y son resistentes a proteasas; las Pyocinas tipo F (*F-type*) son parecidas a la cola no contráctil del bacteriofago *Lambda* y también son resistentes a proteasas; por otro lado, las Pyocinas de tipo S se parecen a las colicinas y son sensibles a proteasas (31).

Pyocinas *R-type*. Esta bacteriocina está codificada por un *loci* que contiene alrededor de 20 genes y ocupa unos 12 kbases. Tienen forma de tubo hueco compuesto por una vaina de 120 nm de largo por 15 nm de ancho compuesta por 34 anillos y que es contráctil, se encuentra rodeando un centro no contractil de 5.7 nm de ancho. Los receptores para esta bacteriocina son lipopolisacaridos de membrana. Una vez que la Pyocina se ha unido a la membrana de la célula sensible hay una rápida contracción de la envoltura y el centro penetra a través de la membrana externa. Esta Pyocina se inserta en la membrana interna despolarizándola y causando una inhibición del transporte, matando a la célula en el lapso

de unos 20 minutos. Ejemplos de estas bacteriocinas son las Pyocinas R1 al R5 (31; 32).

Pyocinas *F-type*. Vistas al microscopio electrónico parecen un tubo flexible de unos 100 nm de largo por 10 nm de ancho compuesto por 23 anillos, cada uno por una proteína de 19 kDa. De uno de los extremos salen una fibras de aproximadamente 40 nm de largo que son usadas por la Pyocina para reconocer a la célula sensible, determinando así el tipo de Pyocina y su especificidad. El *cluster* que codifica para esta Pyocina consta de 12 genes, 6 implicados en la formación del tubo y 6 necesarios para las fimbrias; el *loci* se encuentra río abajo del *cluster* de la Pyocina *R-type* (32).

Pyocinas *S-type*. Estas bacteriocinas se parecen a las Colicinas y, a diferencia de las *R-* y *F-type*, son sensibles a proteasas. Están formadas por dos proteínas, la más grande es la que contiene la actividad bactericida mientras que la pequeña actúa como proteína de inmunidad (su secuencia es similar a la de la proteína de inmunidad de la Colicina E2). La Pyocina *S-type* está constituida por cuatro dominios; su extremo N-terminal de 240 aminoácidos es el responsable del reconocimiento de la célula sensible usando receptores para sideróforo; un segundo dominio de función desconocida; un dominio central de 240 aminoácidos encargado de la translocación a través de la membrana que necesita del sistema TonB o el TolQRA; el último dominio ubicado en el extremo C-terminal de 130 aminoácidos es el responsable de la citotoxicidad de la bacteriocina y tiene actividad de DNAsa (31).

El extremo N-terminal de la proteína de inmunidad interactúa con el dominio C-terminal de la bacteriocina para inhibir su capacidad de reconocimiento de la célula sensible. Bajo condiciones limitantes de hierro, las células resultan más sensibles debido a que deben sintetizar más proteínas de membrana para la captura del hierro, y estos funcionan como receptores para las Pyocinas *S-type* (31).

3.3.4. Bacteriocinas tipo lectina (*Lectin-like*).

Las bacteriocinas tipo lectina (*Lectin-like bacteriocins*) contienen dominios en tándem similares a los dominios de lectinas de unión a manosa (MBL, del inglés *Mannose Binding Lectin*) que ciertas plantas monocotiledoneas usan como mecanismo de defensa para reconocer carbohidratos presentes en bacterias patógenas (33). También se han encontrado algunas MBL que presentan actividad antifúngica, insecticida o nematicida (34). Las bacteriocinas *lectin-like* fueron identificadas inicialmente en las *gama*-proteobacterias *Pseudomonas* y *Xanthomonas*, y posteriormente fueron encontrados en las *beta*-proteobacterias del género *Burkholderia* (35).

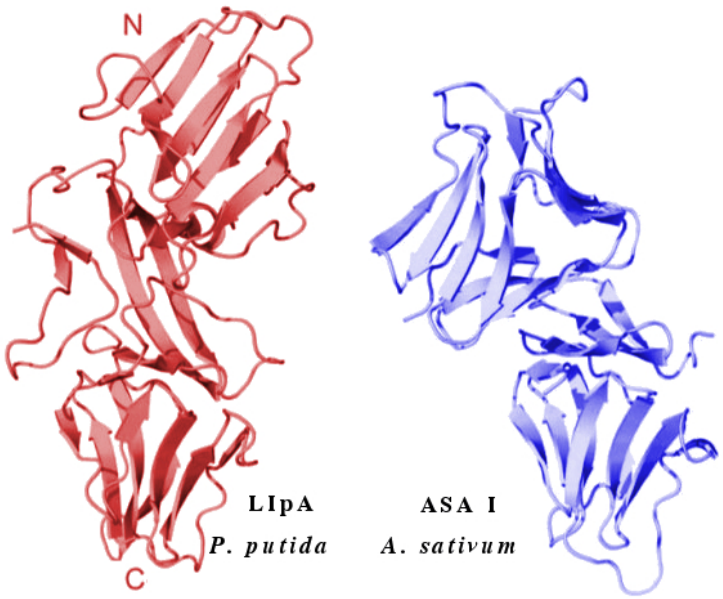

Figura 3.12: Estructura de bacteriocinas *Lectin-like*. En color rojo se muestra la estructura de la bacteriocina tipo Lectina de *Pseudomonas putida*, en azul la Lectina ASA I del ajo *Allium sativum*. Se puede apreciar el arreglo similar del dominio C-terminal de las tres estructuras y diferencias en el plegamiento del dominio N-terminal.

Estas bacteriocinas están compuestas por dos dominios de MBL que se caracterizan por tener varios sitios potenciales para la unión de residuos de manosa (Figura 3.12). Algunas de ellas son sintetizadas como precursores con secuencia señal dependientes del sistema transportador Sec. El dominio N-Terminal está implicado en la especificidad del reconocimiento de la célula sensible mientras que el C-Terminal lleva a cabo la actividad bactericida (35), aunque se desconocen los detalles de este último proceso.

3.3.5. *Bacteriocin-Like Inhibitory Substances* (BLIS).

Los BLIS son substancias inhibitorias semejantes a bacteriocinas en cuanto a su naturaleza protéica, sin embargo, difieren en algunas características como su mecanismo de unión a la célula blanco o que no son activadas por el sistema SOS (36; 37).

3.4. Bacteriocinas producidas por *Archaea*.

Los péptidos antimicrobianos producidas por *Archaea* se denominan Archaeocinas, para diferenciarlas de los producidos por bacterias. Fueron inicialmente descritas en los años 80's. Dentro de este grupo de bacteriocinas, las Halocinas han sido las más estudiadas. Son de tamaños diversos desde 3.6 kDa, como la Microhalocina S8, hasta 35 kDa como la ha-

locina H4. Las microhalocinas son resistentes a ácidos, bases, solventes orgánicos, salinidad y temperaturas extremas, y al igual que las bacteriocinas producidas por bacterias, las halocinas también tienen variados espectros de inhibición. Sin embargo, no se han encontrado halocinas que inhiban el crecimiento de bacterias (38). Por otro lado, pocas de ellas han sido caracterizadas (11; 39).

3.5. Péptidos antimicrobianos en Eukariotas.

Las proteínas y los péptidos antimicrobianos presentes en eukariotas has sido estudiados desde hace algunas décadas y, en concordancia con la nomenclatura de las bacteriocinas, estos péptidos se han denominado Eukariocinas. Suelen ser de tamaño pequeño, resistentes al calor, con alto contenido de residuos de cisteína, carácter anfipático y cuya acción antibacteriana en general se produce atacando a la membrana celular de diversas formas. Estas eukariacinas han sido encontradas en mamíferos, anfibios, artrópodos, insectos, protozoarios y plantas, y parece ser que están ampliamente distribuidas por todo el dominio eukariota (12).

3.6. Bioinformática de bacteriocinas.

La búsqueda experimental de bacteriocinas representa un esfuerzo desafiante debido al enorme espacio de variables que debe explorarse (número de potenciales cepas productoras multiplicado por el número de condiciones posibles de expresión de bacteriocinas, multiplicado por el número de cepas indicadoras usadas). Por otro lado, La disponibilidad de genomas bacterianos completamente secuenciados es una fuente de información para una nueva estrategia de búsqueda de bacteriocinas por medio de predicciones bioinformáticas obtenidas por el análisis genómico. Las bacteriocinas pueden ser rastreadas directamente en el genoma basándose en la identidad de secuencia con bacteriocinas conocidas, con motivos conservados o con los genes biosintéticos. El contexto genómico de las bacteriocinas es de suma importancia en una búsqueda bioinformática, ya que en él se encuentran genes que codifican para modificaciones postraduccionales, regulación, procesamiento de péptido líder, transporte y mecanismos de inmunidad.

Para bacteriocinas relativamente grandes, la búsqueda por similitud de secuencia se puede realizar con la herramienta de BLAST, cuyo algoritmo de alineamiento permite determinar el grado de identidad entre las secuencias comparadas. Sin embargo, las bacteriocinas pequeñas suelen tener una identidad conservada más baja por lo que el uso de BLAST no siempre da buenos resultados. Por ello se necesitan estrategias diferentes, como por ejemplo, utilizar firmas de aminoácidos o los motivos conservados como objeto de búsqueda. Otra

estrategia es generar bases de datos específicas de bacteriocinas en donde se puede realizar búsqueda por BLAST disminuyendo el valor de corte de los datos sin que ésto represente un aumento en el número de secuencias que son falsos positivos.

Una de estas bases de datos especificas para bacteriocinas es BACTIBASE (40). Contiene las bacteriocinas mejor caracterizadas y permite una comparación con secuencias problema (secuencias *query*) por medio de herramientas como BLAST y Modelos ocultos de Markov. Otra base de datos es BAGEL3 (41), que ofrece una plataforma para hacer búsqueda automatizada de bacteriocinas en un determinado replicón.

Otra estrategia que ha dando buenos resultados, sobre todo para rastrear bacteriocinas pequeñas, es la denominada *Genome mining* que se basa en la utilización de secuencias del contexto genómico (genes que codifican para modificaciones postraduccionales, regulación, procesamiento de péptido líder, transporte y/o mecanismos de inmunidad) de las bacteriocinas para hacer una búsqueda en los genomas secuenciados. Una vez que se ha encontrado alguna de estas proteínas (o genes) se procede a hacer un análisis del contexto genómico para determinar la presencia de los genes que codifican para las bacteriocinas. El éxito de esta aproximación radica en que las secuencias usadas como *query* son más grandes y más conservadas, permitiendo un uso más efectivo de BLAST.

Dentro de los estudios bioinformáticos hechos en materia de bacteriocinas podemos citar el trabajo realizado por G. Dirix en bacterias Gram-positivas (42) y Gram-negativas (43). En éstos se utilizó la firma de doble glicina que funciona como secuencia líder en péptidos que son secretados, entre los cuales se encuentran las bacteriocinas, en combinación con el dominio conservado de la peptidasa perteneciente a la familia C39 encargada de su remoción. En el estudio realizado en bacterias Gram-positivas se escanearon 45 genomas secuenciados hasta ese momento en los cuales se encontraron 29 proteínas con dominios C39 y 48 péptidos candidatos a ser procesados con esta proteasa dentro, de los cuales más de la mitad corresponden con bacteriocinas u homólogos de bacteriocinas y el resto son más cercanos a feromonas (42). En cuanto al estudio realizado en Gram-negativas se escanearon 124 genomas y se encontraron 78 proteínas con dominio C39 y 58 péptidos con firma de doble glicina (43).

En 2009 M. Begley y colaboradores realizaron un estudio bioinformático en busca de nuevos lantibióticos usando como secuencia query a la Lantionina sintasa (LanM), encargada de la deshidratación y la ciclización del péptido precursor de lantibioticos de la clase II. En este estudio se encontraron 89 proteínas LanM de las cuales 61 estuvieron presentes en

especies que no habían sido reportadas anteriormente como productoras de lantibioticos (44).

En un estudio realizado sobre lantibióticos de la clase I se utilizó a la proteína NisC, encargada del proceso de ciclización en la Nisina. En este estudio, A. Marsh y colaboradores (45) escanearon 1178 genomas usando una secuencia como *query*. En un primer escaneo, se identificaron 27 nuevos *clusters* de producción de lantibioticos con los cuales se realizó una subsecuente búsqueda encontrando 22 *clusters* de genes adicionales.

En otro estudio, G. Nicolas (46) reportó la detección de nuevas bacteriocinas en *Streptococcus mutans* buscando en su genoma a las Histidine Kinasas con sus Reguladores de Respuesta (HK/RR) seguida de un análisis del contexto genómico usando bases de datos y herramientas disponibles abiertamente a través de internet. En este estudio, Nicolas encontró un nuevo cluster de genes relacionado con la producción de bacteriocinas no reportado antes.

H. Wang y colaboradores (47) determinaron la ocurrencia de bacteriocinas en las cianobacterias escaneando 58 genomas en los que encontraron 145 *clusters* de genes relacionados con bacteriocinas. Para esta búsqueda se utilizaron como secuencias *query* los dominios de los péptidos señal con doble glicina, el dominio C39 y la lantionina sintasa.

En un estudio realizado por M. Singh y D. Sareen (48) se realizó una búsqueda de lantibioticos en genomas secuenciados usando la proteína LanT, encargada de retirar el péptido señal y del transporte de lantibióticos de la clase I al exterior de la célula, activando finalmente a la bacteriocina. En este trabajo se encontraron 24 nuevos clusters de producción de bacteriocinas.

4 ANTECEDENTES

4.1. El género *Burkholderia*

El género bacteriano *Burkholderia* comprende 116 especies de microorganismos (ver Apéndice A). Durante la realización de este trabajo se llevó a cabo la reasignación de algunas de estas especies a los géneros *Paraburkholderia* o *Caballeronia* (49). Entre ellas *B. phymatum*, *B. phytofirmans*, *B. rhizoxinica* y *B. xenovorans* (transferidas al género *Paraburkholderia*), que contaban con organismos con genoma secuenciado y que se habían incluido en la base de datos de este proyecto, razón por la cual algunos resultados pertenecientes a estas especies aparecen en este trabajo. Dentro del género *Burkholderia* encontramos especies capaces de causar enfermedades a plantas, animales y humanos. Por otro lado, también existen bacterias promotoras de crecimiento en plantas, otras que son útiles en procesos de biorremediación y antagonismo con otros microorganismos. Dentro de las especies más versátiles encontramos a *B. cepacia*, importante en todas estas actividades, aunque también ha cobrado mayor importancia como patógeno oportunista en infecciones nosocomiales, principalmente en pacientes con fibrosis quística (50).

4.2. Antagonismo en el género *Burkholderia*

Diferentes especies del género *Burkholderia* han mostrado un comportamiento antagónico frente a diversos micoorganismos, incluyendo *B. pseudomallei*, *Pseudomonas* patógenas, hongos y algunas bacterias Gram-positivas (51; 52; 53; 54).

En un estudio realizado en Papua Nueva Guinea se observó antagonismo específico de tres cepas de *B. ubonensis* contra 27 cepas de *B. pseudomallei* y de varias cepas de *B. thailandensis* con antagonismo contra *Streptococcus pyogenes*, *S. pneumoniae*, *Bacillus cereus*, *Staphylococcus aureus* y *Escherichia coli*. El estudio preliminar de la sustancia antagónica

sugirió que se trataba de una bacteriocina sensible a pepsina (51).

En otro estudio realizado con cepas provenientes de aislados clínicos se encontraron varias cepas del complejo *Burkholderia cepacia* (BCC, del inglés *Burkholderia cepacia* complex) capaces de competir e incluso desplazar a cepas de *Pseudomonas aeruginosa*. En este estudio no se determinó la naturaleza de las sustancias inhibitorias producidas por las cepas de BCC, sin embargo, podrían ser Pyocinas tipo R parecidas a fagos (52).

También se ha reportado una cepa de *Burkholderia gladioli* con un amplio espectro de inhibición que incluye *Tatumella ptyseos, Pseudomonas aeruginosa, Staphylococcus aureus, Raoultella planticola, Ustilago maydis, Candida albicans, Pantoea ananatis, Escherichia coli, Azospirillum brasilense, Gluconacetobacter diazotrophicus*, así como varias especies de *Burkholderia*, *B.cepacia, B. cenocepacia, B. vietnamiensis, B. latens, B. multivorans, B. dolosa,* y *B. stabilis*(53). A pesar de que se ha visto una alta capacidad antagónica en *Burkholderia*, han sido pocas las bacteriocinas que se han caracterizado en estos microorganismos.

4.3. Bacteriocinas en el género *Burkholderia*

Las bacteriocinas que se han caracterizado en *Burkholderia*, incluyen la microcina Capistruina (24), Bacteriocinas tipo CDI (Contact dependent inhibition) (55; 56), bacteriocinas tipo lectina (57) y colicinas tipo M (58).

4.3.1. Capistruina

La Capistruina es un péptido *lasso* reportado en *Burkholderia thailandensis* E264 del cual se conoce su estructura cristalográfica (24). Los péptidos *lasso* se caracterizan por tener una estructura en forma de nudo donde el extemo C-terminal se ensarta en un anillo formado por el extremo N-Terminal (Figura 4.1).

Esta estructura en forma de nudo favorece una mayor estabilidad y le confiere protección contra proteasas. La formación de este nudo se debe a una ciclización entre el grupo amino de una Gly/Cys en el extremo N-terminal y la cadena lateral de un Glu/Asp en la posición 8/9. El extremo C-terminal es insertado en el anillo formado y queda atrapado por impedimentos estéricos debido a la presencia de la cadena lateral grande de una Phe/tyr/Arg (24). La Capistruina está codificada por un *cluster* de cuatro genes en un operon *capABCD*,

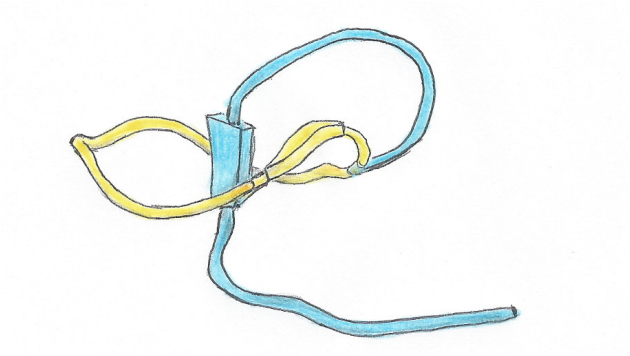

Figura 4.1: Estructura de la capistruina. La Capistruina es un tipo de péptido *lasso* donde el extremo C-terminal se enrolla dentro de un nudo formado por el extremo N-Terminal.

igual que la microcina MccJ25 de *E. coli* (24), (ver Figura 3.8).

4.3.2. *Contact dependent inhibition* CDI

La inhibición de crecimiento dependiente de contacto es una estrategia de competencia interbacteriana que permite tener cierta ventaja para sobrevivir. El mecanismo de CDI está mediado por un sistema de secreción de dos componentes: CdiA y CdiB. En este sistema, la proteína CdiA es exportada a través de la membrana externa usando un canal formado por la proteína CdiB. Las bacterias que contienen este sistema de inhibición (CDI^+) se protegen de una autoinhibición por medio de una proteína de inmunidad llamada CdiI (55).

El *cluster* de genes CDI identificado en *E. coli* está codificado como *cdiBAI*, mientras que en *Burkholderia* se encuentra como *cpbAIOB*. El ORF *cpbO* codifica para un péptido de unos 70 aminoácidos que contiene una señal para unión de lípidos (*lipobox*) en la región N-terminal y se propone que estén localizadas en la región interna de la membrana celular externa (55).

4.3.3. Lectinas

Las bacteriocinas de tipo Lectinas fueron inicialmente descubiertas en *Pseudomonas* y *Xanthomonas* con dominios de lectinas de unión a mannosa (MBL) donde el extremo N-terminal está involucrado de la especificidad para reconocer a la bacteria sensible mientras que el C-terminal es crucial para la muerte celular (57). Las bacteriocinas de tipo Lectinas encontradas en *Burkholderia* tienen baja identidad (menos del 30 %) respecto a las caracterizadas en *Pseudomonas* y *Xanthomonas*, además de diferir en el tamaño del extremo C-terminal.

Sin embargo, conservan los dominios MBL (57).

4.3.4. Colicinas tipo M

Las colicinas de tipo M han sido caracterizadas en *E. coli*, *Pseudomonas* y *Pectobacterium*. Recientemente se identificaron en cepas del complejo *Burkholderia cepacia* y en el grupo de *Burkholderia pseudomallei* bacteriocinas de tipo colicina M que tienen un espectro de inhibición reducido, a las que se denominó BurM (58).

De acuerdo a la proteína de inmunidad, *bmi*, estas bacteriocinas se pueden agrupar en dos categorías: las que poseen una proteína corta (de unos 106 aminoácidos) denominada *BmiA*, con tres hélices transmembranales y sin secuencia detectable para exportación (Sec o Tat); y las que poseen una proteína grande de inmunidad (de alrededor de 126 aminoácidos) con un péptido señal de exportación tipo Sec y una sóla hélice transmembranal, denominada *BmiB* (58).

Ambos tipos de proteínas de inmunidad también se presentan como genes "huérfanos"(sin *burM*) en posiciones genómicas similares en algunas cepas de *Burkholderia*. Los genes húerfanos de tipo *bmiA* son comunes en cepas del grupo de *Pseudomallei*, mientras que las cepas del complejo cepacia codifican genes huérfanos de *bmiB*. Dichos genes podrían representar un reservorio de inmunidad contra este tipo de bacteriocinas (58).

5 MARCO TEÓRICO

5.1. Estrategia de *Genome Mining*.

La secuenciación masiva de genomas, tanto de microorganismos como de organismos eukariotas, ha permitido que conozcamos el contenido genético de esos organismos. Sin embargo, no es suficiente para comprender el funcionamiento de los seres vivos. Para ello es necesario darle sentido a esa información genética y determinar tanto el contenido exacto de genes como los mecanismos de su expresión y regulación.

Este análisis genómico nos permite determinar en qué organismos nos podemos enfocar para estudiar algún proceso en particular o para buscar productos naturales de aplicación médica, industrial, biotecnológica, alimentaria, etc. En el campo de la microbiología, uno de los objetivos más explotados es la búsqueda de metabolitos secundarios útiles en el desarrollo de nuevos materiales, medicamentos y derivados químicos, entre otros.

5.2. Importancia del número de secuencias *query*.

Los trabajos de investigación científica actuales involucran más interdisciplinariedad y eso lleva a la necesidad de trabajar con más información. En materia de bioinformática, anteriormente se realizaban estudios de comparación de secuencias utilizando una sola "secuencia de búsqueda", la secuencia de interés en ese momento en particular. Posteriormente se analizaba otra secuencia relacionada y así sucesivamente. Después de un tiempo se ensamblaba la historia completa de la función biológica. Hoy en día, con el rápido crecimiento de las bases de datos de secuencias y sobre todo de las secuencias de genomas completos, tenemos al alcance una enorme cantidad de información de los organismos que es imposible

explorar de manera experimental a la misma velocidad a la que se generan los datos. Sin embargo, sí es posible analizar de manera teórica (bioinformáticamente) esa información.

Frente a esta situación, se presenta la necesidad de hacer estudios de comparación de secuencias usando muchas secuencias de búsqueda ("***query*** *sequences*") y comparándolas contra las secuencias de las bases de datos ("***subject*** *sequences*") con la intención de determinar la similitud que hay entre ellas.

La bioinformática usa como base la comparación de secuencias, ya sea de ADN, ARN o de proteínas. Para ello se vale de diversas herramientas que usan diferentes tipos de algoritmos. Una de las herramientas más usadas es la de BLAST (*Basic Local Alignment Search Tool*) desarrollada desde 1990 (59). Por medio de la bioinformática podemos realizar estudios de distribución de genes en distintos organismos y con ello extrapolar funciones.

La utilización de muchas secuencias de búsqueda en una sola comparación masiva nos puede brindar, en un único intento, toda la información que necesitamos sobre el sistema biológico que estamos estudiando. Sin embargo, esta búsqueda masiva nos reportará muchos resultados que debemos administrar para poder extraer la información biológica relevante. Por lo tanto, no solo es necesaria una herramienta de comparación masiva de secuencias sino que además es necesaria una herramienta de despliegue masivo de datos que facilite su análisis.

5.3. Necesidad de una herramienta de análisis masivo.

Configurar un programa informático para que una computadora realice, de manera sistemática a gran escala, una comparación de secuencias usando BLAST (o algún otro algoritmo de comparación) es relativamente fácil; la computadora realizará las operaciones en cuestión de algunas horas y generalmente arrojará un número enorme de datos. Sin embargo, organizar y extraer información valiosa de estos resultados es un reto que sin la experiencia y/o la herramienta adecuada puede consumir tanto tiempo que deja de ser rentable.

5.4. Herramientas para BLAST.

Actualmente existen varias herramientas de análisis de resultados de BLAST. Algunos de ellos están acoplados a una interfaz que permite configurar la búsqueda y otros se limitan a incorporar los archivos de salida de BLAST para su análisis, de tal forma que se tienen que

hacer por separado la búsqueda y el análisis.

Entre estos destacan PSAT (60), CGCV (61), ABSYNTE (62) y MULTIGENEBLAST (63). Sin embargo, a pesar de la flexibilidad para configurar ciertos parámetros, el despliegue gráfico usado limita el número de resultados que pueden ser visualizados simultáneamente.

5.5. Lenguajes de programación.

El uso de computadoras para el análisis de datos requiere de lenguajes de programación que sean capaces de procesar la información almacenada en las bases de datos. En bioinformática se puede usar cualquier lenguaje de programación, ya que los datos normalmente están guardados en archivos de *texto plano* (sin formato) (64).

El lenguaje de programación PERL (*Practical Extraction and Report Language*) fue diseñado para trabajar con archivos de texto plano (65), eso lo convierte en una herramienta poderosa para la bioinformática, ya que los datos de secuencias se almacenan en archivos de este tipo. PERL es el lenguaje más usado en bioinformática (64).

El lenguaje **Javascript** fue desarrollado para apoyar el código html en el despliegue de información de manera más dinámica y que trabaja desde el lado del "cliente" (el usuario de la internet) para generar un entorno más gráfico y dinámico (66). Algunas versiones más avanzadas de javascript permiten tener comunicación con el servidor que provee la información al cliente para facilitar el intercambio de información.

El lenguaje **html** no es propiamente un lenguaje de programación a pesar de que muchas veces se le llame así. Este lenguaje fue desarrollado para desplegar información en los navegadores de internet, sus siglas significan *HyperText Markup Language* (lenguaje de marcas de hipertexto) y hace referencia al hecho de que este lenguaje es un interprete de instrucciones de formato que indican como representar la información requerida y cargada a la página a través de *links de hipertexto*, es decir, los archivos que se despliegan en la página, como imágenes o videos, no están guardados en la página sino que son "llamados" desde un directorio. El lenguaje html fue desarrollado en los años 90's en el Centro Europeo de Investigaciones Nucleares (CERN) con el objetivo de compartir información entre la comunidad científica y pronto se convirtió en el lenguaje que hizo posible la existencia de internet como lo conocemos hoy (67).

6 HIPÓTESIS Y OBJETIVOS

6.1. Hipótesis

Un análisis bioinformático de los genomas de *Burkholderia* permitirá identificar genes candidatos que codifiquen para bacteriocinas.

6.2. Objetivos

6.2.1. Objetivo general

Diseñar una estrategia y desarrollar las herramientas que permitan explorar bioinformáticamente, de manera masiva, los genomas de diversas especies del género *Burkholderia* en busca de regiones nucleotídicas de interés, en particular de genes de bacteriocinas.

6.2.2. Objetivos particulares

1.- Escribir *scripts* en PERL para generar una base de datos que permita hacer BLAST local de manera masiva en genomas bacterianos y organice los resultados en función de su posición relativa dentro del genoma.

2.- Desarrollar una gráfica XY interactiva que facilite la visualización y el análisis de los resultados de BLAST.

3.- Localizar, en diferentes fuentes, secuencias de bacteriocinas y proteínas relacionadas para usarlas como secuencias query en la búsqueda bioinformática de potenciales bacteriocinas en *Burkholderia*.

4.- Construir una base de datos propia que permita hacer masivamente búsquedas por BLAST local en los genomas secuenciados del género *Burkholderia*.

5.- Determinar de manera bioinformática la presencia de bacteriocinas potenciales en genomas del género *Burkholderia*.

6.3. Justificación

El análisis de la información en las bases de datos que comprende a los genomas completamente secuenciados permite responder interrogantes de manera relativamente fácil y cada vez con mayor certeza, con antelación a la realización de experimentos. Sin embargo, el manejo masivo de la información generada hace que sea imprescindible el desarrollo de herramientas computacionales que faciliten dicho análisis para explotar mejor esta información.

El desarrollo de la herramienta propuesta en este trabajo permitirá desplegar cantidades enormes de información de manera simultánea, permitiendo hacer estudios bioinformáticos más amplios en menor tiempo.

7 MATERIAL Y MÉTODOS

7.1. Uso de BLAST de manera local.

La disponibilidad de una gran cantidad de información de secuencias de ADN y proteínas en las bases de datos, ya sea individuales o pertenecientes a genomas secuenciados permite realizar estudios bioinformáticos cada vez más productivos. La herramienta más utilizada para realizar comparaciones de secuencias es BLAST (59) desarrollada en 1990 y de la cual se han mejorado los algoritmos de comparación para hacerlos más robustos. Esta herramienta se encuentra disponible a través de internet desde varios sitios web pertenecientes a centros de investigación o universidades. Los sitios más conocidos son el *European Molecular Biology Laboratory* (EMBL, http://www.ebi.ac.uk/Tools/sss/) y el *National Center for Biotechnology Information* (NCBI, http://blast.ncbi.nlm.nih.gov/Blast.cgi). El uso de BLAST a través de internet limita el uso de secuencias de entrada a unas pocas y arroja los resultados de manera individual, de tal forma que es necesario hacer el análisis de los resultados por separado, consumiendo mucho tiempo en la integración de la información.

La herramienta de BLAST también se encuentra disponible para ser utilizada en una computadora de escritorio, no necesita muchos recursos de hardware y puede ser utilizada para hacer búsquedas con muchas secuencias *query*. El programa puede ser descargado desde el sitio del NCBI (http://www.ncbi.nlm.nih.gov/guide/howto/run-blast-local/). La versión para uso local se puede correr por linea de comandos, por lo que es fácil incorporar instrucciones en un *script*[1] de PERL (u otro lenguaje de programación) para configurar y automatizar las búsquedas de acuerdo a las necesidades de cada proyecto.

[1]Un *script* es un conjunto de instrucciones, en un lenguaje de programación, que son interpretados y ejecutados por la computadora. El uso de *scripts* en programación permite codificar, por bloques y de manera ordenada, las acciones necesarias durante un proceso.

El uso de BLAST de manera local necesita la generación de una base de datos. Esta dependerá del uso que se desea, puede ser la base de datos de secuencias no redundantes, los marcadores de secuencia expresada (EST, *Expressed Sequence Tags*), secuencias de referencia, secuencias de proteínas con estructura tridimensional conocida (*Protein Data Bank* o PDB), o genomas secuenciados como es el caso de este trabajo. El programa de BLAST arrojará los resultado en el formato que se le indique (por ejemplo: en forma de tabla, con o sin alineamientos, etc.) y dado que puede llegar a ser mucha información es conveniente estructurar un sistema de carpetas para administrar dichos resultados.

El análisis de la información obtenida normalmente requiere del uso de *scripts* para extraer, organizar, filtrar, etc., los datos a partir de los archivos de salida de BLAST, ya que hacerlo manualmente consumiría mucho tiempo y haría el estudio inviable. Estos *scripts* se escriben con las instrucciones necesarias para realizar las operaciones que cada proyecto en particular demande, por lo que es necesario tener cierto conocimiento de algún lenguaje de programación.

Para este proyecto se utilizó una computadora de escritorio HP pavilion con un procesador Core i7 (3.4 GHz) con 8 GB de memoria RAM y un espacio en disco duro de 2 TB. Se realizó una partición al disco duro e instaló el sistema operativo de Linux, distribución Ubuntu versión 14.04.

Los genomas bacterianos fueron descargados del sitio ftp del NCBI en formato GenBank (*.gbk) y a partir de ellos se generaron las bases de datos necesarios, en este caso tres: una de secuencia de proteínas (denominada faa), otra de secuencias de genes anotados (ffn) y una que permita usar todo el genoma bacteriano sin las anotaciones (fna). Para determinar la identidad de la secuencia *subject* del resultado de BLAST es necesario generar una serie de archivos en un formato tabular (ptt) que contenga los datos de tamaño del gen, su posición real dentro del genoma, el nombre del producto, el identificador GI, entre otros valores, y para este proyecto en partícular fue necesario anexar el valor de su *posición relativa* dentro del genoma.

Para la realización de este trabajo sólo era necesario tener los genomas del género *Burkholderia*, sin embargo, se decidió descargar y generar la base de datos con todos los genomas bacterianos disponibles en el sitio del NCBI. Esto debido a que la estrategía de trabajo permite abordar cualquier tipo de estudio bioinformático que se desee realizar sobre los genomas bacterianos secuenciados y ensamblados, por lo que la construcción de esta base de datos tiene un alcance mayor a este proyecto.

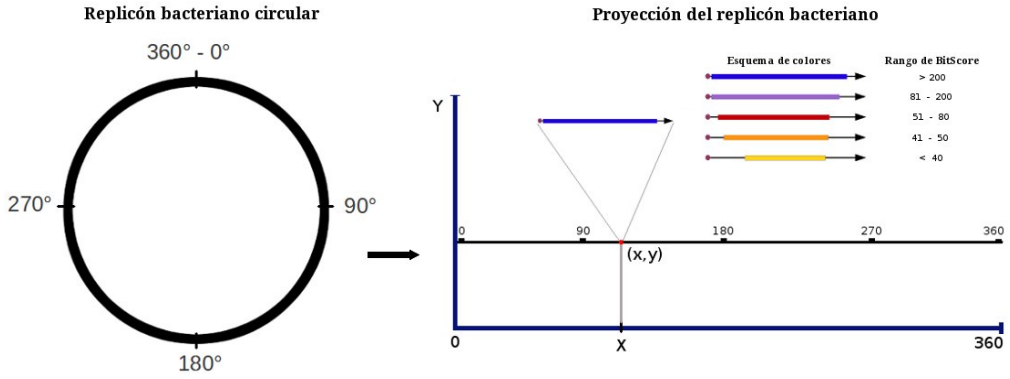

Figura 7.1: Estrategia de representación de datos por *posición relativa*. El replicón bacteriano circular es representado como una linea recta con una longitud definida de 360 unidades (los grados de un círculo). Cada resultado de BLAST es representado mediate un vector con coordenadas "x, y". Adicionalmente se representa una linea gruesa que indica el porcentaje de alineamiento entre las secuencias *query* y *subject*, con un color que refleja el grado de identidad determinado por el valor de BitScore de BLAST.

7.2. Estrategia de representación de datos por *Posición Relativa*

Además de poder administrar la información mediante *scripts*, en muchos casos es necesario representar los resultados para tener una apreciación visual que facilite la interpretación de los mismos. Para este proyecto resulta muy valioso poder determinar de manera rápida si dos o más resultados de BLAST se encuentran cerca uno del otro dentro del genoma de algún microorganismo y es deseable poder visualizar varios genomas simultaneamente.

Para resolver estos dos requerimientos se consideró una estrategia de despliegue de resultados en un formato de gráfica tipo "x, y" que permite representar miles de datos dentro de un espacio definido. En éste, cada resultado de BLAST se puede ubicar muy fácilmente, resaltando automáticamente si dos o más de ellos son cercanos.

Esta estrategia de visualización consistió en hacer una proyección lineal del replicón bacteriano circular y definir un tamaño fijo de 360, que son los grados que tiene un círculo (Figura 7.1). De esta manera se pueden representar en el mismo espacio todos los replicones bacterianos (cromosomas y plásmidos) independientemente de su tamaño real. La proyección lineal permite representar simultaneamente tantos replicones se deseen sin las limitaciones espaciales que tiene la estategia de representación de replicones bacterianos por medio de circulos concéntricos que se utiliza normalmente para la visualización de genomas bacterianos.

La determinación de la *posición relativa* de cada gen se realizó con la siguiente operación aritmética:

$$P_{relativa} = \frac{P_{real} * 360}{Totalpb}$$

Esta se calculó al momento de generar la base de datos, de tal forma que la operación se realiza una sóla vez para cada gen.

7.3. Arquitectura de la herramienta

Para la construcción de esta herramienta de despliegue de datos masivos se optó por una arquitectura modular, de tal forma que cada rutina sea operada por un *script* y sea lo más independiente posible de otros procesos. Esta estructura facilitaría la incorporación a futuro de más módulos que permitirán operaciones requeridas para otros proyectos.

Esta arquitectura modular incluye un *script master* que recibe los datos y canaliza los archivos de entrada a los *scripts* que realizarán las operaciones que el ususario a especificado en un archivo de parámetros de entrada. Para el desarrollo de este proyecto se incluyeron *scripts* que permiten hacer búsquedas a tres niveles: con secuencias de proteínas, de genes y sobre los genomas completos sin las anotaciones. Cada una de estas operaciones se realiza desde un *script* independiente.

Además de los *scripts* necesarios para realizar la búsqueda por BLAST fue necesario escribir una serie de *scripts* que permitieran crear y administrar las bases de datos, renombrar archivos y cambiar formatos, entre otras cosas.

8 RESULTADOS Y DISCUSIÓN

8.1. Búsqueda de secuencias de bacteriocinas

La búsqueda de bacteriocinas de manera bioinformática requiere la utilización de secuencias, ya sea de proteínas o de ADN, que sirvan de "secuencias de búsqueda" (*query sequences*) con las que se pueda rastrear bacteriocinas en los genomas secuenciados. La estrategia usada en este trabajo es de una comparación masiva de secuencias usando la herramienta de BLAST instalada de manera local en una computadora personal de escritorio. Esto nos permite usar una gran cantidad de secuencias de entrada y realizar la comparación de secuencias en poco tiempo.

Para construir una base de datos propia, se obtuvieron un total de 3780 secuencias de bacteriocinas y genes relacionados de las bases de datos de BAGEL (41), BACTIBASE (40), PRODOM (68), CCD (69), MEROPS (70), PIRSF (71) y TC-DB (72). Estas secuencias se organizaron en un sólo archivo para realizar una búsqueda masiva de bacteriocinas, y por otr lado, se organizaron en diversos archivos según su categoría para realizar búsquedas más dirigidas. Adicionalmente, se buscaron y organizaron secuencias de *locus* de bacteriocinas representativas para realizar una búsqueda más sistemática.

8.2. Base de datos de bacteriocinas

Estas secuencias, organizadas en diversos archivos, constituyen una base de datos de bacteriocinas y genes relacionados que incluye 3780 secuencias. La utilidad de esta base de datos radica en que puede ser utilizada para buscar bacteriocinas, muy fácilmente y de manera exhaustiva, en cualquier otro género bacteriano.

8.3. Desarrollo de la herramienta de visualización

La herramienta de búsqueda masiva por BLAST fue desarrollada en lenguaje de programación PERL (v5.18.2). Consta de varios *scripts* que son administrados por un *script master*. Las funciones que lleva a cabo esta herramienta incluyen la lectura de los parámetros de búsqueda, lectura del archivo de secuencias, verificación del formato *fasta* de las secuencias, direccionamiento de la búsqueda según los parámetros, comparación de secuencias *query* y *subject* por BLAST, lectura y organización de los archivos de resultados, filtrado de datos según su valor de E-value/BitScore para eliminar resultados espurios, escritura de resultados en formato *Json* para poder ser visualizados, escritura del archivo html que permitirá ver los resultados en forma de *Gráfica* y de *Tabla*. Además de estos *scripts*, fue necesario escribir otro conjunto que permitiera administrar los genomas de *Burkholderia* y darles el formato requerido para el uso de BLAST.

Generación de la base de datos.

Antes de poder realizar búsquedas por BLAST es necesario configurar y dar un formato a una base de datos de los genomas bacterianos. Este formato lo realiza el mismo programa de BLAST y sólo es necesario hacerlo una vez. Para realizar esta operación se escribió un conjunto de *scripts* de PERL que leyeran los archivos en formato GenBank (*.gbk) y extrajeran de ellos la información necesaria para construir archivos requeridos con otros formatos. Los formatos usados tienen como extensiones *.faa para los archivos con secuencias de proteínas, *.ffn para los archivos que contienen las secuencias de ADN de los genes anotados, *.fna para los archivos con la secuencia del genoma sin las anotaciones de los genes. Los archivos *.ptt para la información sobre las anotaciones y *.ing para las secuencias intergénicas fueron escritos especialmente para esta herramienta (los archivos con las secuencias intergénicas no son necesarios para la realización de este trabajo pero ya que resultarían útiles para otro tipo de proyectos se decidió construirlas). Todos estos archivos se organizaron bajo el siguiente esquema de carpetas.

```
00_NCBIdata/
01_Scripts/
02_faa_files/
02_ffn_files/
02_fna_files/
02_gbk_files/
02_ing_files/
02_ptt_files/
09_QuerySequences/
10_Jobs_folder/
```

La carpeta `00_NCBIdata` es donde se descargan los archivos *.gbk de manera temporal para ser usados en la generación de todos los demás archivos. La carpeta `01_Scripts` contiene todos los *scripts* de PERL que conforman la herramienta BLAST-XYplot Viewer. Las carpetas `02_*_files` contienen los archivos *.faa, *.ffn, *.fna, *.gbk, *.ing y *.ptt respectivamente. En el folder `09_QuerySequences` se localizan los archivos con las secuencias *query* usadas para la búsqueda de bacteriocinas. Finalmente, en la carpeta `10_Jobs_folder` se crean los directorios de cada búsqueda para escribir los archivos de resultados.

Herramienta de BLAST-XYplot Viewer.

En una primera etapa, cada función se realizaba por un *script* independiente, pero en la versión final todos estos *scripts* fueron unificados en sólo 8 (Apéndice B). Un *script* que recibe los datos y las secuencias, un *script* que revisa el formato fasta de las secuencias, un *script master* que direcciona los datos y las secuencias según los parámetros, un *script* para hacer búsqueda con secuencias de proteínas, uno para hacer búsqueda con secuencias de ADN sobre anotaciones de genes, otro para realizar búsqueda con secuencias de ADN sobre el genoma sin anotaciones, un *script* que permite hacer búsqueda de firmas en las regiones intergénicas y un *script* que permite generar la base de datos necesaria para hacer BLAST. En conjunto, todos estos *scripts* suman más de 6000 lineas de código desarrollado en PERL.

El archivo de **Parámetros** especifica las condiciones bajo las que se realizará la búsqueda de secuencias. Tiene un formato de texto plano y puede ser llenado manualmente. Los parámetros incluyen título, tipo de BLAST, base de datos contra la que se hará la búsqueda, tipo de *score* que se calculará y máximo número de resultados permitidos, si es que se desea restringir el tamaño de los archivos de salida.

```
JobTitle: Descripción del trabajo
SubjectDB: Genus /TaxGroup
Genus: Burkholderia
TaxGroup:
BlastType: Proteins/Genes/Genomes
ScoreType: Homology /Proximity
MaxResults: 50000
```

JobTitle. Permite tener una descripción breve de la búsqueda que se está llevando a cabo.

SubjectDB. Permite definir la Base de Datos contra la que se llevará a cabo la búsqueda. Sólo acepta una de dos opciones: *Genus* o *TaxGroup*. La opción de *Genus* realizará una búsqueda de las secuencias *query* en todos los genomas secuenciados y disponibles pertenecientes al género bacteriano especificado en la variable "***Genus:***". Por otro lado, la opción *TaxGroup* realizará una búsqueda de las secuencias *query* en todos los genomas secuenciados y disponibles pertenecientes al grupo taxonómico especificado en "***TaxGroup:***" permitiendo una búsqueda mucho más extensa. Las opciones *Genus* y *TaxGroup* son mutuamente excluyentes, de tal forma que si se coloca *Genus* no es necesario especificar un grupo taxonómico y viceversa.

BlastType. Admite una de tres opciones posibles. *Proteins* hará una búsqueda sobre las proteínas anotadas en los genomas bacterianos, utiliza secuencias de proteínas. *Genes* hará la búsqueda sobre las anotaciones de genes y requiere secuencias de ADN. *Genomes* hará la búsqueda de secuencia de ADN sobre el genoma sin anotaciones. Esta última opción es particularmente útil para buscar genes que puedan estar mal anotados. Los resultados de esta búsqueda no arrojarán información sobre el gen, sólo la posición en nucleótidos del segmento alineado.

ScoreType. Los resultados de BLAST incluyen un valor de E-value y uno de BitScore que permiten valorar numéricamente la calidad de la comparación de secuencias. La herramienta de BLAST-XYplot Viewer utiliza el valor de BitScore para calcular un nuevo score de *proximidad* o de *homología*. El valor de score de *Homology* evaluará cuantas veces una secuencia es encontrada como *subject* durante la búsqueda por BLAST, resaltando la secuencia que tiene identidad con un mayor número de secuencias *query*, mientras que el score de *Proximity* evalua si los resultados de búsqueda se encuentran cerca unos de otros dentro

del genoma, resaltando los posibles *clusters* de genes.

MaxResults. Permite limitar el número de resultados finales en caso de que sea necesario. En este caso, se realiza un ordenamiento de datos de mayor a menor valor de BitScore y se incluirán solamente el número especificado de resultados de BLAST. Colocar un número alto (ej. 100000 o 500000) permitirá incluir todos los resultados de la búsqueda.

El archivo de **secuencias** debe estar en formato *fasta*, un formato estándar para bioinformática que consiste en escribir la secuencia de caracteres sin ningún tipo de formato (por ejemplo, tamaño o tipo de fuente, texto en negritas o itálicas). Esta herramienta requiere que cada secuencia tenga un cabezal para distinguirlo de las demás secuencias, como se muestra a continuación.

```
>Primer_Cabezal
MLGLVRIALRRPYTFVVLAILILIIGPLSALKTPT
DIFPDIRIPVISVVWQYTGLPPDQMAGRITSTFER
SLTTTVNDIQHIEAESVNGF...

>Segundo_Cabezal
MSSEPIVTSQPVPRRRLVIIGVIGIAIAITVVAAG
VTLRAVDARNLKSWTNAQTVPTVTVIHPVSAANGP
TLDLPSHLEAYSRAPIFARV...

>Tercer_Cabezal
MRELTSYELQAVSGGDFFSDLGNAFTAVTNVVASA
TVGALWGGGIGSTMGGRYGATAGGWGFSAISAGVA
LIFGGVLGAAVGAATGA...
```

El flujo de trabajo de la herramienta inicia con la lectura de secuencias y las canaliza según los parámetros especificados para la búsqueda por BLAST. La comparación de secuencias se realiza sistemáticamente comparando todas las secuencias del archivo de secuencias contra un primer genoma (determinado en el archivo de parámetros), una vez terminado este bloque se comparán todas las secuencias contra un segundo genoma y asi sucesivamente hasta terminar. Esto permite generar archivos de salida pequeños para cada conjunto de genes, de tal forma que no se consume memoria RAM (pero sí espacio en disco) en la computadora, dejando libre este recurso para el procesamiento de datos que hace BLAST. Posteriormente se realiza una lectura de estos archivos de salida para generar un archivo único, se filtran los datos por E-value para eliminar resultados espurios y se calculan los valores de *score* por

Homology o Proximity según sea el caso. El archivo final de resultados se escribe en un formato de texto plano organizado por columnas separadas por tabuladores. Puede ser abierto en hoja de cálculo e incluir desde unos pocos resultados de BLAST hasta medio millón o más según el tipo de búsqueda que se haya realizado.

El análisis de datos es posible a través de la terminal de Linux por medio de linea de comandos usando los programas GREP y NAWK que permiten extraer rápidamente información de archivos de texto plano sin necesidad de abrirlos de manera gráfica, por lo que no consumen tantos recursos de memoria RAM. Estos dos programas, instalados automáticamente en el sistema operativo Linux, usan condiciones para explorar dentro de los archivos e imprimen en pantalla el texto que cumple con esas condiciones. De esta forma, se puede determinar como condición la presencia de una secuencia de caracteres que indiquen algún gen/proteína, microorganismo, valor de BitScore/E-value, etc y GREP o NAWK imprimirán en pantalla los resultados de BLAST que cumplan con esa condición.

Los resultados de BLAST también se pueden abrir en programas como GNUPLOT o en una hoja de cálculo tipo Excel o LibreOffice-Calc para ser graficados. En este caso también pueden ser ordenados según las columnas que se especifiquen. Esta forma de análisis permite administrar los datos de manera más interactiva que la linea de comandos de Linux. Sin embargo, estos programas de hoja de cálculo no están optimizados para manejar demasiados datos por lo que tardan mucho tiempo en realizar las operaciones de ordenado de datos y no resultan útiles con búsquedas masivas.

La combinación de estos dos métodos de análisis resulta más versátil ya que con GREP o NAWK se pueden seleccionar subgrupos de resultados de manera específica que pueden ser abiertos en una hoja de cálculo y ser analizados por separado.

Estrategia de visualización de datos por *posición relativa*.

Esta estrategia de visualización de datos consiste en la proyección lineal del cromosoma bacteriano circular. Bajo este esquema, cada replicón es proyectado como una linea recta de longitud fija de 360, que son los grados que tiene un circulo. De esta manera tenemos un espacio delimitado donde cualquier replicon bacteriano (cromosoma o plásmido) puede ser representado independientemente de su tamaño real. Usando esta proyección cada gen puede ser mapeado por su *posición relativa* (de 0 a 360) en vez de su posición real en nucleótidos (Figura 8.1). En un sistema de coordenadas "x,y", el eje "x" representa la longitud del replicon con valores de 0 a 360, mientras que el eje "y" representa el número de replicones usados como base de datos para la búsqueda por BLAST y tiene valores desde 1 (si solamente

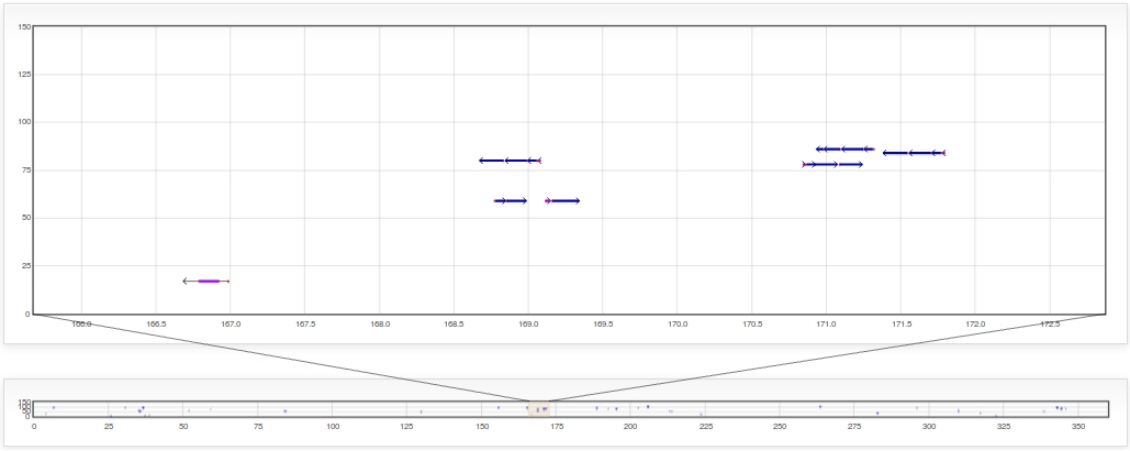

Figura 8.1: Gráfica tipo *x,y* de representación de genes por posición relativa. Resultados de búsqueda por BLAST representados por su posición relativa x, y dentro del genoma. El eje x representa la posición relativa dentro del genoma mientras que la posición en el eje y representa el número del replicón al cual pertenece dicho gen. La zona ampliada en el eje x permite visualizar los *clusters* de genes.

se uso un replicón) hasta el máximo número de replicones usados para realizar la búsqueda. En este trabajo fueron 123, pero la herramienta ha sido desarrollada para incluir todos los replicones bacterianos disponibles en la base de datos del *National Center for Biotechnology Information* (NCBI) que suman mas de 5000.

Al realizar búsquedas por BLAST de algún gen/proteína en particular, cada resultado de esa búsqueda (*subject sequence*) es representado como un punto con coordenadas "x, y", donde el eje "x" representa la posición relativa dentro del genoma y el eje "y" el número de replicon al cual pertenece dicho resultado. Además del punto, ubicado en el origen del gen, se representa un vector con una magnitud proporcional al tamaño del gen y una dirección acorde al sentido de la transcripción. El alineamiento entre la secuencia de búsqueda (*query*) y la encontrada (*subject*) se representa por medio de una linea gruesa sobre el vector mencionado, con una longitud proporcional al largo del alineamiento y un color que depende del valor de BitScore dado por BLAST. En color azul obscuro se representan los valores de BitScore mayores a 200, en morado los valores entre 80 y 200, en rojo los valores entre 50 y 80, y en color amarillo aquellos valores de BitScore menores a 50 (Figura 8.1). De esta manera se tiene una representación visual de la calidad y cobertura del alineamiento entre las dos secuencias.

Para cada *posición relativa* los valores de "x" pueden ser de 0 a 360 con tres decimales para evitar sobrelapamiento de vectores. Los valores de "y" siempre serán números enteros desde 1 hasta el número máximo de replicones usados (más de 5000 en esta base de

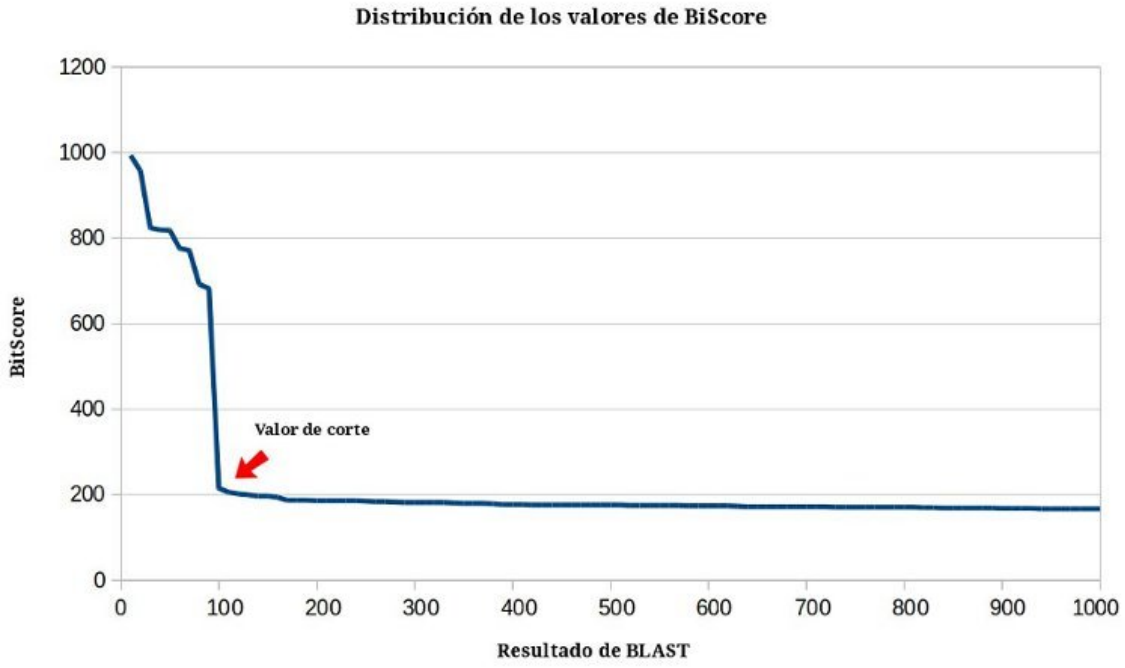

Figura 8.2: Filtrado de datos por valor de BitScore. Los resultados de búsqueda por BLAST son filtrados de acuerdo al valor de BitScore para eliminar datos no significativos y disminuir el número de ellos que será representado en la gráfica.

datos).

Con esta estrategia de visualización se tiene la ventaja de usar la interactividad de un gráfico como el zoom o el barrido para analizar los datos, teniendo en cuenta que se pueden representar miles de resultados de BLAST pertenecientes a muchos genomas al mismo tiempo, es posible visualizarlos todos y/o enfocarse en uno sólo en particular en tiempo real y con pocos recursos computacionales. Otra ventaja que tiene este método es que si dos o más resultados de BLAST se encuentran cerca dentro del genoma bacteriano se pueden identificar muy fácilmente, permitiendo detectar posibles operones (Figura 8.4).

El uso de BLAST en una computadora de escritorio permite el manejo de muchas secuencias de entrada y genera miles de resultados por lo que se vuelve necesario filtrar los que son más significativos y eliminar aquellos que son espurios. Este filtrado se realiza con un *script* que analiza los valores de BitScore de los resultados obtenidos para cada secuencia de entrada. El *script* detecta la caída en los valores de BitScore pertenecientes a los resultados de BLAST menos significativos y los elimina del conjunto de datos útiles (Figura 8.2).

8.4. Desarrollo de una interfaz gráfica

El análisis de datos por linea de comandos y hoja de cálculo es poderosa y permite extraer información necesaria para hacer interpretaciones biológicas de la búsqueda por BLAST, sin embargo, requiere de habilidades en el manejo de sistema operativo Linux así como de los programas GREP y NAWK, que pocas personas tienen. Por esta razón se desarrolló una interfaz gráfica que permita visualizar y analizar los datos sin tener muchas habilidades en computación.

Para el desarrollo de una interfaz gráfica más amigable fue necesario el uso de otros lenguajes de programación como html y javascrip. Se eligió html como base para el despliegue de datos porque permitiría a mediano plazo incorporar esta herramienta a un servidor para ser utilizada a través de internet por cualquier usuario. El uso de javascript permite a las páginas html desplegar información de manera más interactiva ya que codifica rutinas en *scripts* accesorios que son llamados por la página web cuando se requieren.

La interfaz desarrollada para esta herramienta consta de una página web que le permite al usuario configurar los parámetros de búsqueda e introducir las secuencias en un cuadro de texto o cargarlas directamente desde un archivo. Los resultados se despliegan en dos formatos, una Tabla y una Gráfica.

La Tabla (Figura 8.3) tiene la estructura de una hoja de cálculo básica y realiza algunas funciones de ordenamiento de datos que permiten filtrar la información según ciertas condiciones. En el código de esta Tabla se incorporó la opción de mandar a escribir un nuevo archivo de resultados actualizado que pueden visualizarse en la gráfica. El código para desarrollar la Tabla fue adaptado de JQuery (https://www.jquery.com/), una librería de javascript que tiene *plug-ins* disponibles para construir y configurar tablas de datos para páginas web.

La información que contiene la tabla para cada resultado de BLAST es la siguiente: las coordenadas "*x,y*" de la *posición relativa*, el cabezal de la secuencia *query*, los valores de E-value y BitScore, el nuevo valor de score calculado (Homología o Proximidad), el código GI de la secuencia *subject*, el microorganismo al cual pertenece la secuencia *subject* y el producto del gen. La Tabla permite ordenar los resultados de BLAST en función de cada uno de estos valores.

El desarrollo de la Gráfica (Figura 8.4) representó un reto mayor ya que no existen *plug-ins* que permitan generar un gráfico con las características necesarias para esta

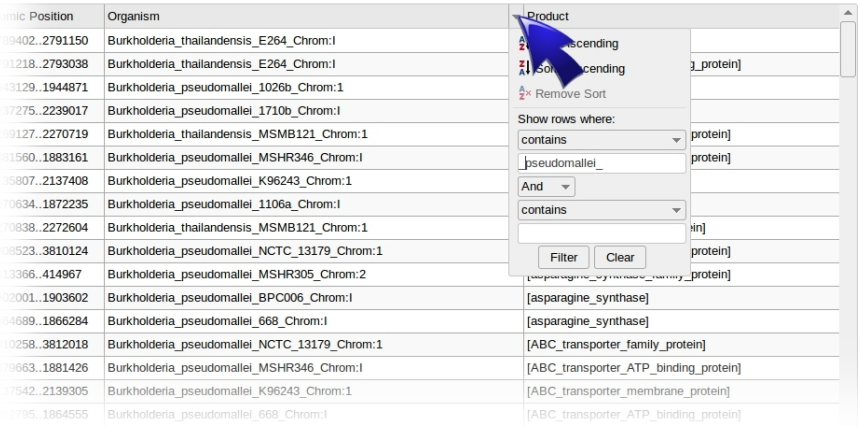

Figura 8.3: Detalle de la Tabla con los datos de los resultados de BLAST**.** Los resultados de búsqueda son enlistados en una tabla que tiene funciones básicas de hoja de cálculo útiles para ordenar y/o filtrar datos.

herramienta. Sin embargo, fue posible adaptar algunas opciones básicas disponibles en Flot (http://www.flotcharts.org/), una librería de javascript que permite incorporar gráficas en páginas web. Sobre esta estructura se desarrolló el código específico para esta herramienta que permitió el despliegue de los datos acorde a la estrategia de representación de los genes por *posición relativa*.

8.5. Disponibilidad de la herramienta BLAST-XYplot Viewer

Ya con la herramienta en el punto de desarrollo donde los datos podían ser proyectados en un Tabla y una Gráfica que permitieran desplegar y analizar los resultados de BLAST, se inició un proyecto de colaboración con el Dr. Gustavo Rubín de la Facultad de Ciencias de la Computación de la BUAP para poner la herramienta en un servidor web que la hiciera disponible para la comunidad científica. En dicho proyecto de colaboración, el alumno Rodrigo Alberto Cuevas Vede fue quien se encargó de desarrollar el código necesario para dicho servidor.

La herramienta se encuentra disponible para su uso desde cualquier computadora con internet en la dirección ***www.blast-xyplot-viewer.icuap.buap.mx***. Cuenta con un tutorial que describe las características del programa y su forma de uso. Puede realizar búsquedas de secuencias de proteína o de ADN en el género bacteriano o grupo taxonómico que se indique.

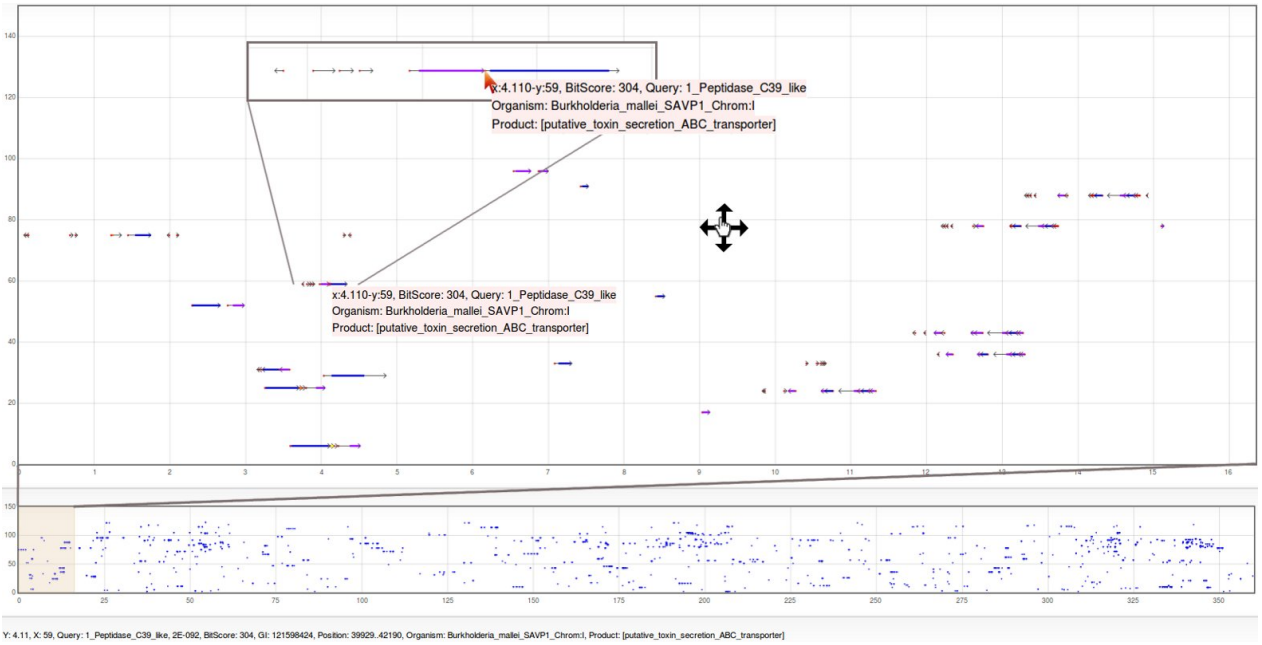

Figura 8.4: Gráfica tipo x,y de representación de genes por posición relativa. Resultados de búsqueda por BLAST representados por su posición relativa x, y dentro del genoma. El eje x representa la posición relativa dentro del genoma mientras que la posición en el eje y representa el número del replicón al cual pertenece dicho gen.

8.6. Potenciales bacteriocinas encontradas en este trabajo

En este trabajo se usaron 38 genomas disponibles de *Burkholderia* que suman 123 replicones y que en conjunto incluyen 248982 genes anotados. La compración de secuencias se llevó a cabo con el programa BLASTp, que usa secuencias de proteína. Cada una de las 3780 secuencias de la base de datos de bacteriocinas se comparó contra las secuencias de proteína codificadas por los 248982 genes del género *Burkholderia*. Esto representa más de 940 millones de comparaciones individuales.

La búsqueda de bacteriocinas potenciales se llevó a cabo de dos maneras. Por un lado se usaron las 3780 secuencias obtenidas de las bases de datos de internet en una comparación masiva de secuencias que implicó alrededor de dos horas de computo y arrojó miles de resultados que fueron concatenados y ordenados en un único archivo de salida de manera automática por la herramienta. Por otro lado, se realizaron búsquedas sistemáticas y dirigidas con secuencias de proteínas pertenecientes a los *loci* de bacteriocinas caracterizadas en mayor o menor medida y que fueron rastreadas en bases de datos y en publicaciones.

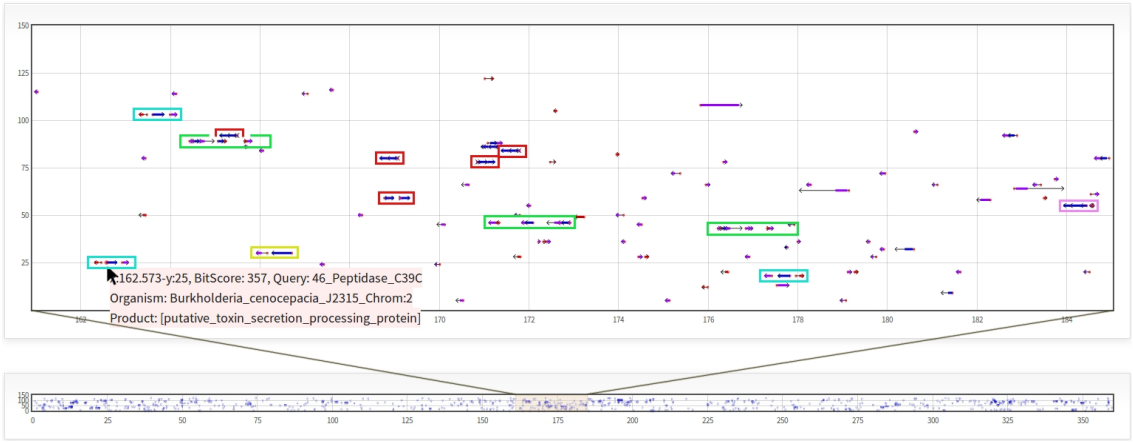

Figura 8.5: Gráfica XY de las potenciales bacteriocinas en el género *Burkholderia*. Para buscar genes que codifiquen para potenciales bacteriocinas se usaron 3780 secuencias *query* y se compararon contra las secuencias pertenecientes a 123 replicones bacterianos de los genomas de *Burkholderia*. Se muestra una ampliación de la zona central del gráfico y con recuadros de colores se resaltan los distintos tipos de *clusters* de genes que la herramienta despliega automáticamente.

Se encontraron 175 *clusters* de genes y 45 bacteriocinas de tipo lectina (Figura 8.5). Cada uno de los 38 genomas tuvo al menos una bacteriocina, en promedio se observaron 1.8 bacteriocinas por replicón (en un rango de 0 a 8) y un promedio de 5.9 *clusters* de genes por genoma. La mayoría de los *clusters* de genes (46.18 %) fueron detectados en especies pertenecientes al grupo de *Pseudomallei*. Todo esto representa una alta incidencia de bacteriocinas en *Burkholderia* no reportado antes, probablemente porque la mayoría de los estudios en bacteriocinas se centran en estudiar sólo una o pocas bacteriocinas.

Estos *clusters* de genes fueron categorizados en seis grupos dependiendo de su estructura genética. El grupo I lo conforman 58 *clusters* que contenían una peptidasa de la famila C39 rodeada de genes que codifican para una maquinaria de exportación del tipo ABC. El grupo II está constituido por las 45 bacteriocinas tipo lectina. El grupo III lo forman 40 *clusters* de genes del tipo CDI (*Contact-dependent inhibition*). En el grupo IV tenemos a 37 bacteriocinas tipo *phage* de las cuales 27 fueron similares a las Pyocinas tipo R y 10 fueron similares a las Pyocinas tipo F. En el grupo V están 22 *clusters* de genes que tienen una proteína cinasa del tipo TOMM (*Thiazole/Oxazole Modified Microcins*). Finalmente, en el grupo VI se encuentran 18 *clusters* de genes que codifican para la microcina tipo *lasso* reportada como Capistruina (Cuadro 8.1).

En este estudio no se detectaron bacteriocinas de tipo Colicina o Pyocinas de tipo S. Sin embargo, tres proteínas hipotéticas fueron observadas recurrentemente como *subject* de secuencias de colicina. Estos resultados tenían un valor de E-value menor al punto de corte

Grupo principal	Tipo	Número	Porcentage
Grupo I	Peptidasas C39	58	24.6 %
Grupo II	Tipo Lectinas	45	20.5 %
Grupo III	CDI	40	18.2 %
Grupo IV	Tipo *Phage*	37	16.8 %
Grupo V	TOMM	22	10.0 %
Grupo VI	Capistruina	18	8.2 %
Total		220	100.0 %

Cuadro 8.1: Número total de potenciales *clusters* de genes de bacteriocinas en *Burkholderia*. El tipo de *cluster* más abundante es el que contiene a las Peptidasas C39. Abreviaciones CDI:*Contact-dependent inhibition*, TOMM:*Thiazole/Oxazole Modified Microcins*

establecido para la búsqueda masiva (E-value de e^{-14}), pero fueron observados en la búsqueda sistemática donde se podían ajustar los valores de corte dependiendo del número de resultados obtenidos para cada búsqueda en partícular. Estas tres proteínas son BamMC406_0333 de *B. ambifaria* MC40-6, Bamb_0324 de *B. ambifaria* AMMD y GEM_3116 de *B. cepacia* GG4. Un análisis de secuencias más detallado reveló que su dominio C-terminal guarda similitud con el dominio citotóxico de las colicinas. Estas tres secuencias fueron usadas como secuencias *query* para buscar más homólogos en *Burkholderia*, sin embargo, no se hallaron secuencias adicionales.

Incorporación de datos no pertenecientes a resultados de BLAST.

Esta herramienta de visualización de datos también permite incorporar información que no proviene de resultados de BLAST. Estos datos pueden ser extraídos directamente de los archivos GenBank (**.gbk*) de los genomas bacterianos y, una vez adapatado su formado de texto (por ejemplo, añadir columnas en blanco para los valores de E-value y BitScore) se pueden anexar al archivo de resultados de BLAST y ser proyectados junto con los resultados de la búsqueda por comparación de secuencias.

Esta flexibilidad de la herramienta nos permitió incorporar a todos aquellos péptidos que tuvieran una firma de doble glicina (-GG-) o glicina-alanina (-GA-). Estas dos firmas son reconocidas por las Peptidasas de la famila C39 y por las enzimas TOMM cinasas (43). Para extraer estos péptidos se usaron *scripts* de PERL que utilizan *expresiones reglares* para identificar caracteres dentro de los archivos de texto. Una vez incorporados estos péptidos al archivo de resultados de BLAST se eliminaron aquellos que no estaban cerca de algún *cluster* de genes (peptidos aislados), ésto permitió obtener las potenciales bacteriocinas asociadas a

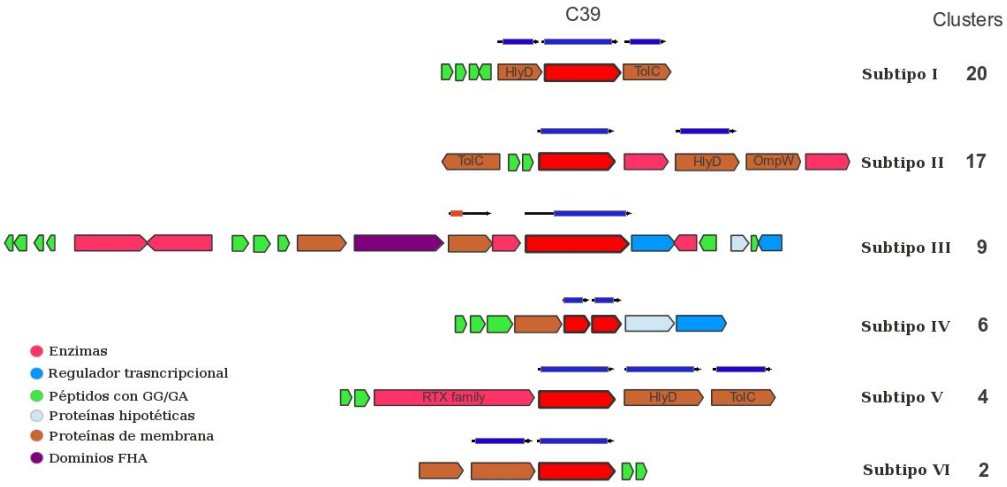

Figura 8.6: Representación esquemática de los *clusters* que contienen Peptidasa C39. Las barras de color azul intenso representan los resultados de BLAST como fueron observados en el gráfico x, y. Los polígonos en diferentes colores representan el contexto genómico real que está conservado dentro del genoma.

Peptidasas C39 y a TOMM cinasas sin tener que hacer búsquedas por BLAST de péptidos pequeños que tuvieran alguna firma GG o GA, ahorrando tiempo de cómputo y de análisis de datos.

8.6.1. *Loci* de potenciales bacteriocinas encontradas en este trabajo

Los *clusters* de genes de potenciales bacteriocinas encontrados en *Burkholderia* se ordenaron en seis grupos de acuerdo a su contenido de genes y se describen a continuación.

Grupo I. *Clusters* de Peptidasas de la familia C39

Los 58 clusters de genes que contienen a una Peptidasa de la familia C39 fueron agrupados en seis subtipos dependiendo de la organización de su contexto genómico (Figura 8.6). El **Subtipo I** está compuesto por un homólogo de HlyD, una proteína con el dominio C39 fusionada con un dominio ABC, y una proteína de membrana externa de tipo TolC. Esta estructura de genes conforma la unidad básica de un mecanismo de transporte a través de las dos membranas en bacterias Gram-Negativas. Junto a este *cluster* se hallaron péptidos que contienen doble glicina o glicina-alanina (GG/GA), lo que los convierte en candidatos a ser bacteriocinas, pues se ha reportado que esta firma es reconocida por la Peptidasa C39 como sitio de corte (43). Este subtipo fue el más abundante, se encontraron 20 *clusters* de genes.

El **subtipo II** tiene un homólogo a TolC, una proteína de fusión con los dominios de Peptidasa C39 y de transportadores ABC, una aciltransferasa, una proteína HlyD, una proteína de membrana externa OmpW, y una polifosfáto cinasa. En medio de este cluster se hallaron péptidos con GG/GA. Este cluster también forma un sistema de transporte completo a través de las dos membranas. Las enzimas aciltransferasa y polifosfáto cinasa podrían tomar parte en el procesamiento de los péptidos para convertirlos en bacteriocinas activas.

El **subtipo III** fue el *cluster* más grande que contiene un dominio C39, está constituido por una proteína ABC, una HlyD y una proteína de membrana externa de la familia NodT formando el sistema de transporte a través de las dos membranas. En la periferia de este grupo de genes hay dos reguladores transcripcionales, una proteína con dominio FHA (*ForkHead-Associated*), y cuatro enzimas: una fosfatasa, una isomerasa, una peptidasa y una cinasa. La función exacta de estas últimas proteínas se desconoce ya que no tienen ortólogos caracterizados experimentalmente, pero es posible que participen en la maduración de los péptidos con firma GG/GA.

El **subtipo IV** está compuesto por una proteína de unión a ATP, la peptidasa C39, dos proteínas hipotéticas, y un regulador transcripcional. Una de las proteínas hipotéticas tiene identidad de secuencia con la Peptidasa C39, mientras que la otra no tiene homólogos con función conocida o propuesta. Aparentemente, este *cluster* codifica para un sistema ABC incompleto ya que carece de los genes que codifican para homólogos a HlyD y TolC. Es posible que los péptidos translocados al espacio intermembrana usen algún sistema de exportación codificado en otra región del genoma y que no esté especializado en el transporte de los péptidos de este *cluster*.

El **subtipo V** está compuesto por las tres proteínas, Peptidasa C39, HlyD y TolC, necesarias para formar un mecanismo de transporte a través de las dos membranas. Río arriba de estos genes se localiza una proteína de la familia de las toxinas RTX.

El **subtipo VI** contiene una proteína TolC, una HlyD y una peptidasa C39 fusionada con un transportador ABC. Este contenido de genes es el mismo que el subtipo I pero difieren en el orden.

Grupo II. Bacteriocinas de tipo Lectina

Se encontraron 45 genes que codifican para bacteriocinas de tipo lectina. Todas las cepas tienen al menos una copia, excepto las cepas de *B. cenocepacia* AU-1054, HI2424 y

MC0-3 que tienen tres copias. Todas las bacteroicinas de tipo lectina fueron encontradas en el cromosoma 1, excepto las de *B. glumae* BGR1 y *B. rhizoxinica* HKI-454 que están presentes en el plásmido 1.

Grupo III. CDI (*Contact-dependent inhibition*)

Las CDI están codificadas por un operón tipicamente compuesto por tres genes *cdiAIB*. La proteína CdiA es la que presenta actividad citotóxica en su dominio C-terminal, CdiI es una proteína variable muy pequeña que actúa como proteína de inmunidad, y CdiB es una proteína de membrana externa encargada de exportar a CdiA (55). Se ha reportado la presencia de un *cluster* de genes, denominado *bcpAIOB*, en *Burkholderia* que codifica para CDI. Se usaron las secuencias de proteína de este *cluster* para rastrear su presencia en las demás cepas de *Burkholderia*. Se encontraron 40 *clusters* de CDI en 21 cepas. En 30 de estos *clusters*, CdiA se encuentra ubicado río arriba de CdiB tal como está descrito por Anderson *et. al.* (55), mientras que en las 10 restantes CdiA está localizado río abajo de CdiB, en una disposición parecida a la que presenta *E. coli* (55).

Grupo IV. Bacteriocinas tipo *Phage*

Se encontraron 37 *clusters* de bacteriocinas tipo *phage*, de los cuales 27 resultaron ser más similares a las Pyocinas de tipo R y los diez restantes fueron similares a las Pyocinas tipo F. La mayoría de las tipo R se localizó en el cromosoma 1, mientras que las tipo F tuvieron una distribución similar, tanto en el cromosoma 1 como en el 2. Solamente un *cluster* se encontró en cromosoma 3 y ninguno en plásmido.

Grupo V. Bacteriocinas con Cinasas tipo TOMM

Los *clusters* de genes que se detectaron de bacteriocinas potenciales con Cinasas tipo TOMM en *Burkholderia* están compuestos por una proteína con dominio FHA, una serine/treonine cinasa, dos precursores TOMM y una proteína de fusión entre una cyclodehidratasa y la proteína necesaria para la formación del anillo de azolina (73). La presencia de este tipo de *cluster* de bacteriocinas fue única del grupo de *Pseudomallei*, que lo porta en el cromosoma 1, y del Complejo *Burkholderia cepacia* (BCC) que lo tiene en el cromosoma 2.

Grupo VI. Microcinas tipo Capistruina

La Capistruina es un péptido tipo *lasso* caracterizado por tener una estructura de nudo, similar a la Microcina J25 de *Escherichia coli* (74). Se ha reportado a *Burkholderia thailandensis* E264 como productora de Capistruina (24), por lo que se usaron las secuencias

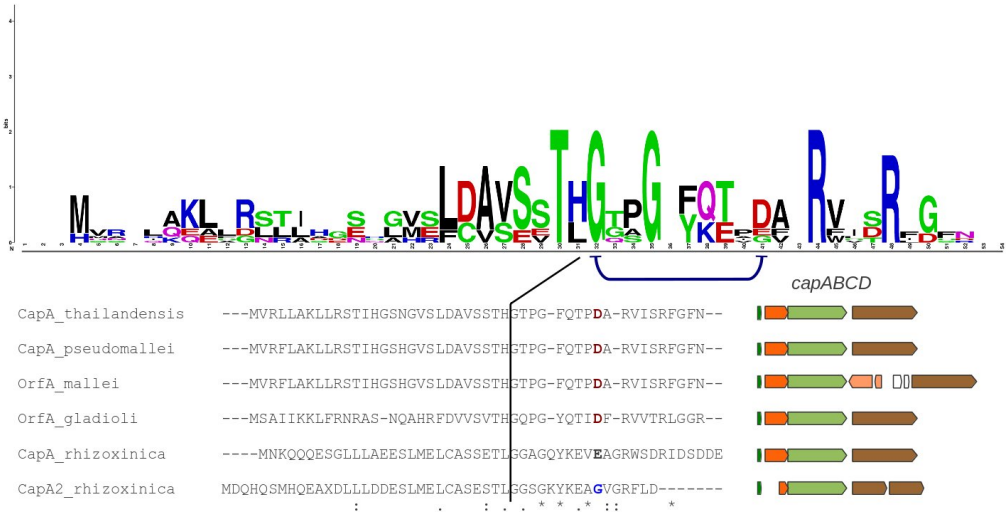

Figura 8.7: Alineamiento de las secuencias de los péptidos de Capistruina. La línea curva muestra los aminoácidos necesarios para la ciclización del péptido. La línea recta muestra el sitio de corte del péptido lider. Se observa que el péptido CapA2_rhizoxinica no contiene el ácido aspártico necesario para la ciclización de la bacteriocina.

de las proteínas implicadas en la síntesis de esta bacteriocina como *cluster* de búsqueda. Este *cluster* se encontró solamente en miembros del grupo *Pesudomallei*, en *B. gladioli* BSR3, y en *B. rhizoxinica* HKI 454. El operón completo *capABCD* involucrado en la síntesis de esta bacteriocina está presenten todos los miembros de grupo *Pesudomallei* y en *B. gladioli*, sin embargo, en *B. rhizoxinica* hay dos copias, uno similar al de *B. thailandensis* E264 y uno que difiere en las anotaciones para *capB* y *capC*. Este último *cluster* podría no ser funcional o no producir un péptido *lasso* debido a que el ORF predicho para *capA* codifica un péptido que carece del ácido aspártico necesario para la ciclización (Figura 8.7). Además de estos operones, en todas las cepas de *B. mallei* se encontró un *cluster* interrumpido por la inserción de una transposasa entré los genes *capC* y *capD*. De los 18 *clusters* de Capistruina encontrados en este trabajo solo ocho incluían la anotación para el gen de *capA*.

La búsqueda del operón para Capistruina nos muestra un claro ejemplo del potencial de buscar al mismo tiempo todas las secuencias involucradas en un proceso biológico, en este caso la producción de una bacteriocina. Si la búsqueda se ubiera dirigido sólo con la secuencia del precursor CapA se habrían encontrado únicamente ocho operones.

Por otro lado, dirigir la búsqueda sólo con alguna de las otras proteínas (CapB, CapC o CapD) habría hecho difícil encontrar todos los clusters, inclusive con un escrutínio del contexto genómico, ya que las proteínas CapB, CapC y CapD tienen porcentages de identidad diversos con sus homólogos de otras cepas bacterianas, de tal forma que los valores de corte

del programa BLAST descartarían algunos resultados.

El hecho de incluír a todas las secuencias del operón permitió realizar una búsqueda exahustiva en una sóla corrida y la estrategia de visualización nos permitió detectar los posibles *clusters* de genes, tanto los que estaban completos como aquellos que aparecían como *incompletos* en los resultados de BLAST.

Es notable el hecho de que las cepas que contienen el *cluster* de Capistruina también presentan el *cluster* de cinasas TOMM, el de bacteriocinas tipo lectina y el subtipo I de las Peptidasas C39. Además, las cepas que contienen a las Peptidasas C39 subtipo I también presentan el subtipo II pero en diferente cromosoma. Aunado a esto, todas las cepas de *B. pseudomallei* tienen un cluster de CDI y algunas de ellas tienen una bactericocina tipo *phage*. En promedio, las cepas de *B. pseudomallei* tienen 6.9 probables *clusters* de bacteriocinas por genoma.

Finalmente, 15 cepas de *Burkholderia* presentaron a la Linocina M18. Este péptido fue inicialmente reportado en *Brevibacterium linens* con actividad antagonista contra un amplio rango de cepas (75). Sin embargo, una caracterización estructural de un homólogo en *Thermotoga maritima* reveló la participación de esta proteína en procesos enzimáticos encapsulados y no se encontró actividad bactericida en la Linocina purificada (76). En este trabajo se detectó la presencia de Linocina en los genomas de *Burkholderia* porque hay secuencias de esta proteína en las bases de datos de bacteriocinas de BAGEL y BACTIBACSE que fueron usadas en este proyecto.

9 CONCLUSIONES

A lo largo de este trabajo se encontró que la mayoría (44.6 %) de las potenciales bacteriocinas presentes en el género *Burkholderia* fueron de tipo microcinas (Tabla 8.1) que incluyen los *clusters* de peptidasas C39, los *clusters* de las cinasas TOMM y a la Capistruina. Seguidas de las bacteriocnias tipo lectina (20.45 %), las CDI (18.2 %) y las bacteriocinas tipo *Phage* (10.0 %). Se encontró que hay una preferencia marcada de algunas bacteriocinas por un cromosoma en particular. Por ejemplo, los clusters de Capistruina y las bacteriocinas de tipo lectina están ubicados en el cromosoma 1 (excepto para las bacteriocinas tipo lectina de *B. glumae* BGR1 y *B. rhizoxinica* KHI-454 ubicadas en el cromosoma 2). Otro ejemplo claro es el de los *clusters* con peptidasas C39, subtipos I y II, que siempre estuvieron codificados en el cromosoma 2.

El estudio de genomas bacterianos usando herramientas bioinformáticas se está convirtiendo en una estrategia cada vez más productiva para entender las funciones microbianas y, en este sentido, la búsqueda de nuevos genes por *genome mining* nos ha aportado información valiosa. La estrategia más usada para buscar determinados genes en los nuevos genomas secuenciados es usar aquellos genes/proteínas más conservados como secuencias *query* y, una vez encontrados en los genomas, hacer un barrido del contexto genómico para encontrar el resto de los genes involucrados en el fenómeno de estudio. La caracterización de este contexto genómico normalmente involucra una comparación de secuencias por BLAST o un alineamiento de secuencias para determinar la función. Realizar una busqueda por BLAST de manera masiva permite incluir como secuencias *query* a todos los genes/proteínas que se sabe (o sospecha) que participan en algún proceso biológico, ahorrando mucho tiempo en la caracterización bioinformática de *clusters* de genes.

Una búsqueda masiva genera miles de resultados de BLAST que necesitan ser ordenados y filtrados antes de ser analizados. En este sentido, la representación de datos por *posición*

relativa nos brinda un espacio ordenado donde colocarlos, resaltando de manera automática si dos o más resultados se encuentran cerca formado *clusters* de genes.

Esta estrategia de visualización de genes permite tener de manera ordenada, interactiva e intuitiva los resultados de BLAST y es útil para manejar y analizar de manera simultanea una basta cantidad de información genómica, ya que es posible trabajar con cientos de genomas que incluyen miles de genes de manera muy sencilla y con bajos recursos computacionales. Con esta estrategia de visualización fue posible realizar una búsqueda masiva por BLAST de bacteriocinas en los genomas de *Burkholderia*, sin embargo, el desarrollo de la heramienta no se limita la búsqueda de bacteriocinas sino que puede ser usada para mapear cualquier proceso biológico, no sólo en *Burkholderia* sino en cualquier género bacteriano o incluso en todos los genomas bacterianos simultaneamente.

10 PERSPECTIVAS

La estrategia de visualización por *posición relativa* permite la visualización de una enorme cantidad de datos de manera simultánea y permite hacer estudios bioinformáticos que demandan el análisis tanto de pocas como de múltiples secuencias de manera simultánea. Por ahora esta herramienta está desarrollada para trabajar con genomas bacterianos completamente ensamblados. Las perspectivas de desarrollo de BLAST-XYplot Viewer están orientadas hacia la adaptación del código para trabajar con genomas parcialmente completos (*draft genomes*), es decir, que contienen practicamente todos los genes pertenecientes al genoma pero que no ha logrado circularizar.

Así mismo, pretendemos incluir la posibilidad de desplegar contexto genómico de los resultados de BLAST. Esta opción no es viable directamente sobre la Gráfica de despliegue de resultados de BLAST ya que incrementaría el número de datos proyectados sobre el gráfico, disminuyendo la capacidad de representar los resultados de BLAST. Sin embargo, sí es posible escribir código en javascript y html que permitan desplegar el contexto genómico de algún resultado en partícular en una ventana diferente.

Dentro de la arquitectura de BLAST-XYplot Viewer existe la posibilidad de seguir añadiendo código en PERL que permita filtrar los datos de manera específica, según las necesidades de cada proyecto, por ejemplo, la detección de dominios múltiples en las proteínas.

Otra oportunidad de desarrollo para BLAST-XYplot Viewer es ofrecer la posibilidad de que los usuarios pudieran usar la herramienta para analizar los genomas que han sido secuenciados durante la realización de sus propios proyectos (genomas que no están disponibles en las bases de datos públicas). Para ello es necesario desarrollar un sistema de *cuentas de usuario* para que la base de datos de genomas, así como los resultados de búsqueda, estén localizados en un sistema de directorios independientes de la herramienta principal.

10.1. Otras aplicaciones de la estrategia de representación de datos

Esta estrategia de visualización de genomas puede ser usada para otros propósitos, además de proyectar resultados de BLAST. Por ejemplo, una de las aplicaciones que se puede implementar es la representación de los cortes con enzimas de restricción sobre el replicón completo. Con esta herramienta se podrían visualizar todos los puntos de corte con diversas enzimas sobre muchos genomas de manera simultánea, permitiendonos detectar posibles sitios de inserción de genes/transposones/marcadores de manera masiva.

Por otro lado, una vez que se tengan filtrados los resultados de alguna búsqueda se pueden usar como atlas o bibliotecas de la presencia y distribución de operones o genes relacionados con algún proceso. Este atlas se podría ir actualizando cada vez que existan nuevos genomas bacterianos disponibles, de tal forma que la busqueda de genes se lleva a cabo sólo en los nuevos genomas y la información analizada y filtrada se añade al archivo original para poder visualizarla. De esta manera es posible tener bases de datos de presencia y distribución de rutas biosintéticas completas en genomas secuenciados y acceder a ellos de manera rápida e interactiva para visualizarlos/analizarlos.

11

APÉNDICES

11.1. Apéndice A

Listado de las especies pertenecientes al género *Burkholderia*.

.

1. *B. acidipaludis*
2. *B. alpina*
3. *B. ambifaria*
4. *B. andropogonis*
5. *B. anthina*
6. *B. arationis*
7. *B. arboris*
8. *B. arvi*
9. *B. aspalathi*
10. *B. bannensis*
11. *B. bryophila*
12. *B. caballeronis*
13. *B. caledonica*
14. *B. calidae*
15. *B. caribensis*
16. *B. caryophylli*
17. *B. catudaia*
18. *B. cenocepacia*
19. *B. cepacia*
20. *B. choica*
21. *B. cocovenenans*
22. *B. concitans*
23. *B. contaminans*
24. *B. cordobensis*
25. *B. denitrificans*
26. *B. diazotrophica*
27. *B. diffusa*
28. *B. dilworthii*
29. *B. dipogonis*
30. *B. dolosa*
31. *B. eburnea*
32. *B. endofungorum*
33. *B. ferrariae*
34. *B. fortuita*
35. *B. fungorum*
36. *B. ginsengisoli*
37. *B. ginsengiterrae*
38. *B. gladioli*
39. *B. glathei*
40. *B. glathei*
41. *B. glebae*
42. *B. glumae*
43. *B. graminis*
44. *B. grimmiae*
45. *B. heleia*
46. *B. hospita*
47. *B. humi*
48. *B. humisilvae*
49. *B. humptydooensis*
50. *B. hypogeia*
51. *B. insulsa*
52. *B. jiangsuensis*
53. *B. jirisanensis*
54. *B. kirstenboschensis*
55. *B. kururiensis*
56. *B. lata*
57. *B. latens*
58. *B. mallei*
59. *B. megalochromosomata*
60. *B. megapolitana*
61. *B. metallica*
62. *B. metalliresistens*
63. *B. mimosarum*
64. *B. monticola*
65. *B. multivorans*
66. *B. nodosa*
67. *B. norimbergensis*
68. *B. oklahomensis*
69. *B. oxyphila*
70. *B. panaciterrae*
71. *B. pedi*
72. *B. peredens*
73. *B. phenazinium*
74. *B. phenoliruptrix*
75. *B. pickettii*
76. *B. plantarii*
77. *B. pseudomallei*
78. *B. pseudomultivorans*
79. *B. ptereochthonis*
80. *B. puraquae*
81. *B. pyrrocinia*
82. *B. pyrrocinia*
83. *B. rhizosphaerae*
84. *B. rhynchosiae*
85. *B. sabiae*
86. *B. sacchari*
87. *B. sartisoli*
88. *B. sediminicola*
89. *B. seminalis*
90. *B. silvatlantica*
91. *B. singularis*
92. *B. solanacearum*
93. *B. soli*
94. *B. solisilvae*
95. *B. sordidicola*
96. *B. sprentiae*
97. *B. stabilis*
98. *B. stagnalis*
99. *B. susongensis*
100. *B. symbiotica*
101. *B. telluris*
102. *B. temeraria*
103. *B. terrae*
104. *B. terrestris*
105. *B. terricola*
106. *B. territorii*
107. *B. thailandensis*
108. *B. tropica*
109. *B. tuberum*
110. *B. turbans*
111. *B. ubonensis*
112. *B. udeis*
113. *B. unamae*
114. *B. vandii*
115. *B. vietnamiensis*
116. *B. zhejiangensis*

Listado de las especies con genoma secuenciado usadas en este trabajo que fueron transferidas al género *Paraburkholderia*.

1. *B. phymatum*
2. *B. phytofirmans*
3. *B. rhizoxinica*
4. *B. xenovorans*

11.2. Apéndice B

Conjunto de scripts de la herramienta de MIBlaST Versión 1.0.

Nombre del *script*	**Versión**	**Funciones**
MIBlaST	1.0	Leer el archivo de parámetros. Leer el archivo de secuencias. Asignación de fecha de trabajo. Apertura de directorio para escribir resultados. Mandar correr el script de MASTER.
Verif_Secuencias	1.7	Leer el archivo de secuencias. Verificar presencia de cabezal. Identificar cabezales repetidos. Verificar presencia de secuencia. Remover espacios en blanco. Remover números. Remover interlineados. Convertir a mayúsculas los caracteres. Identificar tipo de secuencia (ADN o proteínas).
MASTER	1.2	Leer el archivo de parámetros. Leer el archivo de secuencias revisadas. Direccionar los archivos al *script* necesario según los parámetros.
BLAST_Proteinas	2.5	Leer el archivo de parámetros. Leer el archivo de secuencias revisadas. Hacer búsqueda por sc blast según los parámetros. Organizar los resultados. Filtrar resultados por BitScore. Calcular un nuevo score de homología o proximidad. Escribir el archivo Json y el html para visualización.
BLAST_Genes	2.5	Leer el archivo de parámetros. Leer el archivo de secuencias revisadas. Hacer búsqueda por sc blast según los parámetros. Organizar los resultados. Filtrar resultados por BitScore. Calcular un nuevo score de homología o proximidad. Escribir el archivo Json y el html para visualización.
BLAST_Genoma	2.5	Leer el archivo de parámetros. Leer el archivo de secuencias revisadas. Hacer búsqueda por sc blast según los parámetros. Organizar los resultados. Filtrar resultados por BitScore. Calcular un nuevo score de homología o proximidad.

Continúa en la página siguiente

Continúa de la página anterior

Nombre del *script*	**Versión**	**Funciones**
		Escribir el archivo Json y el html para visualización.
BLAST_Intergen	2.5	Realiza una búsqueda de *string* usando expresiones regulares. Organizar los resultados. Filtrar resultados por BitScore. Calcular un nuevo score de homología o proximidad. Escribir el archivo Json y el html para visualización.
Genera_DataBase	1.1	Lee archivos en formato gbk de los genomas bacterianos. Extrae información sobre el genoma. Identifica los genes y escribe las secuencias de proteínas. Identifica los genes y escribe las secuencias de ADN. Identifica los genes y escribe las secuencias intergenicas. Escribe el archivo .ptt con las coordenadas "x,y" de posición relativa. Da el formato a las secuencias de Proteínas, ADN y genomas para poder realizar BLAST.

11.3. Apéndice C

Listado de replicones disponibles de *Burkholderia* utilizados en este estudio.

Especie	Cepa	UID	Archivo
B. ambifaria	AMMD	uid58303	NC_008385.gbk
B. ambifaria	AMMD	uid58303	NC_008390.gbk
B. ambifaria	AMMD	uid58303	NC_008391.gbk
B. ambifaria	AMMD	uid58303	NC_008392.gbk
B. ambifaria	MC40-6	uid58701	NC_010551.gbk
B. ambifaria	MC40-6	uid58701	NC_010552.gbk
B. ambifaria	MC40-6	uid58701	NC_010553.gbk
B. ambifaria	MC40-6	uid58701	NC_010557.gbk
B. sp	CCGE1001	uid42975	NC_015136.gbk
B. sp	CCGE1001	uid42975	NC_015137.gbk
B. sp	CCGE1002	uid42523	NC_014117.gbk
B. sp	CCGE1002	uid42523	NC_014118.gbk
B. sp	CCGE1002	uid42523	NC_014119.gbk
B. sp	CCGE1002	uid42523	NC_014120.gbk
B. sp	CCGE1003	uid46253	NC_014539.gbk
B. sp	CCGE1003	uid46253	NC_014540.gbk
B. cenocepacia	AU-1054	uid58371	NC_008060.gbk
B. cenocepacia	AU-1054	uid58371	NC_008061.gbk
B. cenocepacia	AU-1054	uid58371	NC_008062.gbk
B. cenocepacia	HI2424	uid58369	NC_008542.gbk
B. cenocepacia	HI2424	uid58369	NC_008543.gbk
B. cenocepacia	HI2424	uid58369	NC_008544.gbk
B. cenocepacia	HI2424	uid58369	NC_008545.gbk
B. cenocepacia	J2315	uid57953	NC_011000.gbk
B. cenocepacia	J2315	uid57953	NC_011001.gbk
B. cenocepacia	J2315	uid57953	NC_011002.gbk
B. cenocepacia	J2315	uid57953	NC_011003.gbk
B. cenocepacia	MC0-3	uid58769	NC_010508.gbk
B. cenocepacia	MC0-3	uid58769	NC_010512.gbk
B. cenocepacia	MC0-3	uid58769	NC_010515.gbk
B. cepacia	GG4	uid173858	NC_018513.gbk
B. cepacia	GG4	uid173858	NC_018514.gbk
B. gladioli	BSR3	uid66301	NC_015376.gbk
B. gladioli	BSR3	uid66301	NC_015377.gbk
B. gladioli	BSR3	uid66301	NC_015378.gbk
B. gladioli	BSR3	uid66301	NC_015381.gbk
B. gladioli	BSR3	uid66301	NC_015382.gbk
B. gladioli	BSR3	uid66301	NC_015383.gbk
B. glumae	BGR1	uid59397	NC_012718.gbk
B. glumae	BGR1	uid59397	NC_012720.gbk
B. glumae	BGR1	uid59397	NC_012721.gbk
B. glumae	BGR1	uid59397	NC_012723.gbk
B. glumae	BGR1	uid59397	NC_012724.gbk
B. glumae	BGR1	uid59397	NC_012725.gbk
B. sp.	KJ006	uid165871	NC_017920.gbk
B. sp.	KJ006	uid165871	NC_017921.gbk
B. sp.	KJ006	uid165871	NC_017922.gbk
B. sp.	KJ006	uid165871	NC_017923.gbk

Continúa en la página siguiente

Continúa de la página anterior

Especie	Cepa	UID	Archivo
B. lata		uid58073	NC_007509.gbk
B. lata		uid58073	NC_007510.gbk
B. lata		uid58073	NC_007511.gbk
B. mallei	ATCC 23344	uid57725	NC_006348.gbk
B. mallei	ATCC 23344	uid57725	NC_006349.gbk
B. mallei	NCTC 10229	uid58383	NC_008835.gbk
B. mallei	NCTC 10229	uid58383	NC_008836.gbk
B. mallei	NCTC 10247	uid58385	NC_009079.gbk
B. mallei	NCTC 10247	uid58385	NC_009080.gbk
B. mallei	SAVP1	uid58387	NC_008784.gbk
B. mallei	SAVP1	uid58387	NC_008785.gbk
B. multivorans	ATCC 17616	uid58697	NC_010070.gbk
B. multivorans	ATCC 17616	uid58697	NC_010084.gbk
B. multivorans	ATCC 17616	uid58697	NC_010086.gbk
B. multivorans	ATCC 17616	uid58697	NC_010087.gbk
B. multivorans	ATCC 17616	uid58909	NC_010801.gbk
B. multivorans	ATCC 17616	uid58909	NC_010802.gbk
B. multivorans	ATCC 17616	uid58909	NC_010804.gbk
B. multivorans	ATCC 17616	uid58909	NC_010805.gbk
B. phenoliruptrix	BR3459a	uid176370	NC_018672.gbk
B. phenoliruptrix	BR3459a	uid176370	NC_018695.gbk
B. phenoliruptrix	BR3459a	uid176370	NC_018696.gbk
B. phymatum	STM815	uid58699	NC_010622.gbk
B. phymatum	STM815	uid58699	NC_010623.gbk
B. phymatum	STM815	uid58699	NC_010625.gbk
B. phymatum	STM815	uid58699	NC_010627.gbk
B. phytofirmans	PsJN	uid58729	NC_010676.gbk
B. phytofirmans	PsJN	uid58729	NC_010679.gbk
B. phytofirmans	PsJN	uid58729	NC_010681.gbk
B. pseudomallei	1026b	uid162511	NC_017831.gbk
B. pseudomallei	1026b	uid162511	NC_017832.gbk
B. pseudomallei	1106a	uid58515	NC_009076.gbk
B. pseudomallei	1106a	uid58515	NC_009078.gbk
B. pseudomallei	1710b	uid58391	NC_007434.gbk
B. pseudomallei	1710b	uid58391	NC_007435.gbk
B. pseudomallei	668	uid58389	NC_009074.gbk
B. pseudomallei	668	uid58389	NC_009075.gbk
B. pseudomallei	BPC006	uid174460	NC_018527.gbk
B. pseudomallei	BPC006	uid174460	NC_018529.gbk
B. pseudomallei	K96243	uid57733	NC_006350.gbk
B. pseudomallei	K96243	uid57733	NC_006351.gbk
B. pseudomallei	MSHR305	uid213227	NC_021877.gbk
B. pseudomallei	MSHR305	uid213227	NC_021884.gbk
B. pseudomallei	MSHR346	uid55259	NC_012695.gbk
B. pseudomallei	NCTC 13179	uid226109	NC_022658.gbk
B. pseudomallei	NCTC 13179	uid226109	NC_022659.gbk
B. rhizoxinica	HKI 454	uid60487	NC_014718.gbk
B. rhizoxinica	HKI 454	uid60487	NC_014722.gbk
B. rhizoxinica	HKI 454	uid60487	NC_014723.gbk
B. sp	RPE64	uid205541	NC_021287.gbk
B. sp	RPE64	uid205541	NC_021288.gbk

Continúa en la página siguiente

Continúa de la página anterior

Especie	**Cepa**	**UID**	**Archivo**
B. sp	RPE64	uid205541	NC_021289.gbk
B. sp	RPE64	uid205541	NC_021294.gbk
B. sp	RPE64	uid205541	NC_021295.gbk
B. thailandensis	E264	uid58081	NC_007650.gbk
B. thailandensis	E264	uid58081	NC_007651.gbk
B. thailandensis	MSMB121	uid201037	NC_021173.gbk
B. thailandensis	MSMB121	uid201037	NC_021174.gbk
B. vietnamiensis	G4	uid58075	NC_009226.gbk
B. vietnamiensis	G4	uid58075	NC_009227.gbk
B. vietnamiensis	G4	uid58075	NC_009228.gbk
B. vietnamiensis	G4	uid58075	NC_009229.gbk
B. vietnamiensis	G4	uid58075	NC_009230.gbk
B. vietnamiensis	G4	uid58075	NC_009254.gbk
B. vietnamiensis	G4	uid58075	NC_009255.gbk
B. vietnamiensis	G4	uid58075	NC_009256.gbk
B. xenovorans	LB400	uid57823	NC_007951.gbk
B. xenovorans	LB400	uid57823	NC_007952.gbk
B. xenovorans	LB400	uid57823	NC_007953.gbk
B. sp.	YI23	uid81081	NC_016589.gbk
B. sp.	YI23	uid81081	NC_016590.gbk
B. sp.	YI23	uid81081	NC_016591.gbk
B. sp.	YI23	uid81081	NC_016592.gbk
B. sp.	YI23	uid81081	NC_016625.gbk
B. sp.	YI23	uid81081	NC_016626.gbk

11.4. Apéndice D

La herramienta BLAST-XYplot Viewer desarrollada en este proyecto fue aceptada para su publicación en la revista G3 Genes, Genomes, Genetics con el título: *BLAST-XYplot viewer: a tool for performing BLAST in whole-genome sequenced bacteria/archea and visualize whole results simultaneously* (77).

La estrategia de visualización masiva de datos y la herramienta BLAST-XYplot Viewer También fueron presentadas en los siguientes congresos:

International scientific conference on bacteriocins and antimicrobial peptides. Celebrado en Univercity city of Kosice, Slovakia en Mayo de 2013. Bajo la modalidad de conferencia plenaria con el título: *A new bioinformatic strategy to search for bacteriocins in sequenced bacterial genomes.*

IV Congreso de bioquímica y biología molecular de bacterias. Celebrado en Atlixco, Puebla en Octubre de 2015. Bajo la modalidad de conferencia plenaria con el título: *Development of a tool to perform massive BLAST search in whole-sequenced bacterial genomes.*

Fifth Meeting on Biochemistry and Molecular Biology of Bacteria. Celebrado en Chautla, Puebla en Octubre de 2017. Bajo la modalidad de conferencia plenaria con el título: *Tracing full biosynthetic pathways in whole-genome sequenced bacterial genomes with a single hit.*

3er simposio internacional de bioinformática. Celebrado en Cuernavaca, Morelos en Mayo de 2018. Bajo la modalidad de cartel con el título: *BLAST-XYplot Viewer: una herramienta de búsqueda de multi-secuencias en genomas bacterianos.*

SOFTWARE AND DATA RESOURCES

BLAST-XYPlot Viewer: A Tool for Performing BLAST in Whole-Genome Sequenced Bacteria/Archaea and Visualize Whole Results Simultaneously

Yagul Pedraza-Pérez,* Rodrigo Alberto Cuevas-Vede,† Ángel Bernardo Canto-Gómez,* Liliana López-Pliego,* Rosa María Gutiérrez-Ríos,‡ Ismael Hernández-Lucas,‡ Gustavo Rubín-Linares,† Ygnacio Martínez-Laguna,* Jesús Francisco López-Olguín,* and Luis Ernesto Fuentes-Ramírez*,[1]

*Instituto de Ciencias, and †Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, Puebla, Pue., México CP 72570, and ‡Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Mor., México CP 62210

ORCID IDs: 0000-0002-3873-4611 (Y.P.-P.); 0000-0002-0779-0987 (R.A.C.-V.); 0000-0001-6339-9491 (L.L.-P.); 0000-0002-0603-3810 (R.M.G.-R.); 0000-0002-8053-7742 (L.E.F.-R.)

ABSTRACT One of the most commonly used tools to compare protein or DNA sequences against databases is BLAST. We introduce a web tool that allows the performance of BLAST-searches of protein/DNA sequences in whole-genome sequenced bacteria/archaea, and displays a large amount of BLAST-results simultaneously. The circular bacterial replicons are projected as horizontal lines with fixed length of 360, representing the degrees of a circle. A coordinate system is created with length of the replicon along the *x*-axis and the number of replicon used on the *y*-axis. When a query sequence matches with a gene/protein of a particular replicon, the BLAST-results are depicted as an "*x,y*" position in a specially adapted plot. This tool allows the visualization of the results from the whole data to a particular gene/protein in real time with low computational resources.

KEYWORDS
large scale BLAST
bacterial genomes
operon search
BLAST-XYPlot viewer

Thousands of completely sequenced bacterial and archaeal genomes are currently available on public repositories and this number is increasing rapidly. This information allows the accomplishment of exhaustive comparative genomic studies. One of the most widely used tools for searching sequence similarity is BLAST (Altschul *et al.*, 1990), available to run from several web servers or locally with a stand-alone version. Running BLAST from a web server is limited to comparing a small number of query sequences at the same time. When running a BLAST local-version it is possible to include as many query sequences as desired. Nevertheless, additional programming skills are required to extract information.

doi: https://doi.org/10.1534/g3.118.200220
Manuscript received March 6, 2018; accepted for publication May 9, 2018; published Early Online May 22, 2018.

Supplemental material available at Figshare: https://doi.org/10.25387/g3.6288914.
[1]Corresponding author: luis.fuentes@correo.buap.mx

Genomic information is a powerful source for getting insight into microbial traits and functions. Considering the large quantity of inherent data, its study depends on bioinformatics tools. Genomic mining is used to find the genes or clusters of genes that code for a specific biological function in sequenced genomes. One of the most successful approaches to search for these genes is the use, as a query sequence, of either the most conserved gene/protein or of a representative one with known function. That result is then used as a starting point for scanning its genomic context to find the rest of the genes involved in the feature of study. That genomic context characterization usually involves BLAST or sequence alignment comparisons to forecast their functions. Nevertheless, that strategy is time consuming. Alternatively, performing a large-scale BLAST search allows the inclusion, as query, of multiple nucleotide or amino acid sequences of those genes/proteins. It saves time for bioinformatic characterization, but also generates many BLAST results that need to be analyzed.

Many biological processes are encoded in gene clusters, so multiple searches are required to determine the presence of a full biosynthetic pathway or bacterial operon. Tools to perform these searches in a single run, and to sort and display results in an easy way to analyze them are needed as well. Currently, several tools are available and quite useful to search single and multiple genes/proteins by sequence similarity (Fong

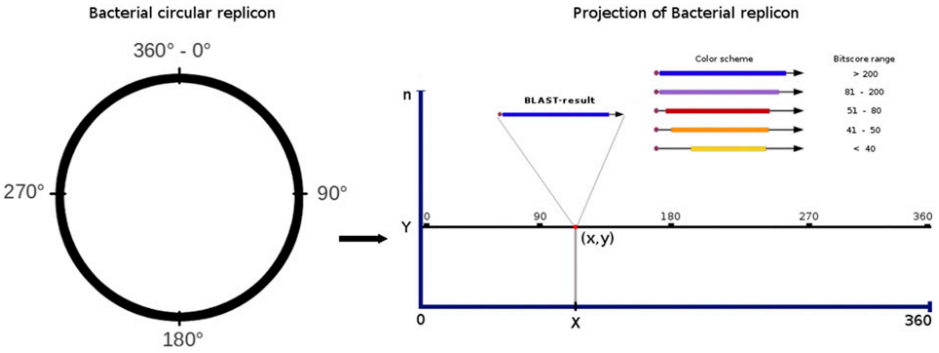

Figure 1 Linear projection of the circular bacterial replicon. The bacterial replicon is represented as a line with fixed length of 360 arbitrary units. An (*x,y*) plot scheme is used to depict BLAST-results where the *x*-axis represents the position of subject sequence into the bacterial replicon (with any value ranging from 0 to 360), and the *y*-axis corresponds to the number of replicons used in the search. See text for description.

et al. 2008; Revanna *et al.* 2009; Despalins *et al.* 2011; Medema *et al.* 2013; Neumann *et al.* 2014). These tools include flexible configuration options, but only display a limited number of data. Therefore, they require extra steps to fully browse through the results. Hence, the number of results that can be displayed simultaneously is limited.

In this study we introduced a platform-independent, free and open to use web tool (available at http://www.blast-xyplot-viewer.icuap.buap.mx) that allows multiple BLAST searches against whole-sequenced bacterial/archaea genomes, and a novel strategy for visualizing vast data. To display an extensive number of BLAST-results simultaneously, we used the advantages of an (*x,y*) plot (Figure 1). This tool can be used to search for the presence, completeness and distribution of single genes/proteins, operons or full biosynthetic pathways in a particular taxon or biological hierarchy, or even in all sequenced bacteria/archaea, in a single run.

MATERIALS AND METHODS

The visualization of massive BLAST-results uses an (*x,y*) dot plot scheme. To perform this representation, the circular bacterial replicons were projected as straight lines with fixed length equal to 360 (the degrees in a circle), providing a delimited space where any chromosome or plasmid can be represented independently of their actual size. In an (*x,y*) plot coordinate system, the *x*-axis represents the length of the replicons, and the *y*-axis represents the quantity of replicons (chromosomes or plasmids) used as database in the search. That number ranges from 1 (if only one replicon is used) to the maximum number of replicons used (*e.g.*, the chromosomes and plasmids of all sequenced bacteria/archaea). Using this projection, each BLAST-result is mapped by their *relative position* (0-360) instead of their real (nucleotide) position (Figure 1).

When searching for a particular gene/protein in a genome(s) with BLAST, each result (subject sequence) is represented by a dot with an (*x,y*) coordinate. In addition to the dot, representing the origin of the subject gene, a vector with a magnitude proportional to its length, and with a direction according to the transcription sense is plotted. The BLAST-output (alignment between query and subject sequences) is represented by a thicker line over the vector, with a longitude proportional to the length of the region aligned and a color that depends on the significance defined by the BitScore value given by BLAST. That color varies from dark blue for the most significant to yellow for the least (Figure 1). The advantages of this visualization method are: 1) the interactivity of the plot, like zooming or dragging, can be used to visualize/analyze vast data from the whole results and focusing into a particular one in real time; 2) it is easily noticed if two or more BLAST-results are close to each other; and 3) BLAST-results that

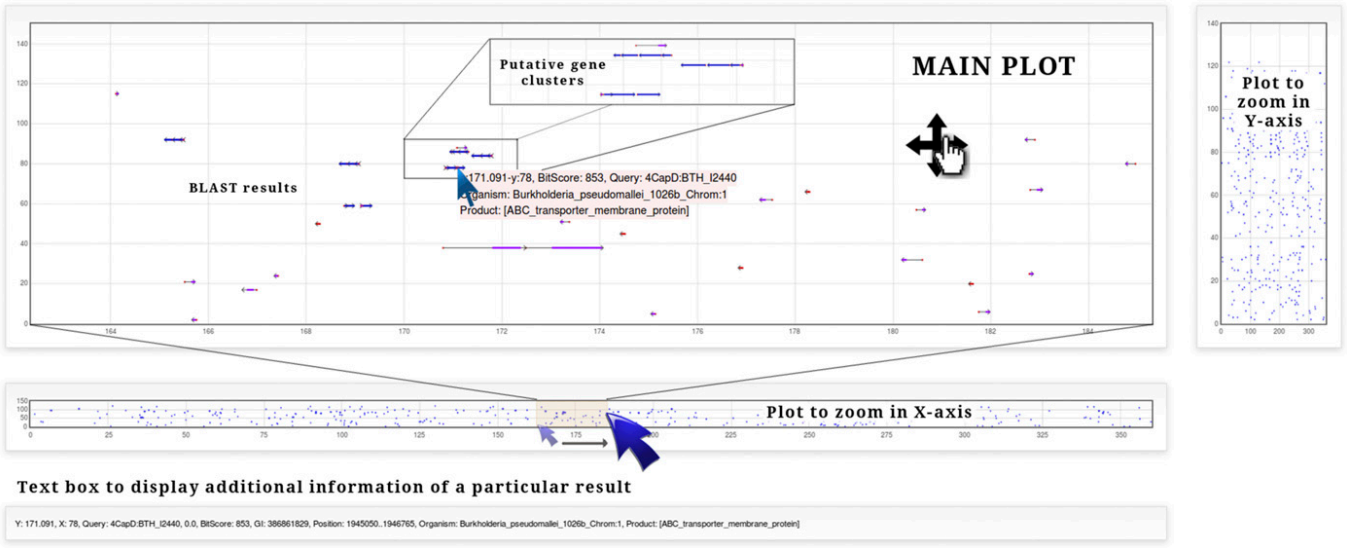

Figure 2 Screenshots of the plot used to display BLAST-results. Three plots and a text box are used to visualize data. The main plot allows zoom in/out, drag and displays information about the subject sequence on mouseover of BLAST-results. The two additional plots can be used to select a subrange of data in the *x*-axis (bottom plot) or in the *y*-axis (right plot). The text box displays additional information when clicking a particular BLAST-result.

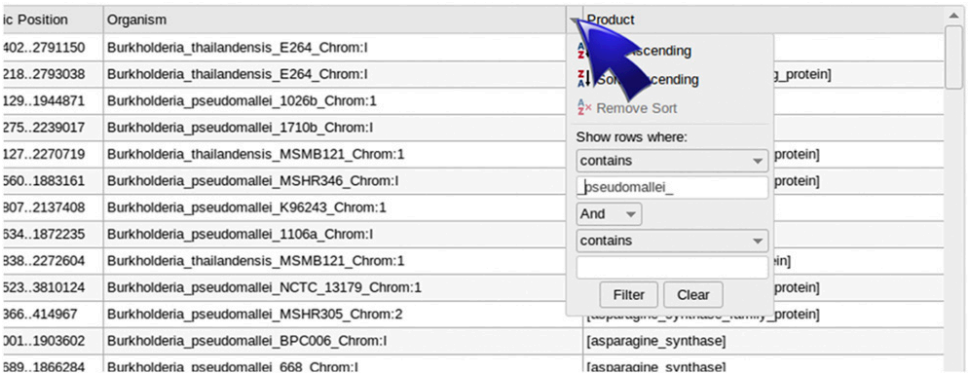

Figure 3 Table of data. The table lists BLAST-results in a spreadsheet format, and includes basic sort and filter options.

belong to many different replicons can be schematized simultaneously (Figure 2).

The input of BLAST-XYplot viewer is a set of sequences (or a single one) in FASTA format, including headers, that can be pasted in a text box or loaded as a file from the user's computer, in plain text format. Users must type a Job title, provide a valid e-mail address and configure several parameters. BLAST searches can be performed against protein, genes or whole-genome bacterial/archaeal databases, depending of configuration setup. The BLAST Sequence comparison could be performed against genomes belonging to a particular genus (*e.g.*, *Escherichia*, *Enterococcus*, *Archaeoglobus*), a taxonomic group (*e.g.*, *Enterobacteriaceae*, *Proteobacteria*, *Euryarchaeota*) or all genomes. Currently, the genome list used by BLAST-XYplot viewer includes 5138 replicons, downloaded from the NCBI ftp server. Depending on the number of query sequences, databases used and jobs queue, the sequence comparison could take anywhere from a few minutes (search for less than twenty query sequences in a few genomes) to several hours (hundreds of query sequences in many genomes).

BLAST-results are recorded in a plain-text file in Json format and loaded into a specially adapted FLOT chart (http://www.flotcharts.org) to be visualized. In addition to the plot, a table of data in spreadsheet style allows users to sort and filter results (see http://www.blast-xyplot-viewer.icuap.buap.mx/tutorial for a graphical tutorial). The maximum number of data displayed on the web tool is limited to 50 thousand results because more than this quantity saturates the plot, hindering analysis. Nevertheless, the raw data file available for download could include up to one million BLAST-results.

The *Graph* results page consist of three plots and a text box (Figure 2). The main plot is used to visualize the BLAST results; it can be dragged and zoomed in or out. On mouseover of a particular result, a pop up text will appear with basic information about the result: relative position, BitScore value, query header, organism and gene product. By clicking on a particular result, additional information will appear on the text box below the plot. The two small plots allow users to zoom in on a particular region: on the *x*-axis, to visualize a section of all replicons; or in the *y*-axis to visualize whole replicons of a subgroup of microorganisms. To browse through results, a zoom could be applied using the small plot to select a range of around 25 degrees on the *x*-axis and drag the main plot with the left mouse button.

The *Table* page shows the BLAST-results in a basic spreadsheet. On mouseover of the header of a column, a small triangle will appear. By left clicking, a drop-down menu will be displayed to configure filtering or sorting options (Figure 3). After filter BLAST-results, data can be updated, in order to be visualized on the plot, by clicking on the "Actualize data" button bellow the table, and then "refresh" the *Graph* page.

The data can be downloaded from the *Graph* page as a Job-folder containing files with raw data (*_Sorted), filtered results (*_Sorted_Scored), *Table* and *Graph* in html format and their respective Json files. The raw file is a plain text and would be used for deeper scrutiny of the BLAST-result in the user's computer.

The main advantages of this tool, when compared with other similar tools (Table 1), are the possibility to visualize several thousand BLAST-results simultaneously, highlighting *relative position* into the bacterial genome, and the interactivity of a plot that allows focusing on a particular result in real time using low computational resources. Additionally, this tool is capable of scanning many genomes simultaneously when dragging the plot through a zoomed portion of the *x*-axis, a feature not supported by any other tool.

Data availability

The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article. Supplemental material available at Figshare: https://doi.org/10.25387/g3.6288914.

RESULTS AND DISCUSSION

Case studies that can be performed easily with this tool include search for a single gene/protein, a small group of genes/proteins (small

Table 1 Comparison of XY-plot viewer with similar tools

	Large scale search	Precalculated results	Web Server	Display vast results simultaneously	Displays gene order	Real time zoom	Displays relative position in the genome	Displays gene context	Filters Data	Custom databases	Ref.
XYplot viewer	X		X	X	X	X	X		X	X	this work
PSAT	X	X	X		X		X	X		X	Fong *et al.* 2008
MultiGeneBlast	X				X			X		X	Mederma *et al.* 2013
BLASTgrabber	X			X					X	X	Neumann *et al.* 2014
Absynte			X		X			X		X	Despalins *et al.* 2011

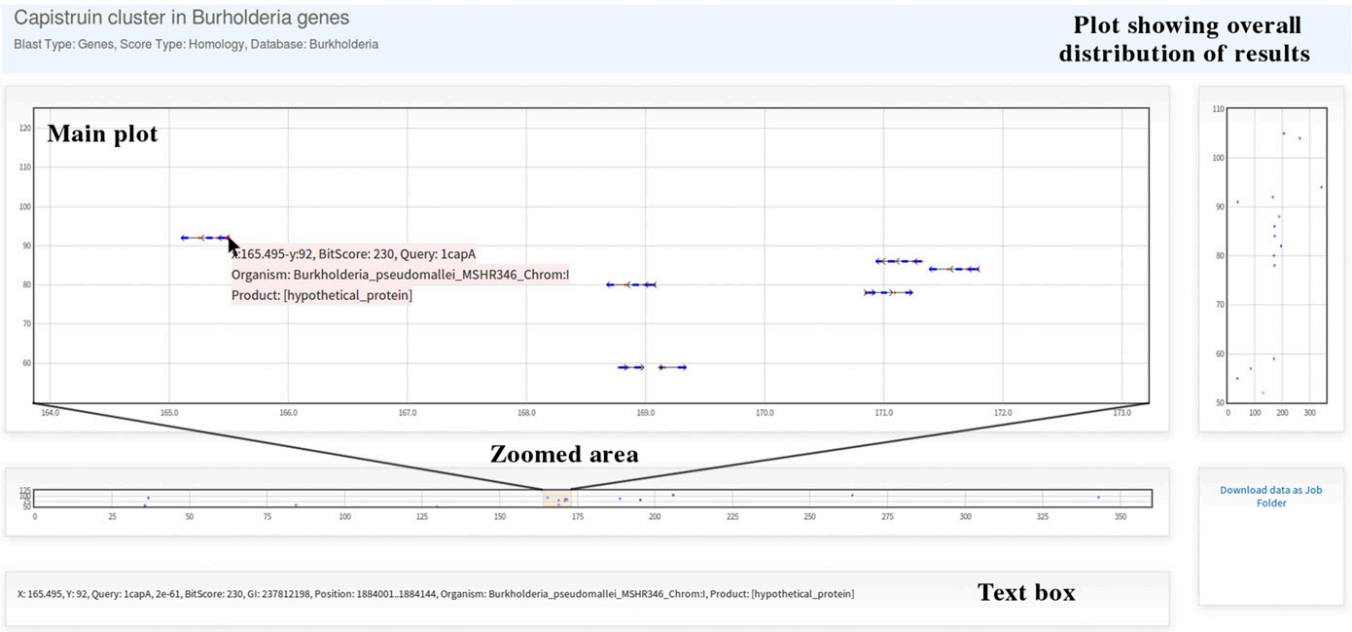

Figure 4 Plot of BLAST-results for Capistruin gene cluster search. The plot shows overall distribution of BLAST-results. By zooming data on the *x*-axis with the bottom plot, it is possible to see several clusters, and get information of a particular result when clicking on it.

operon), a large operon or a full biosynthetic/degradative pathway, or multiple operons. All of these can be compared against a few, many or all of the sequenced genomes in the database. Some examples are depicted here (for an extended description see supplementary material).

One example of a small group of genes is the cluster that codes for Capistruin, a bacteriocin reported in *Burkholderia thailandensis* E264 and characterized by a knotted structure (Knappe *et al.* 2008). The Capistruin locus is composed by four genes (*capABCD*) that code for the bacteriocin precursor, two proteins involved in maturation and an exporter (Knappe *et al.* 2008). To search for this gene cluster in the remaining genomes of the *Burkholderia* genus, parameters can be defined as follows:

Job title: Search for Capistruin cluster in Burkholderia genomes
Sequence File: Capistruin.sec
Subject Data Base: Genus
Genus: Burkholderia
BlastType: Genes
E-value: 1e-08
MaxResults: 30000

A header for each sequences must be specified and an informative name is recommended, like those shown below:

>capA
ATGGTTCGACTTTTGGCGAAGC...
>capB
ATGCAACGGTCGCGCTATTTTC...
>capC
ATGGCGAAATCTATCGAACGCC...
>capD
ATGGCCCTTCCCATCCGAAACG...

The links to the *Graph* and *Table* pages are shown when the job is finished. If an e-mail address is provided, these links will be sent to it. In the *Graph* web page, the main plot shows the overall distribution of the BLAST-results. By selecting a region in the small plot, a range of data can be visualized more closely allowing the visualization of genes clustering (Figure 4).

An example of a multiple operon case is the search of the six types of secretion systems. Here we used 80 sequences of representative secretion systems (Abby *et al.* 2016) and ran a BLAST search in all genomes. To browse through the huge amount of results generated by large jobs like this example, it may be more convenient to filter them by some criteria with the *Table*. To facilitate this process, the informative names of headers become relevant. The most useful way to name the sequence header could be: >SSI_HlyD, >SSI_TolC. . ., >SSII_GspC, >SSII_GspE. . ., and so on. By using this systematic nomenclature, it should be easy to filter results belonging to a specific type of secretion system. For example, to show only the results of the secretion system type I, we can type "SSI_" on the "contains" option of the drop down filter menu of the *Query* column in the *Table*, and then actualize results to be plotted in the *Graph*. Another option to filter results is to write the name of a desired organism in the drop down menu or a combination of both filter conditions. For example, to see the distribution of secretion systems in *Burkholderia*, data can be filtered by typing "Burkholderia_" in the *Organism* column and clicking on the "Actualize data" button to update the plot. When browsing the results it would be possible to find different kinds of gene clusters (Figure 5). Similarly, by filtering data by "SSII_" in the *Query* column, and by "gladioli_" in the *Organism* column the table will show only the type II secretion system in *B. gladioli*, and then can be displayed in the plot to see the distribution.

An example of multiple biosynthetic pathway searches may include all kinds of bacteriocin gene clusters. Bacteriocins are proteinaceous compounds synthesized by bacteria that inhibit the growth of other microorganisms. They are diverse in sequence, structure, mechanisms of action, and genetic regulation (Riley and Wertz 2002). This diversity makes laborious to find them by classical sequence similarity analysis. A strategy that has been used for finding them is to search for bacteriocin-related and conserved genes in a first step, and then browse the genomic context in order to find the non-conserved ones (Lee *et al.* 2008; Begley *et al.* 2009; Marsh *et al.* 2010; Wang *et al.* 2011; Singh and Sareen 2014). As an alternative of that strategy, we downloaded over 2700 sequences of bacteriocins and related proteins (*e.g.*, those related to maturation or exportation) from

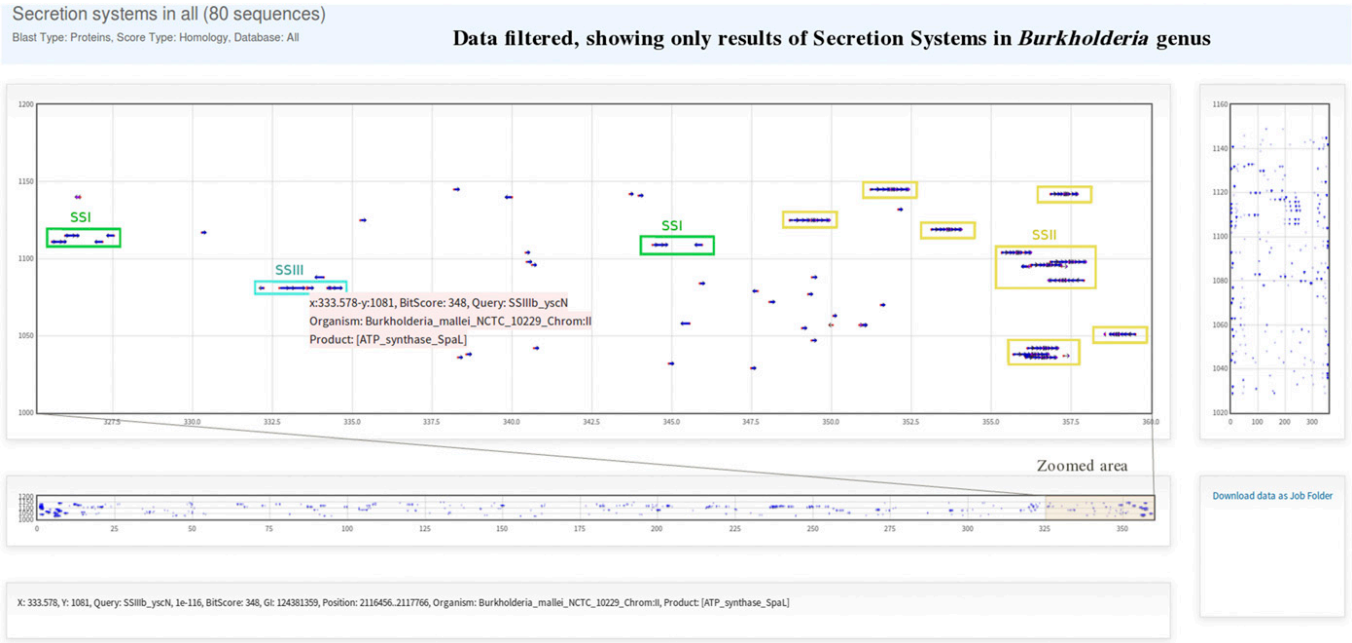

Figure 5 Plot of BLAST-results for Secretion Systems gene cluster search. The plot shows a section of the distribution of BLAST in a zoomed area from the degrees 325 to 360.

databases (Bru *et al.* 2005; Nikolskaya *et al.* 2006; Rawlings *et al.* 2006; Saier *et al.* 2006; Hammami *et al.* 2010; Marchler-Bauer *et al.* 2010; van Heel *et al.* 2013), and performed a large scale BLAST-search of bacteriocins in the *Burkholderia* genus. The BLAST-results plotted in the *Graph* highlights different putative bacteriocin gene cluster (Figure 6).

Conclusion

The study of bacterial genomes using bioinformatic tools is becoming a more common strategy to understand microbial functions. Particularly, genome mining is used to find specific genes or clusters of genes in new sequenced genomes. The genomic characterization usually involves BLAST or sequence alignment comparison of genes/proteins to forecast their function. Performing a large scale BLAST search allows to include multiple query sequences of all genes/proteins involved in a particular biological function, in order to trace them in a single hit. This new strategy of data visualization by *relative position* provides a space where the enormous data generated by large-scale BLAST searches can be placed. This approach projects the data in an ordinate, interactive and intuitive format that is useful to handle and analyze vast genomic information simultaneously.

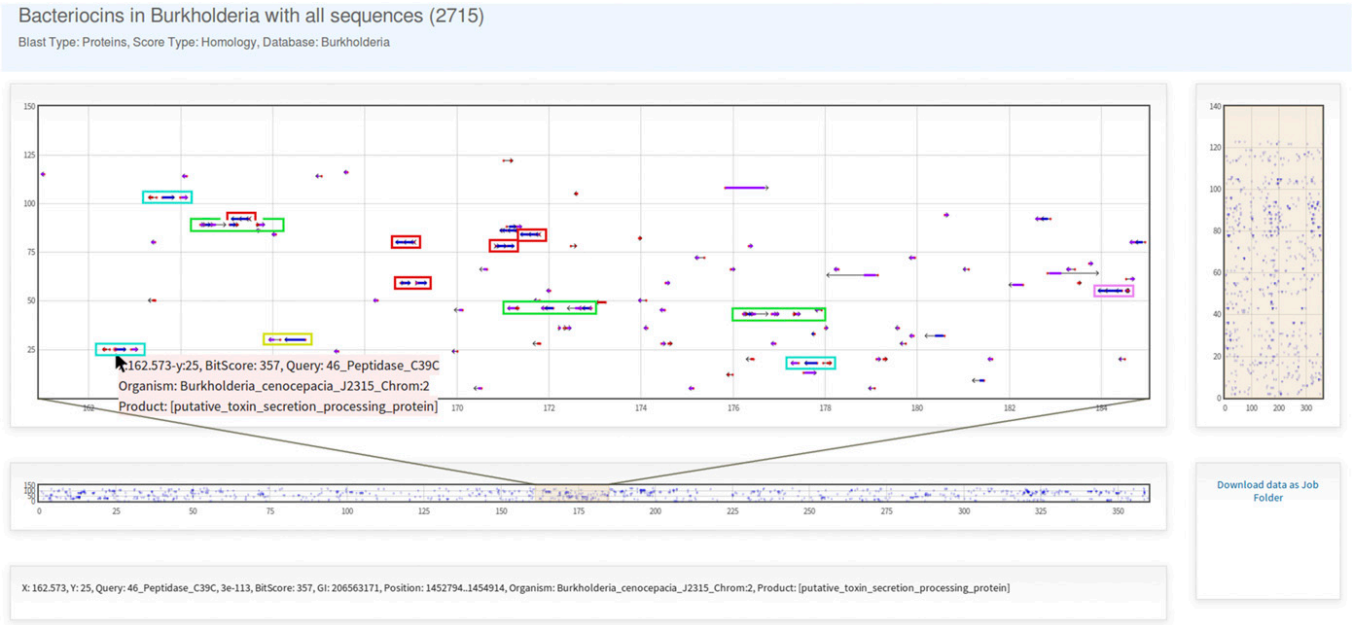

Figure 6 Plot of BLAST-results for the search of bacteriocins in *Burkholderia*. BLAST-results for more than 2700 query sequences are plotted together, highlighting different kind of putative bacteriocin gene cluster in the *Burkholderia* genus.

ACKNOWLEDGMENTS

This work was partially funded by CONACYT CB-2009-128235-Z. Y. Pedraza scholarship was supported by VIEP-BUAP. We are grateful for English edition to Joseph Bradley (English Teaching Assistant, Fulbright García-Robles). We acknowledge Winter Genomics (www.wintergenomics.com) for help in debugging code.

LITERATURE CITED

Abby, S. S., J. Cury, J. Guglielmini, B. Néron, M. Touchon *et al.*, 2016 Identification of protein secretion systems in bacterial genomes. Sci. Rep. 6: 23080. https://doi.org/10.1038/srep23080

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Begley, M., P. D. Cotter, C. Hill, and R. P. Ross, 2009 Identification of a novel two-peptide lantibiotic, lichenicidin, following rational genome mining for LanM proteins. Appl. Environ. Microbiol. 75: 5451–5460. https://doi.org/10.1128/AEM.00730-09

Bru, C., E. Courcelle, S. Carrère, Y. Beausse, S. Dalmar *et al.*, 2005 The ProDom database of protein domain families: more emphasis on 3d. Nucleic Acids Res. 33: D212–D215. https://doi.org/10.1093/nar/gki034

Despalins, A., S. Marsit, and J. Oberto, 2011 Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. Bioinformatics 27: 2905–2906. https://doi.org/10.1093/bioinformatics/btr473

Fong, C., L. Rohmer, M. Radey, M. Wasnick, and M. J. Brittnacher, 2008 PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes. BMC Bioinformatics 9: 170. https://doi.org/10.1186/1471-2105-9-170

Hammami, R., A. Zouhir, C. Le Lay, J. B. Hamida, and I. Fliss, 2010 BACTIBASE second release: a database and tool platform for bacteriocin characterization. BMC Microbiol. 10: 22. https://doi.org/10.1186/1471-2180-10-22

Knappe, T. A., U. Linne, S. Zirah, S. Rebuffat, X. Xie *et al.*, 2008 Isolation and structural characterization of capistruin, a lasso peptide predicted from the genome sequence of *Burkholderia thailandensis* E264. J. Am. Chem. Soc. 130: 11446–11454. https://doi.org/10.1021/ja802966g

Lee, S. W., D. A. Mitchell, A. L. Markley, M. E. Hensler, D. Gonzalez *et al.*, 2008 Discovery of a widely distributed toxin biosynthetic gene cluster. Proc. Natl. Acad. Sci. USA 105: 5879–5884. https://doi.org/10.1073/pnas.0801338105

Marchler-Bauer, A., S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire *et al.*, 2010 CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Res. 39: D225–D229. https://doi.org/10.1093/nar/gkq1189

Marsh, A. J., O. O'Sullivan, R. P. Ross, P. D. Cotter, and C. Hill, 2010 In silico analysis highlights the frequency and diversity of type 1 lantibiotic gene clusters in genome sequenced bacteria. BMC Genomics 11: 679. https://doi.org/10.1186/1471-2164-11-679

Medema, M. H., E. Takano, and R. Breitling, 2013 Detecting sequence homology at the gene cluster level with MultiGeneBlast. Mol. Biol. Evol. 30: 1218–1223. https://doi.org/10.1093/molbev/mst025

Neumann, R. S., S. Kumar, T. H. A. Haverkamp, and K. Shalchian-Tabrizi, 2014 BlastGrabber: a bioinformatic tool for visualization, analysis and sequence selection of massive blast data. BMC Bioinformatics 15: 128. https://doi.org/10.1186/1471-2105-15-128

Nikolskaya, A. N., C. N. Arighi, H. Huang, W. C. Barker, and C. H. Wu, 2006 PIRSF family classification system for protein functional and evolutionary analysis. Evol. Bioinform. Online 2: 197.

Rawlings, N. D., F. R. Morton, and A. J. Barrett, 2006 MEROPS: the peptidase database. Nucleic Acids Res. 34: D270–D272. https://doi.org/10.1093/nar/gkh071

Revanna, K. V., V. Krishnakumar, and Q. Dong, 2009 A web-based software system for dynamic gene cluster comparison across multiple genomes. Bioinformatics 25: 956–957. https://doi.org/10.1093/bioinformatics/btp078

Riley, M. A., and J. E. Wertz, 2002 Bacteriocins: evolution, ecology, and application. Annu. Rev. Microbiol. 56: 117–137. https://doi.org/10.1146/annurev.micro.56.012302.161024

Saier, M. H., Jr, C. V. Tran, and R. D. Barabote, 2006 TCDB: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Res. 34: D181–D186. https://doi.org/10.1093/nar/gkj001

Singh, M., and D. Sareen, 2014 Novel LanT associated lantibiotic clusters identified by genome database mining. PLoS One 9: e91352. https://doi.org/10.1371/journal.pone.0091352

van Heel, A. J., A. de Jong, M. Montalban-Lopez, J. Kok, and O. P. Kuipers, 2013 BAGEl3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Res. 41: W448–W453. https://doi.org/10.1093/nar/gkt391

Wang, H., D. P. Fewer, and K. Sivonen, 2011 Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. PLoS One 6: e22384. https://doi.org/10.1371/journal.pone.0022384

Communicating editor: M. Cherry

REFERENCIAS

[1] Gratia A. Sur un remarquable exemple d'antagonisme entre deux souches de coilbacille. Comptes Rendus Des Seances De La Societe De Biologie Et De Ses Filiales. 1925;93:1040–1041.

[2] Riley MA, Gordon DM. The ecological role of bacteriocins in bacterial competition. Trends in Microbiology. 1999;7(3):129–133.

[3] Drider D, Rebuffat S. Prokaryotic antimicrobial peptides: from genes to applications. Springer Science & Business Media; 2011.

[4] Lancaster LE, Wintermeyer W, Rodnina MV. Colicins and their potential in cancer treatment. Blood Cells, Molecules, and Diseases. 2007;38(1):15–18.

[5] Settanni L, Corsetti A. Application of bacteriocins in vegetable food biopreservation. International Journal of Food Microbiology. 2008;121(2):123–138.

[6] Yang SC, Lin CH, Sung CT, Fang JY. Antibacterial activities of bacteriocins: application in foods and pharmaceuticals. Frontiers in Microbiology. 2014;5:241.

[7] López-Cuellar MdR, Rodríguez-Hernández AI, Chavarría-Hernández N. LAB bacteriocin applications in the last decade. Biotechnology & Biotechnological Equipment. 2016;30(6):1039–1050.

[8] Johnson EM, Jung DYG, Jin DYY, Jayabalan DR, Yang DSH, Suh JW. Bacteriocins as food preservatives: Challenges and emerging horizons. Critical Reviews in Food Science and Nutrition. 2017;p. 1–25.

[9] Heng NC, Wescombe PA, Burton JP, Jack RW, Tagg JR. The diversity of bacteriocins in Gram-positive bacteria. In: Bacteriocins. Springer; 2007. p. 45–92.

[10] Gordon DM, Oliver E, Littlefield-Wyer J. The diversity of bacteriocins in Gram-negative bacteria. In: Bacteriocins. Springer; 2007. p. 5–18.

[11] O'connor E, Shand R. Halocins and sulfolobicins: the emerging story of archaeal protein and peptide antibiotics. Journal of Industrial Microbiology and Biotechnology. 2002;28(1):23–31.

[12] Hancock RE, Scott MG. The role of antimicrobial peptides in animal defenses. Proceedings of the National Academy of Sciences. 2000;97(16):8856–8861.

[13] Alvarez-Sieiro P, Montalbán-López M, Mu D, Kuipers OP. Bacteriocins of lactic acid bacteria: extending the family. Applied Microbiology and Biotechnology. 2016;100(7):2939–2951.

[14] Meindl K, Schmiederer T, Schneider K, Reicke A, Butz D, Keller S, et al. Labyrinthopeptins: a new class of carbacyclic lantibiotics. Angewandte Chemie International Edition. 2010;49(6):1151–1154.

[15] Rea MC, Sit CS, Clayton E, O'Connor PM, Whittal RM, Zheng J, et al. Thuricin CD, a posttranslationally modified bacteriocin with a narrow spectrum of activity against *Clostridium difficile*. Proceedings of the National Academy of Sciences. 2010;107(20):9352–9357.

[16] Booker SJ, Grove TL. Mechanistic and functional versatility of radical SAM enzymes. F1000 Biology Reports. 2010;2.

[17] Cascales E, Buchanan SK, Duché D, Kleanthous C, Lloubes R, Postle K, et al. Colicin biology. Microbiology and Molecular Biology Reviews. 2007;71(1):158–229.

[18] Gillor O, Vriezen JA, Riley MA. The role of SOS boxes in enteric bacteriocin regulation. Microbiology. 2008;154(6):1783–1792.

[19] Bonsor DA, Hecht O, Vankemmelbeke M, Sharma A, Krachler AM, Housden NG, et al. Allosteric β-propeller signalling in TolB and its manipulation by translocating colicins. The European Molecular Biology Organization journal. 2009;28(18):2846–2857.

[20] Chauleau M, Mora L, Serba J, de Zamaroczy M. FtsH-dependent processing of RNase colicins D and E3 means that only the cytotoxic domains are imported into the cytoplasm. Journal of Biological Chemistry. 2011;286(33):29397–29407.

[21] El Ghachi M, Bouhss A, Barreteau H, Touzé T, Auger G, Blanot D, et al. Colicin M exerts its bacteriolytic effect via enzymatic degradation of undecaprenyl phosphate-linked peptidoglycan precursors. Journal of Biological Chemistry. 2006;281(32):22761–22772.

[22] Barreteau H, El Ghachi M, Barnéoud-Arnoulet A, Sacco E, Touzé T, Duché D, et al. Characterization of colicin M and its orthologs targeting bacterial cell wall peptidoglycan biosynthesis. Microbial Drug Resistance. 2012;18(3):222–229.

[23] Clarke DJ, Campopiano DJ. Maturation of McjA precursor peptide into active microcin MccJ25. Organic & Biomolecular Chemistry. 2007;5(16):2564–2566.

[24] Knappe TA, Linne U, Zirah S, Rebuffat S, Xie X, Marahiel MA. Isolation and structural characterization of capistruin, a *lasso* peptide predicted from the genome sequence of *Burkholderia thailandensis* E264. Journal of the American Chemical Society. 2008;130(34):11446–11454.

[25] Yorgey P, Davagnino J, Kolter R. The maturation pathway of microcin B17, a peptide inhibitor of DNA gyrase. Molecular Microbiology. 1993;9(4):897–905.

[26] Allali N, Afif H, Couturier M, Van Melderen L. The highly conserved TldD and TldE proteins of *Escherichia coli* are involved in microcin B17 processing and in CcdA degradation. Journal of Bacteriology. 2002;184(12):3224–3231.

[27] Severinov K, Semenova E, Kazakov A, Kazakov T, Gelfand MS. Low-molecular-weight post-translationally modified microcins. Molecular Microbiology. 2007;65(6):1380–1394.

[28] Metlitskaya A, Kazakov T, Kommer A, Pavlova O, Praetorius-Ibba M, Ibba M, et al. Aspartyl-tRNA synthetase is the target of peptide nucleotide antibiotic Microcin C. Journal of Biological Chemistry. 2006;281(26):18033–18042.

[29] Kazakov T, Vondenhoff GH, Datsenko KA, Novikova M, Metlitskaya A, Wanner BL, et al. Escherichia coli peptidase A, B, or N can process translation inhibitor microcin C. Journal of Bacteriology. 2008;190(7):2607–2610.

[30] Lagos R, Villanueva JE, Monasterio O. Identification and properties of the genes encoding microcin E492 and its immunity protein. Journal of Bacteriology. 1999;181(1):212–217.

[31] Michel-Briand Y, Baysse C. The pyocins of *Pseudomonas aeruginosa*. Biochimie. 2002;84(5-6):499–510.

[32] Nakayama K, Takashima K, Ishihara H, Shinomiya T, Kageyama M, Kanaya S, et al. The R-type pyocin of *Pseudomonas aeruginosa* is related to P2 phage, and the F-type is related to *lambda* phage. Molecular Microbiology. 2000;38(2):213–231.

[33] Barre A, Bourne Y, Van Damme EJ, Peumans WJ, Rougé P. Mannose-binding plant lectins: different structural scaffolds for a common sugar-recognition process. Biochimie. 2001;83(7):645–651.

[34] Ghequire MG, Loris R, e De Mot R. MMBL proteins: from lectin to bacteriocin. Biochemical Society Transactions. 2012;40(part 6):1553–1559.

[35] Ghequire MG, Garcia-Pino A, Lebbe EK, Spaepen S, Loris R, De Mot R. Structural determinants for activity and specificity of the bacterial toxin LlpA. PLoS Pathogens. 2013;9(2):e1003199.

[36] Tagg J. Bacterial BLIS. ASM News. 1991;57(611).

[37] Jack RW, Tagg JR, Ray B. Bacteriocins of Gram-positive bacteria. Microbiological Reviews. 1995;59(2):171–200.

[38] Li Y, Xiang H, Liu J, Zhou M, Tan H. Purification and biological characterization of halocin C8, a novel peptide antibiotic from *Halobacterium* strain AS7092. Extremophiles. 2003;7(5):401–407.

[39] Naor A, Yair Y, Gophna U. A halocin-H4 mutant *Haloferax mediterranei* strain retains the ability to inhibit growth of other halophilic archaea. Extremophiles. 2013;17(6):973–979.

[40] Hammami R, Zouhir A, Le Lay C, Hamida JB, Fliss I. BACTIBASE second release: a database and tool platform for bacteriocin characterization. BMC Microbiology. 2010;10(1):22.

[41] van Heel AJ, de Jong A, Montalban-Lopez M, Kok J, Kuipers OP. BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Research. 2013;41(W1):W448–W453.

[42] Dirix G, Monsieurs P, Marchal K, Vanderleyden J, Michiels J. Screening genomes of Gram-positive bacteria for double-glycine-motif-containing peptides. Microbiology. 2004;150(5):1121–1126.

[43] Dirix G, Monsieurs P, Dombrecht B, Daniels R, Marchal K, Vanderleyden J, et al. Peptide signal molecules and bacteriocins in Gram-negative bacteria: a genome-wide

in silico screening for peptides containing a double-glycine leader sequence and their cognate transporters. Peptides. 2004;25(9):1425–1440.

[44] Begley M, Cotter PD, Hill C, Ross RP. Identification of a novel two-peptide lantibiotic, lichenicidin, following rational genome mining for LanM proteins. Applied and Environmental Microbiology. 2009;75(17):5451–5460.

[45] Marsh AJ, O'Sullivan O, Ross RP, Cotter PD, Hill C. In silico analysis highlights the frequency and diversity of type 1 lantibiotic gene clusters in genome sequenced bacteria. BMC Genomics. 2010;11(1):679.

[46] Nicolas GG. Detection of putative new mutacins by bioinformatic analysis using available web tools. BioData Mining. 2011;4(1):22.

[47] Wang H, Fewer DP, Sivonen K. Genome mining demonstrates the widespread occurrence of gene clusters encoding bacteriocins in cyanobacteria. PloS One. 2011;6(7):e22384.

[48] Singh M, Sareen D. Novel LanT associated lantibiotic clusters identified by genome database mining. PLoS One. 2014;9(3):e91352.

[49] Depoorter E, Bull MJ, Peeters C, Coenye T, Vandamme P, Mahenthiralingam E. *Burkholderia*: an update on taxonomy and biotechnological potential as antibiotic producers. Applied Microbiology and Biotechnology. 2016;100(12):5215–5229.

[50] Woods DE, Sokol PA. The genus *Burkholderia*. In: The Prokaryotes. Springer; 2006. p. 848–860.

[51] Marshall K, Shakya S, Greenhill AR, Padilla G, Baker A, Warner JM. Antibiosis of *Burkholderia ubonensis* against *Burkholderia pseudomallei*, the causative agent for melioidosis. Southeast Asian Journal of Tropical Medicine and Public Health. 2010;41(4):904.

[52] Bakkal S, Robinson SM, Ordonez CL, Waltz DA, Riley MA. Role of bacteriocins in mediating interactions of bacterial isolates taken from cystic fibrosis patients. Microbiology. 2010;156(7):2058–2067.

[53] Marín-Cevada V, Muñoz-Rojas J, Caballero-Mellado J, Mascarúa-Esparza M, Castañeda-Lucio M, Carreño-López R, et al. Antagonistic interactions among bacteria inhabiting pineapple. Applied Soil Ecology. 2012;61:230–235.

[54] Rojas-Rojas FU, Salazar-Gómez A, Vargas-Díaz ME, Vásquez-Murrieta MS, Hirsch AM, De Mot R, et al. Broad-spectrum antimicrobial activity by Burkholderia cenocepacia TAtl-371, a strain isolated from the tomato rhizosphere. Microbiology. 2018;.

[55] Anderson MS, Garcia EC, Cotter PA. The *Burkholderia bcpAIOB* genes define unique classes of two-partner secretion and contact dependent growth inhibition systems. PLoS Genetics. 2012;8(8):e1002877.

[56] Nikolakakis K, Amber S, Wilbur JS, Diner EJ, Aoki SK, Poole SJ, et al. The toxin/immunity network of *Burkholderia pseudomallei* contact-dependent growth inhibition (CDI) systems. Molecular Microbiology. 2012;84(3):516–529.

[57] Ghequire MG, Canck E, Wattiau P, Winge I, Loris R, Coenye T, et al. Antibacterial activity of a lectin-like *Burkholderia cenocepacia* protein. Microbiology Open. 2013;2(4):566–575.

[58] Ghequire MG, De Mot R. Distinct colicin M-like bacteriocin-immunity pairs in *Burkholderia*. Scientific Reports. 2015;5:17368.

[59] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215(3):403–410.

[60] Fong C, Rohmer L, Radey M, Wasnick M, Brittnacher MJ. PSAT: a web tool to compare genomic neighborhoods of multiple prokaryotic genomes. BMC bioinformatics. 2008;9(1):170.

[61] Revanna KV, Krishnakumar V, Dong Q. A web-based software system for dynamic gene cluster comparison across multiple genomes. Bioinformatics. 2009;25(7):956–957.

[62] Despalins A, Marsit S, Oberto J. Absynte: a web tool to analyze the evolution of orthologous archaeal and bacterial gene clusters. Bioinformatics. 2011;27(20):2905–2906.

[63] Medema MH, Takano E, Breitling R. Detecting sequence homology at the gene cluster level with MultiGeneBlast. Molecular Biology and Evolution. 2013;30(5):1218–1223.

[64] Moorhouse M, Barry P. Bioinformatics biocomputing and Perl: an introduction to bioinformatics computing skills and practice. John Wiley & Sons; 2005.

[65] Wall L, et al.. The Perl programming language. Prentice Hall Software Series; 1994.

[66] Flanagan D. JavaScript: the definitive guide. OŔeilly Media, Inc.; 2006.

[67] Krause J. HTML: Hypertext Markup Language. In: Introducing Web Development. Springer; 2016. p. 39–63.

[68] Bru C, Courcelle E, Carrère S, Beausse Y, Dalmar S, Kahn D. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Research. 2005;33(suppl_1):D212–D215.

[69] Marchler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Research. 2010;39(suppl_1):D225–D229.

[70] Rawlings ND, Morton FR, Barrett AJ. MEROPS: the peptidase database. Nucleic Acids Research. 2006;34(suppl_1):D270–D272.

[71] Wu CH, Nikolskaya A, Huang H, Yeh LSL, Natale DA, Vinayaka C, et al. PIRSF: family classification system at the Protein Information Resource. Nucleic Acids Research. 2004;32(suppl_1):D112–D114.

[72] Saier Jr MH, Tran CV, Barabote RD. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Research. 2006;34(suppl_1):D181–D186.

[73] Dunbar KL, Melby JO, Mitchell DA. YcaO domains use ATP to activate amide backbones during peptide cyclodehydrations. Nature Chemical Biology. 2012;8(6):569.

[74] Rosengren KJ, Clark RJ, Daly NL, Göransson U, Jones A, Craik DJ. Microcin J25 has a threaded sidechain-to-backbone ring structure and not a head-to-tail cyclized backbone. Journal of the American Chemical Society. 2003;125(41):12464–12474.

[75] Valdés-Stauber N, Scherer S. Isolation and characterization of Linocin M18, a bacteriocin produced by *Brevibacterium linens*. Applied and Environmental Microbiology. 1994;60(10):3809–3814.

[76] Sutter M, Boehringer D, Gutmann S, Günther S, Prangishvili D, Loessner MJ, et al. Structural basis of enzyme encapsulation into a bacterial nanocompartment. Nature Structural and Molecular Biology. 2008;15(9):939.

[77] Pedraza-Pérez Y, Cuevas-Vede RA, Canto-Gómez ÁB, López-Pliego L, Gutiérrez-Ríos RM, Hernández-Lucas I, et al. BLAST-XYPlot Viewer: A tool for performing BLAST in whole-genome sequenced bacteria/archaea and visualize whole results simultaneously. G3:Genes, Genomes, Genetics. 2018;8(7):2167–2172.