UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

Minería Genómica de NRPS en Bacterias: Sus Mecanismos Evolutivos y Relaciones Filogenéticas.

TESIS

QUE PARA OPTAR POR EL GRADO DE:
Maestro en Ciencias

PRESENTA:
Biol. Andrés De Sandozequi Mijares

TUTOR PRINCIPAL
Dr. Lorenzo Patrick Segovia Forcella
Instituto de Biotecnología, UNAM

MIEMBROS DEL COMITÉ TUTOR
Dra. María del Refugio Trejo
Centro de Investigación en Biotecnología, UAEM
Dr. Enrique Merino Pérez
Instituto de Biotecnología, UNAM

Cuernavaca, México. Agosto, 2018

A mis padres,

Son mi más grande inspiración.

Este proyecto de Maestría se realizo en el Laboratorio de Diseño y Evolución de Proteínas del Instituto de Biotecnología de la UNAM, bajo la tutela del Dr. Lorenzo Segovia.

Agradezco a los apoyos económicos que permitieron la realización de este proyecto. Los estudios de maestría se realizaron con la beca CONACYT 596818.

Agradecimientos Personales

Al Dr. Lorenzo, por darme la oportunidad de realizar este proyecto y enseñarme tanto.

A Martín, por ser mi maestro siempre y nunca perder la confianza en mí.

A mis padres, por apoyarme tanto en todos mis estudios.

A todos en el laboratorio, por hacer mi estancia de lo más amena y tener la paciencia para enseñarme.

A la Dra. María del Refugio y al Dr. Enrique Merino por sus recomendaciones a lo largo del proyecto.

A la Dra. Claudia, por apoyarme para continuar.

A Daniela y a Jonathan, por su gran ayuda y confianza.

Resumen

Las Sintetasas de Péptidos No Ribosomales (NRPS) son responsables de sintetizar una gran parte de los antibióticos conocidos y otros productos naturales importantes: incluyendo la penicilina, la vancomicina y el lipopéptido surfactina producido por cepas de *Bacillus subtilis*, que es un potente biosurfactante. Los NRPS están compuestos de múltiples módulos funcionales con tres dominios básicos: un dominio de condensación (C); un dominio de adenización (A); y un dominio de portador de péptidos (T). Cada módulo cataliza la incorporación del siguiente aminoácido a la cadena peptídica en crecimiento, funcionando como una línea de ensamblaje de pequeños péptidos. Estas megasintetasas pueden usar aminoácidos comunes, aminoácidos no proteinógenicos y decorar el péptido usando dominios enzimáticos adicionales. En este estudio, utilizamos modelos ocultos de Markov para explorar 8,756 Proteomas de referencia de Bacterias de UniProt, identificando 5,687 proteínas NRPS. Posteriormente, analizamos sus relaciones con herramientas bioinformáticas para descubrir los mecanismos evolutivos que impulsan su variabilidad. Descubrimos que diversos mecanismos pueden estar involucrados en la evolución de NRPS. De este modo, las bacterias pueden explorar nuevas estructuras de Péptidos No Ribosomales y sus diferentes bioactividades. Estos mecanismos podrían ocurrir más a menudo de lo que se pensaba, pero los arreglos multimodulares complejos son escasos, y las arquitecturas modulares similares pueden ser polifiléticas. También encontramos grupos de similitud de secuencias altamente conservados que se correlacionan con la especificidad del sustrato del módulo que podría ayudar a los esfuerzos de bioingeniería. Además, presentamos evidencia de que estos genes podrían haberse diseminado a través de los taxa bacterianos a partir de antiguas cianobacterias por varios eventos de transferencia horizontal de genes y radiación adaptativa posterior dentro del taxón receptivo.

Palabras clave:

NRPS - Péptidos no ribosomales - Minería de genomas - Evolución - Arquitectura de proteínas - Predicción de sustrato - Herramientas bioinformáticas - Filogenómica - HGT

Tabla de contenidos

1. Introducción

La búsqueda e investigación de los productos naturales producidos por microorganismos ha sido uno de los mayores focos de interés en la biotecnología actual, debido a su amplio espectro de actividades biológicas, y usos en la industria farmacéutica y biotecnológica, siendo utilizados como antibióticos, enzimas, vitaminas, medicamentos, biosurfactantes, entre otros (1, 2). De entre estos destacan los antibióticos, que con la creciente aparición de cepas resistentes a diferentes fármacos de última generación, se le ha dado una prioridad en los esfuerzos mundiales de investigación, así como nuevos tipos de fármacos que combatan el cáncer (3).

Una característica de los microorganismos es que son ubicuos, por lo que presentan estilos de vida y metabolismos únicos que les permiten sobrevivir en cualquier ecosistema del planeta y consecuentemente producir nuevos tipos de enzimas y metabolitos. Asimismo, los microorganismos son uno de los mayores recursos para el descubrimiento de nuevas clases de productos naturales.

Recientemente, ha crecido el interés en una clase de productos naturales producidos por bacterias y hongos, llamados Péptidos No Ribosomales (NRP por sus siglas en inglés), que son sintetizados por grandes complejos enzimáticos multifuncionales llamados Sintetasas de Péptidos No Ribosomales (NRPS por sus siglas en inglés), que utilizan reacciones regioespecíficas y estereoespecíficas complejas para montar estructural y funcionalmente diversos péptidos que tienen importantes aplicaciones biotecnológicas (4). Las NRPS son enzimas multimodulares que reconocen, activan, modifican y enlazan los aminoácidos intermediarios al péptido producto de forma secuencial. La biosíntesis de los NRP se produce a través de la función de la unidad catalítica, o módulo, el cual es responsable de añadir un nuevo aminoácido a la cadena peptídica en crecimiento (5). Varios módulos forman una NRPS, en la que el orden de los módulos de la o las NRPS involucradas generalmente es colineal a la secuencia del péptido producto (Figura 1). La principal diferencia entre péptidos sintetizados por el ribosoma, y los NRP, es que aparte de integrar aminoácidos convencionales a su vez pueden utilizar aminoácidos no

proteinogénicos como la Ornitina o el Di-aminobutirato para generar péptidos con estructuras más variadas (6, 7). En este estudio se analizan las relaciones estructurales y filogenéticas entre NRPS bacterianos, con lo cual se infieren los mecanismos por los cuales han surgido y evolucionado.

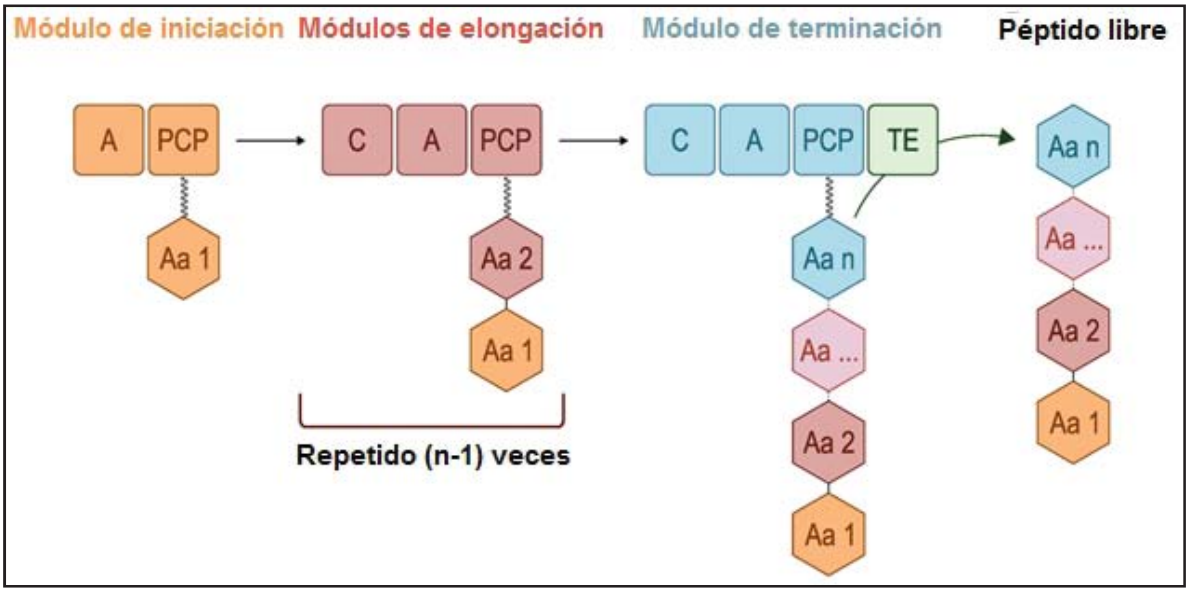

Figura 1. Las NRPS funcionan como una línea de ensamblaje de péptidos en donde cada módulo añade un nuevo aminoácido a la cadena de elongación del péptido (adaptado de referencia 8).

1.1 Péptidos no Ribosomales

Los NRP han sido ampliamente estudiados en los últimos 30 años y se han descubierto una gran variedad de usos biotecnológicos como antibióticos, antifúngicos, antitumorales, inmunosupresores, control de plagas en plantas, e incluso como biosurfactantes industriales, cosméticos y su uso en la biorremediación (9). Los NRP tienen una inmensa variabilidad estructural variando entre 2 hasta 18 aminoácidos (Tabla 1): pueden ser péptidos comunes, como la Nostopeptolida producida por una especie del género *Nostoc*, que es un péptido de 9 aminoácidos, este péptido tiene la función de regulación de la diferenciación celular y de la quimiotaxis (10); otro tipo de NRP, son los glucopéptidos, que por modificaciones decorativas contienen aminoácidos glucosilados, un ejemplo con importancia médica es el antibiótico de último recurso Vancomicina, que se usa para tratar cepas de bacterias gram positivas (ej. MRSA) resistentes a antibióticos comunes

(11); otro tipo de NRP son los lipopéptidos, que presentan una cadena lipídica de cadena variable comúnmente ligada al primer aminoácido de la secuencia de esamblaje. por ejemplo, la Surfactina sintetizada por cepas de *Bacillus* (12), la cual consiste de una cadena cíclica de 7 residuos de aminoácidos y una cadena de alquilos de 8 a 16 carbonos dependiendo de la cepa productora. La mayoría de los lipopéptidos descritos hasta la fecha los producen cepas de los géneros de *Bacillus, Pseudomonas* y *Streptomyces*, los cuales se han caracterizado como potentes agentes bactericidas, fungicidas e insecticidas.

Muchas de estas cepas se han descrito como agentes controladores de enfermedades en plantas (13). Otro ejemplo importante es el uso de la Daptomicina como antibiótico de último recurso para tratar cepas de *Staphylococcus aureus* resistentes a Meticilina (14). Ha crecido el interés en los lipopéptidos como biomoléculas con actividades farmacológicas (Tabla 1), debido a la amplia gama de estructuras (Figura 2) pueden presentar actividades variadas desde antitumorales, antimicóticas, propiedades inmunosupresoras, antibacterianas, etc. (15). Estos compuestos son anfifílicos, teniendo una parte hidrófila, así como una hidrofóbica. Esto les permite reducir las fuerzas repulsivas entre fases disímiles, lo que provoca que las fases puedan interactuar y mezclarse con mayor facilidad, lo que los clasifica como moléculas biosurfactantes. La efectividad de un surfactante se mide por la energía necesaria por unidad de área para traer una molécula de la fase masiva a la superficie. Los surfactantes efectivos reducen la tensión superficial entre el agua y el aire de 72 a 35 mN/m y la tensión interfacial entre el agua y n-hexadecano de 40 a 1 mN/m (16). Los lipopéptidos presentan, además de sus propiedades tensoactivas, actividades farmacológicas como por ejemplo antimicrobiana, antitumoral y propiedades inmunosupresoras (14, 17). También se ha encontrado que pueden tener otras utilidades, como por ejemplo, como emulsificadores, detergentes, insecticidas, para el control de plagas en plantas, dispersión y solubilización de compuestos hidrofóbicos (1), etc.

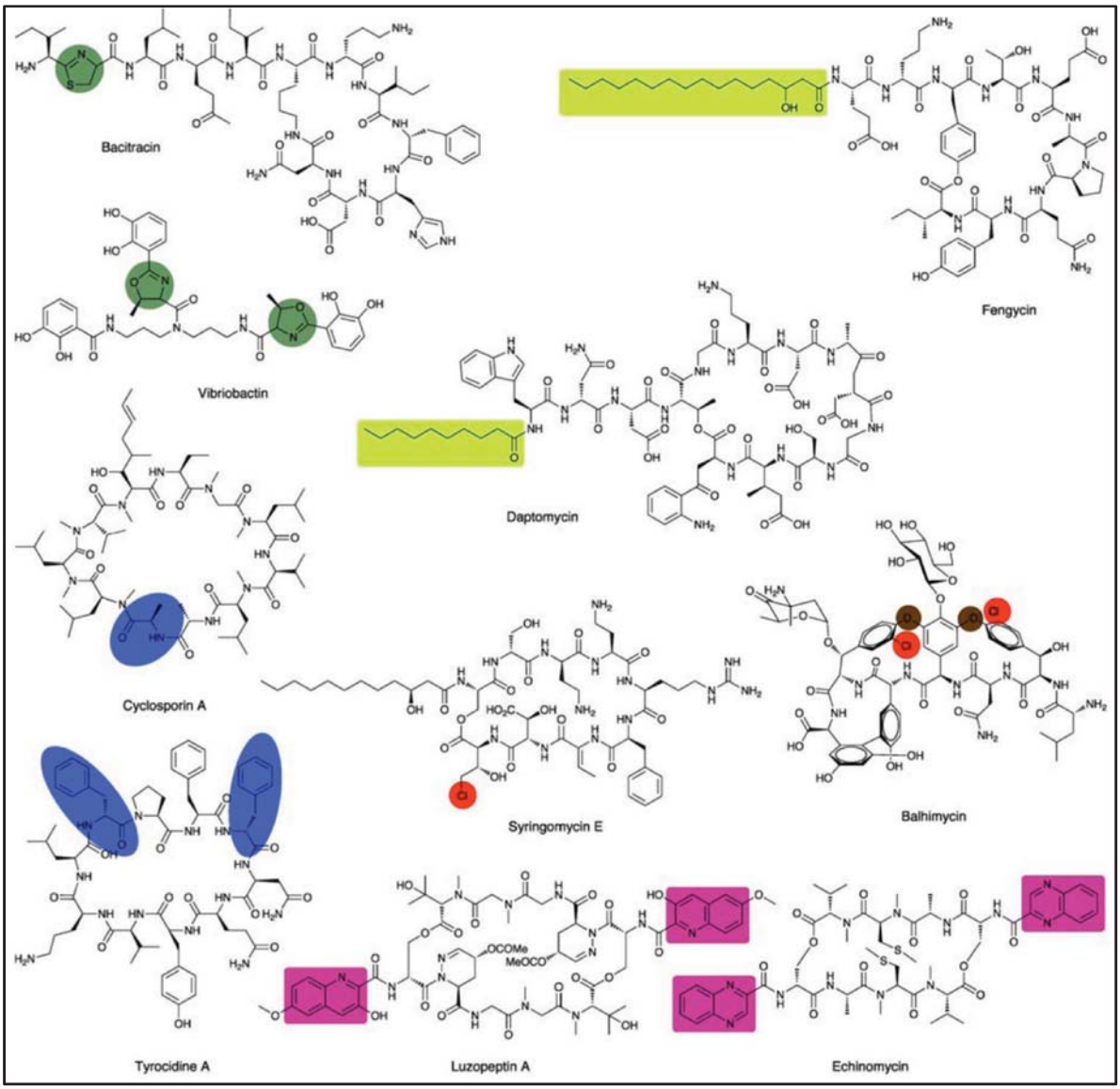

Figura 2. Las estructuras dentro de los NRPs que les confieren su bioactividad. Verde obscuro, anillo aromático heterocíclico (oxazol). Azul, residuos de aminoácidos convencionales. Verde claro, cadena alifática. Rojo, residuos de cloro u otras halogenaciones. Morado, quinolinas con uno o dos heteroátomos (adaptado de referencia 13).

La cadena lipídica generalmente es lo que les confiere su bioactividad, como es el caso de la Daptomicina o la Fengicina (Figura 2, recuadro verde claro), aunque otras estructuras como anillos aromáticos y decoraciones especiales como la metilación o halogenación también juegan un papel importante.

Tabla 1. Péptidos no ribosomales con importancia biotecnólogica.

Nombre	Especie	Tipo	Función	Cadena	Referencias
A54145	*Streptomyces fradiae*	Lipopéptido cíclico parcial	Antibiótico.	FA,Trp,D-Glu,OH-Asn,Thr,NMe-Gly,Ala,Asp,D-Lys,OMe-Asp,Gly,D-Asn,Glu,Ile	(18–20)
Arthrofactina	*Pseudomonas sp.* MIS38	Lipopéptido cíclico	Biosurfactante, Antibiótico.	FA, D-Leu,D-Asp,D-Thr,D-Leu,D-Leu,D-Ser,Leu,D-Ser,Ile,Ile,Asp	(21–24)
Bacitracina	*Bacillus licheniformis*	Péptido cíclico parcial	Antibiótico.	Ile,Cys,Leu,D-Glu,Ile,Lys,D-Orn,Ile,D-Phe,His,D-Asp,Asn	(25, 26)
CDA 4b	*Streptomyces coelicolor*	Lipopéptido cíclico parcial	Antibiótico dependiente de Calcio.	FA,Ser,Thr,D-Trp,Asp,Asp,D-Hpg,Asp,Gly,D-OH-Asn,3Me-Glu,Trp	(27–30)
Daptomicina	*Streptomyces filamentosus, Streptomyces lividans*	Lipopéptido cíclico parcial	Antibiótico.	FA,Trp,D-Asn,Asp,Thr,Gly,Orn,Asp,D-Ala,Asp,Gly,D-Ser,3Me-Glu,Kyn	(31–34)
Enduracidina	*Streptomyces fungicidicus*	Glucolipopéptido cíclico	Antibiótico.	FA,Asp,Thr,D-Hpg,D-Orn,D-aThr,Hpg,D-Hpg,aThr,Cit,D-End,Hpg,D-Ser,Cl_2-Hpg,Gly,End,D-Ala,Hpg	(35, 36)
Fengycina	*Bacillus species*	Lipopéptido cíclico	Antibiótico, Antifúngico.	FA,Glu,D-Orn,Tyr,D-aThr,Glu,D-Ala,Pro,Gln,D-Tyr,Ile	(37–39)
Fusaricidina	*Paenibacillus polymyxa*	Lipopéptido cíclico	Antibiótico, Antifúngico.	Guanidinil-FA,Thr,D-Val,Val,Thr,D-Asn,D-Ala	(40–42)
Gramicidina	*Brevibacillus brevis*	Péptido	Antibiótico para MRSA.	Val,Orn,Leu,D-Phe,Pro,Val,Orn,Leu,D-Phe,Pro	(43–45)
Lichenisina	*Bacillus licheniformis*	Lipopéptido cíclico	Antibiótico, Antifúngico, Biosurfactante.	FA,Gln,Leu,D-Leu,Val,Asp,D-Leu,Ile	(46–48)
Massetolida	*Pseudomonas fluorescens*	Lipopéptido cíclico parcial	Antibiótico, control de oomycteos patógenos de plantas.	FA,Leu,D-Glu,D-aThr,D-alle,Leu,D-Ser,Leu,D-Ser,Ile	(49–51)
Nostopeptolida	*Nostoc sp. GSV224*	Péptido cíclico	Diferenciación celular, quimioatrayente.	Ile,Ser,Me-Pro,Leu,Leu,Gly,Asn,Tyr,Pro	(10, 52)
Orfamida	*Pseudomonas fluorescens*	Lipopéptido cíclico parcial	Biosurfactante, lisis de zoosporas de oomycetos, biocontrol de *Rhizoctonia,* insecticida.	FA,Leu,Leu,Leu,D-Glu,D-aThr,D-alle,D-Ser,Leu,D-Ser,Val	(53)
Plipastatina	*Paenibacillus elgii*	Lipopéptido cíclico	Antibiótico.	FA,Glu,D-Orn,Tyr,D-allo-Thr,Glu,D-Ala,Pro,Gln,D-Tyr,Ile	(54, 55)
Polymixina	*Parnibacillus polymyxa*	Lipopéptido cíclico parcial	Antibiótico para gram negativas.	FA,Dab,Thr,Dab,Dab,Dab,D-Phe,Leu,Dab,Dab,Thr	(56, 57)
Pyoverdina	*Especies de Pseudomonas*	Sideróforos	Ingestión de Hierro.	Ser,Arg,Ser,FoOHOrn,[Lys,FoOHOrn,Thr,Thr]	(58–60)
Serrawettina	*Serratia marscecens*	Lipopéptido cíclico	Biosurfactante, Antibiótico, Antitumoral, control de plagas en plantas.	FA,Ser,FA,Ser	(61–64)
Surfactina	*Bacillus subtilis*	Lipopéptido cíclico	Biosurfactante, Antibiótico, Antifúngico, Antitumoral.	FA,Glu,Leu,D-Leu,Val,Asp,D-Leu,Leu	(12, 37, 65–67)
Syringomicina	*Pseudomonas syringae pv. Syringae*	Lipopéptido cíclico	Factor de virulencia, Fitotoxina, Antifúngico.	FA,Ser,D-Ser,Dab,D-Dab,Arg,Phe,dhAbu,OH-Asp,4Cl-Thr	(68–71)
Syringopeptina	*Pseudomonas syringae pv. Syringae*	Lipopéptido cíclico parcial	Antibiótico, Fitotoxina.	FA,dhAbu,D-Pro,D-Val,Val,D-Ala,D-Ala,D-Val,D-Val,dhAbu,D-Ala,D-Val,Ala,D-Ala,dhAbu,D-aThr,D-Ser,D-Ala,dhAbu,Ala,Dab,D-Dab,Tyr	(68, 69, 72)
Vancomycin	*Amycolatopsis orientalis*	Glucopéptido	Antibiótico.	Asn,bOH-Cl-Tyr,NMe-Leu,Hpg,D-Glc,Van,bOH-Cl-Tyr,Dhpg,Hpg	(11, 73)

1.2 Biosíntesis

Los NRP son sintetizados independientemente del ribosoma, por grandes complejos de proteínas multimodulares, las NRPS. Estas megasintetasas están compuestas por una serie de módulos iterativos. Cada módulo se compone a su vez de dominios catalíticos que llevan a cabo una función específica en la biosíntesis de NRP. Los dominios básicos que contiene un módulo NRPS son el dominio de condensación (C), que cataliza el enlace peptídico; el dominio de adenilación (A), que selecciona y activa el próximo aminoácido; el dominio acarreador de péptidos o 'PCP' (T); y uno o dos dominios tioesterasa (TE) al final del complejo enzimático, responsables de expulsar el péptido ya sintetizado y en algunos casos llevar a cabo la ciclación del péptido. Este dominio también puede encontrarse como un gen propio (4, 75). Debido a su naturaleza proteica, los módulos de las NRPS han evolucionado de tal forma que también son capaces de sintetizar péptidos que contienen aminoácidos inusuales, incluyendo D-aminoácidos, β-aminoácidos, aminoácidos hidroxilados o N-metilados, etc. (76).

Dominios adicionales ligados a las NRPS catalizan reacciones decorativas como la acilación, glucosilación, metilación, epimerización, oxidación, heterociclización, formilación, halogenación, etc. Estos dominios enzimáticos adornan al péptido producto, lo que le confiere características únicas que pueden llegar a ser relevantes en su actividad biológica (18). Debido a la arquitectura modular de estas proteínas, se ha hecho un considerable esfuerzo en tratar de reprogramarlas usando ingeniería genética, para producir nuevos agentes farmacéuticos, al combinar, agregar o modificar módulos para conferir estructuras nuevas intentando mejorar la actividad biológica del péptido, aunque con pocos ejemplos de éxito, y generalmente reduciendo la eficiencia de síntesis (77, 78). Esta compleja organización modular de las NRPS puede ser una razón para el número limitado de estudios sobre la producción a gran escala de NRP (17).

El estudiar cómo ha sido su historia evolutiva podría ayudar a estos esfuerzos de ingeniería genética, al entender e imitar los eventos de duplicación de genes, translocación de módulos y divergencia genética que han producido esta enorme

diversidad de arquitecturas modulares y substratos, las cuales, moldeadas por los procesos de selección, producen una variedad prácticamente infinita de NRP (79). Al tener la capacidad de utilizar como unidades de síntesis más de 35 aminoácidos proteinogénicos y no proteinogénicos diferentes, tan solo un péptido de 7 residuos tiene más de 65 mil millones de posibles configuraciones, sin mencionar las modificaciones adicionales que se puedan añadir. Esto hace complicado el estudio funcional de los NRP, por lo cual las herramientas bioinformáticas han sido desarrolladas para predecir la estructura de los productos de las NRPS a partir de su secuencia, así como para inferir su posible función, importancia y novedad, para ultimadamente tener las bases para priorizar el análisis bioquímico y su caracterización molecular (2, 7).

Existen tres tipos de NRPS (Figura 3), clasificándose por su mecanismo de síntesis. La más común es donde el orden de los módulos dicta la secuencia de aminoácidos del péptido producto, éste mecanismo es el más abundante y mayormente estudiado dentro de las bacterias y los hongos unicelulares, generalmente utilizada por NRPS multimodulares (5). Por otro lado, existen aquellas NRPS que utilizan un mecanismo iterativo de síntesis, en el cual uno o dos módulos son los responsables de añadir varias unidades de síntesis repetidas, como la Serrawettina donde un solo módulo se encarga de sintetizar un dipéptido compuesto por dos serinas (65). Otro tipo es aquel que sintetiza de una forma no linear, es decir, cada módulo es responsable de añadir una unidad de síntesis, pero ni su orden ni su cantidad corresponden a la secuencia de modular. Por otro lado, existen las proteínas híbridas NRPS/PKS que es una fusión entre módulos NRPS con módulos de otras proteínas multimodulares que sintetizan Policétidos (PKS), las cuales funcionan como una línea de ensamblaje en el cual se usa malonil-CoA como unidad de síntesis, en lugar de aminoácidos (80). Algunos ejemplos de policétidos con relevancia son el antibiótico Eritromicina, o la toxina carcinógena Aflatoxina B1 (81). Estas proteínas híbridas añaden complejidad y variabilidad a las estructuras peptídicas, consecuentemente explorando y mejorando las funciones biológicas de estos productos naturales.

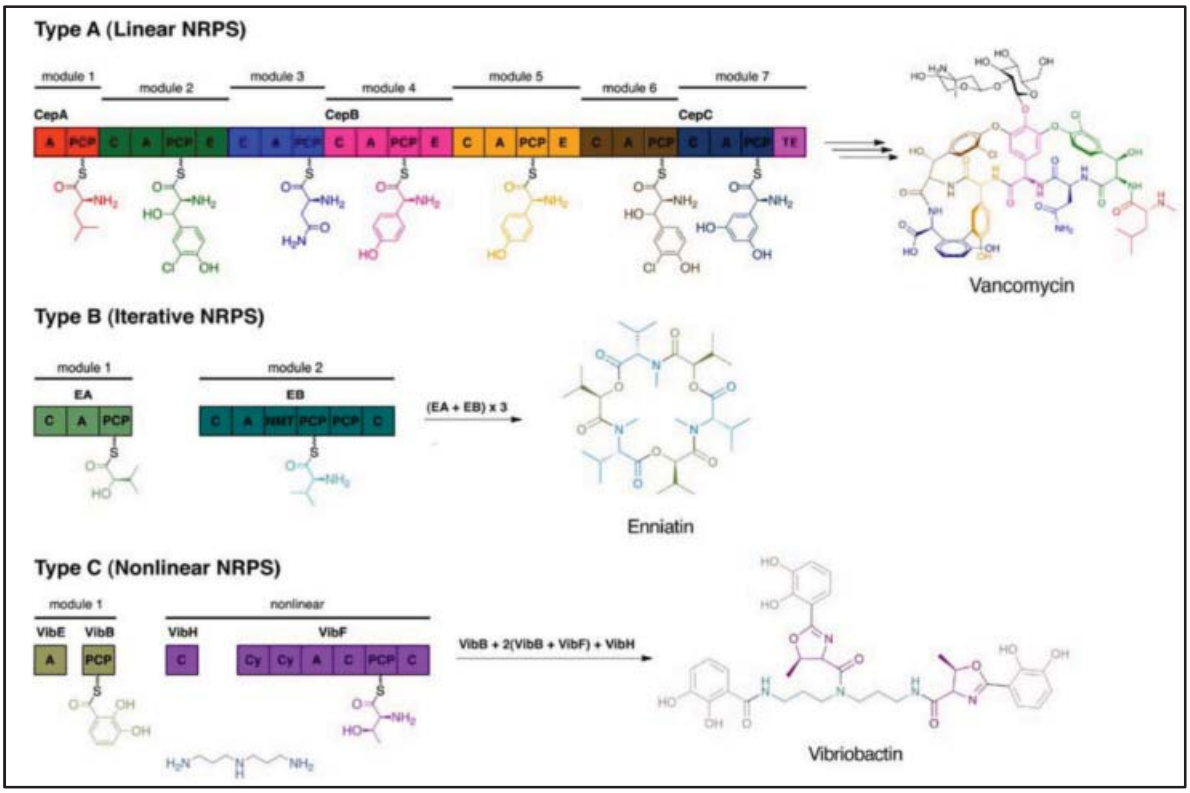

Figura 3. Tres tipos de ensamblaje de péptidos por las NRPS. Tipo A, línea de ensamblaje clásica linear, donde se correlaciona el arreglo modular con la secuencia de aminoácidos del péptido producto, la presentan la mayoría de las NRPS multimodulares de bacterias y hongos; Tipo B, línea de ensamblaje iterativa, en donde se utiliza el mismo substrato para la síntesis de péptidos simétricos, generalmente utilizada por NRPS de uno o dos módulos; Tipo C, ensamblaje no linear, en donde no se correlaciona el número de módulos con la secuencia del producto (adaptado de referencia 13).

1.3 Estructura y BGCs

Los genes de las NRPSs generalmente se encuentran en un clúster de genes biosintético (BGC por sus siglas en inglés, Figura 4). Un BGC se define como un grupo físicamente agrupado de dos o más genes en un genoma particular, que juntos expresan las enzimas de una ruta biosintética para la producción de un metabolito especializado (incluyendo sus variantes químicas). Con respecto a la vía de síntesis de los NRPs, los BGCs comprenden a las proteínas NRPSs, así como las proteínas accesorias como transportadores intermembranales, enzimas decorativas, reguladores transcripcionales, etc. Cabe destacar que los genes de NRPSs son algunos de los genes de mayor tamaño en los genomas bacterianos. Dado que un solo módulo asume aproximadamente 1,050 aminoácidos, algunos NRPS pueden llegar a tener más de 10,000 aminoácidos, una proteína enorme comparada con el promedio de longitud para una proteína bacteriana de 300 aminoácidos (82).

Uno de los BGC más estudiados, es aquel responsable de la biosíntesis de la Surfactina (Figura 5A), proveniente de especies del género *Bacillus*. El estudio de esta molécula lleva más de 45 años, donde a lo largo de toda su historia se han descubierto nuevos usos. Este lipopéptido fue identificado por sus capacidades para inhibir la formación de agregados de fibrina y lisar eritrocitos. Fue llamada Surfactina por sus propiedades surfactantes, ya que reduce la tensión superficial del agua de 72 mN/m a un intervalo de entre 32 y 27 mN/m (dependiendo de su concentración), lo que la hace un potente biosurfactante, en comparación con un surfactante químico como el Triton X-100, que reduce la tensión superficial del agua a 31 mN/m (12, 83). Dentro de sus múltiples usos destacan sus actividades antibacteriales y antitumorales; también se ha propuesto como un substituto a surfactantes químicos, para su uso en la bioremediación de ambientes contaminados por hidrocarburos, aunque actualmente los medios de producción de este biosurfactante hacen que su costo sea muy elevado.

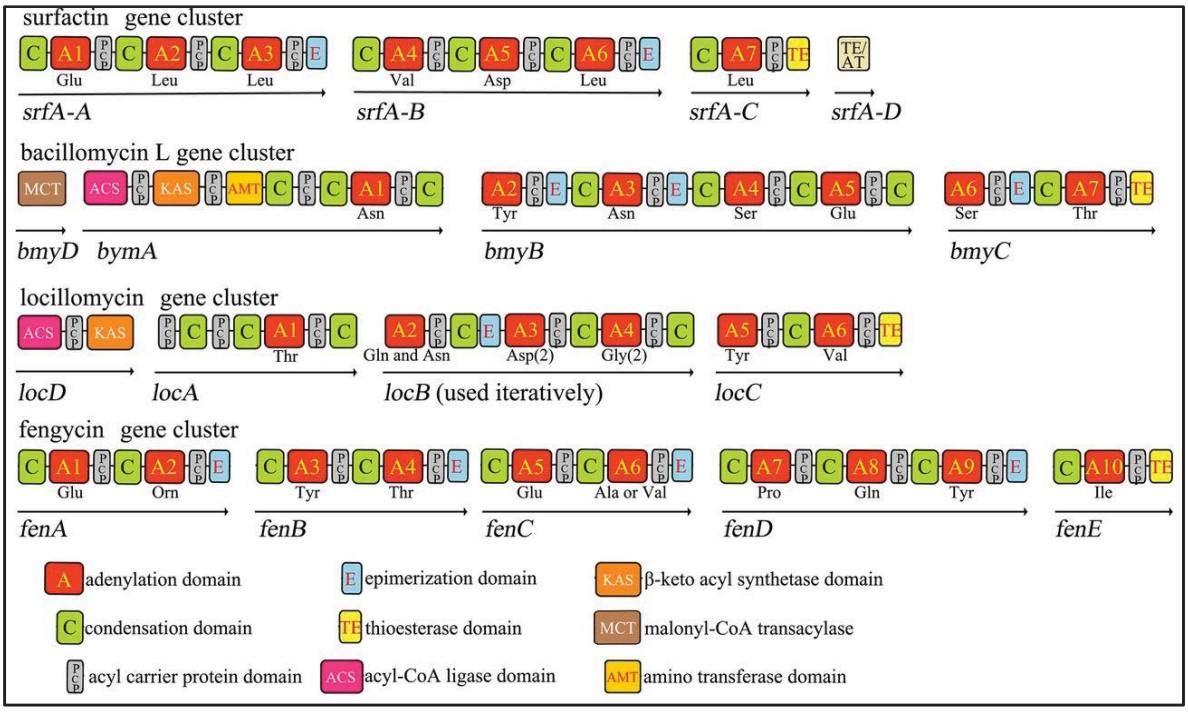

Figura 4. Esquema de la organización de algunos BCG que expresan NRPS, las cuales sintetizan lipopéptidos. La bacillomicina L y locillomicina contienen dominios propios de las PKS por lo que se clasifican como híbridos PKS/NRPS (adaptado de referencia 84).

La surfactina es un lipopéptido cíclico de 7 aminoácidos, el cual es sintetizado por tres proteínas NRPS: SrfA-A, SrfA-B y SrfA-C, que incorporan los residuos [FA-Glu-Leu-Leu], [Val-Asp-Leu] y [Leu] respectivamente, al péptido producto (Figura 5B). El primer dominio de condensación (C-starter) se encarga de iniciar la síntesis al catalizar la transferencia de un ácido graso-CoA al primer aminoácido de la cadena [Glu]. El último módulo (SrfA-C) contiene un dominio de thioesterasa, que junto con un dominio de thioesterasa externo se encargan de ciclizar el péptido y liberarlo. Cabe destacar que los dominios TE pueden estar en genes separados, como el gen *srfTE*, que tiene un papel importante en la iniciación y la liberación del péptido, también se ha reportado que aumenta la eficiencia de la unión de la cadena lipídica al péptido producto en la biosíntesis de la surfactina (85).

La proteína SrfA-A es la inicial, que además de incorporar los primeros tres aminoácidos, añade un ácido graso (FA). Esta reacción la cataliza el primer dominio de condensación, llamado C-starter, el cual tiene diferencias en sus regiones conservadas comparado con

los dominios de condensación intermedios. Además de los principales genes de síntesis, los BGCs de las NRPSs contienen genes para el transporte, para la regulación, y que pueden llevar a cabo modificaciones adicionales al péptido producto.

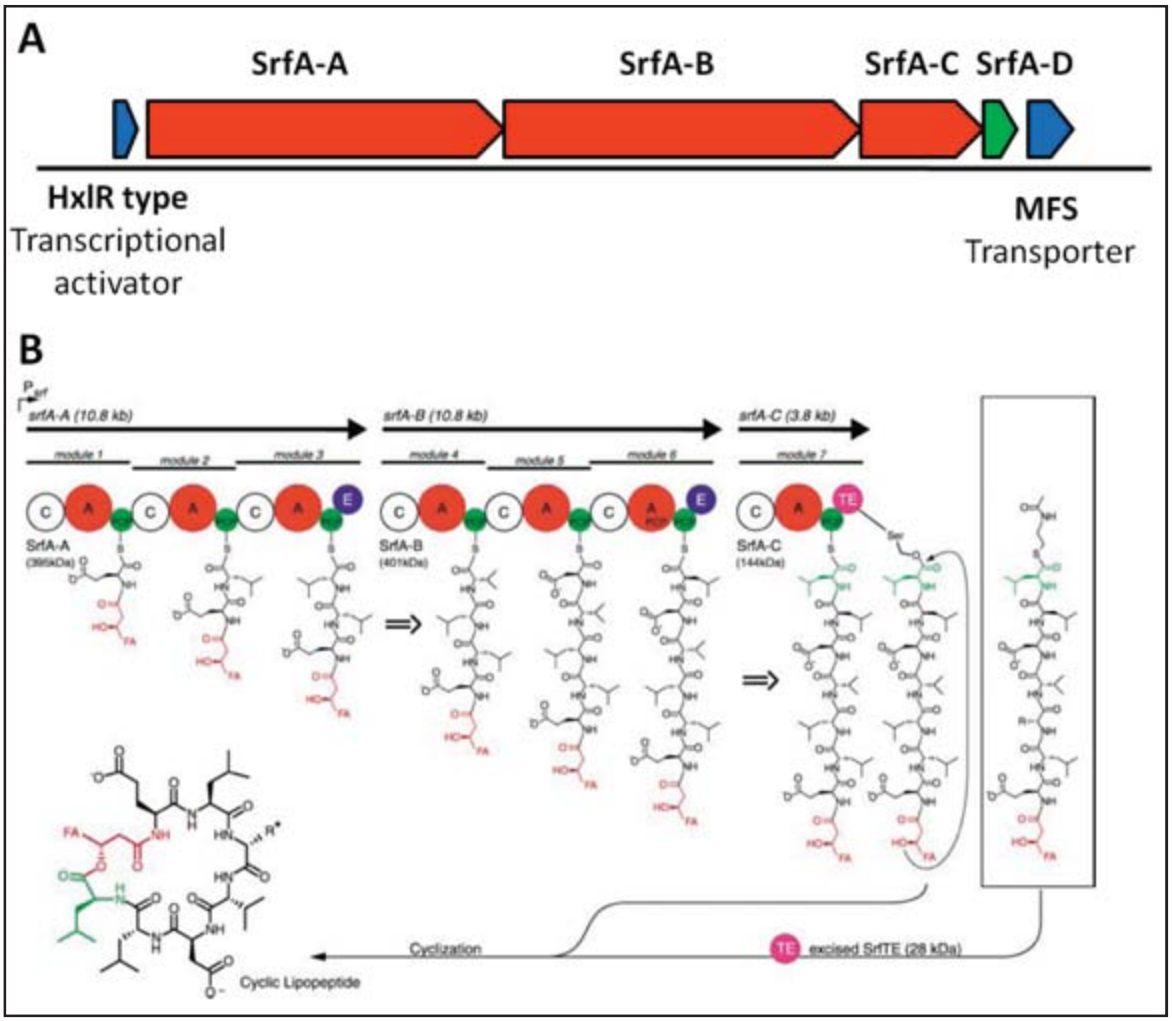

Figura 5. Biosíntesis de la Surfactina. A, BGC involucrado en la producción de la Surfactina A de Bacillus subtilis (Strain 168). Rojo, genes NRPS. Verde, SrfA-D con un Dominio TE involucrado en la iniciación de la biosíntesis, el BGC de la surfactina comprende alrededor de 25 Mb. Azul, genes de regulación y transporte a través de la membrana; B, Vía NRPS para la producción de surfactina, en donde cada modulo es responsable de la elongación de la cadena del péptido con el aminoácido correspondiente, siendo su secuencia co-lineal a la arquitectura modular en rojo tienen una afinidad a un aminoácido en específico y se encargan de activar al aminoácido a costa de un ATP. Los dominios PCP en verde, se encargan de acarrear al aminoácido en forma de amino-acil o peptidil-S-fosfopantetenina (adaptado de referencia 5).

2. Dominios NRPS

2.1 Dominio de Adenilación

Los dominios de adenilación (A) tienen la función de seleccionar al péptido y activarlo para la consecuente elongación del péptido. Estos dominios pertenecen a la familia de sintetasas/ligasas dependientes de AMP, dentro de esta familia también se encuentran la luciferasa, las ligasas de ácidos grasos de cadena larga-CoA, la sintetasa de acetil-CoA, entre otras. Este dominio se compone de alrededor de 500 aminoácidos, el cual se divide en dos subdominios, el subdominio principal N-terminal comprende los primeros 400 aminoácidos, responsable de la catálisis, y el subdominio C-terminal que comprende los últimos 100 aminoácidos, que se ha reportado que puede tener un papel estructural en la holoenzima NRPS, asimismo se han descrito 10 motivos conservados (86, 87), y aunque solo unas pocas estructuras terciarias de los dominios A (Figura 6B) han sido caracterizadas, se tiene una buena noción de cómo funcionan.

El dominio A lleva a cabo su función en una reacción de dos pasos, en un mecanismo bi-uni uni-bi ping-pong (Figura 6A). Es decir, entran dos substratos a la primera reacción y sale un producto (bi-uni); para la segunda reacción entra un substrato y salen dos productos (uni-bi), el prefijo ping-pong se refiere a que para que comience una nueva ronda enzimática, se tiene que haber terminado la reacción anterior (88). Primero, la fase de formación de adenilato, donde se une selectivamente el aminoácido correspondiente, une un AMP al carboxilo del aminoácido y lo convierte en un adenilato-aminoacil a expensas de un Mg-ATP. Cuando el PPi abandona el sitio catalítico esto produce un cambio conformacional (Figura 6D), donde el sub-dominio C-terminal gira para dar lugar a la segunda reacción. El dominio catalítico ahora provoca un ataque nucleofílico del grupo sulfhidrilo del brazo 4'-fosfopanteteína del dominio T adyacente hacia el carboxilo adenilado del aminoácido, liberando el AMP (18, 87).

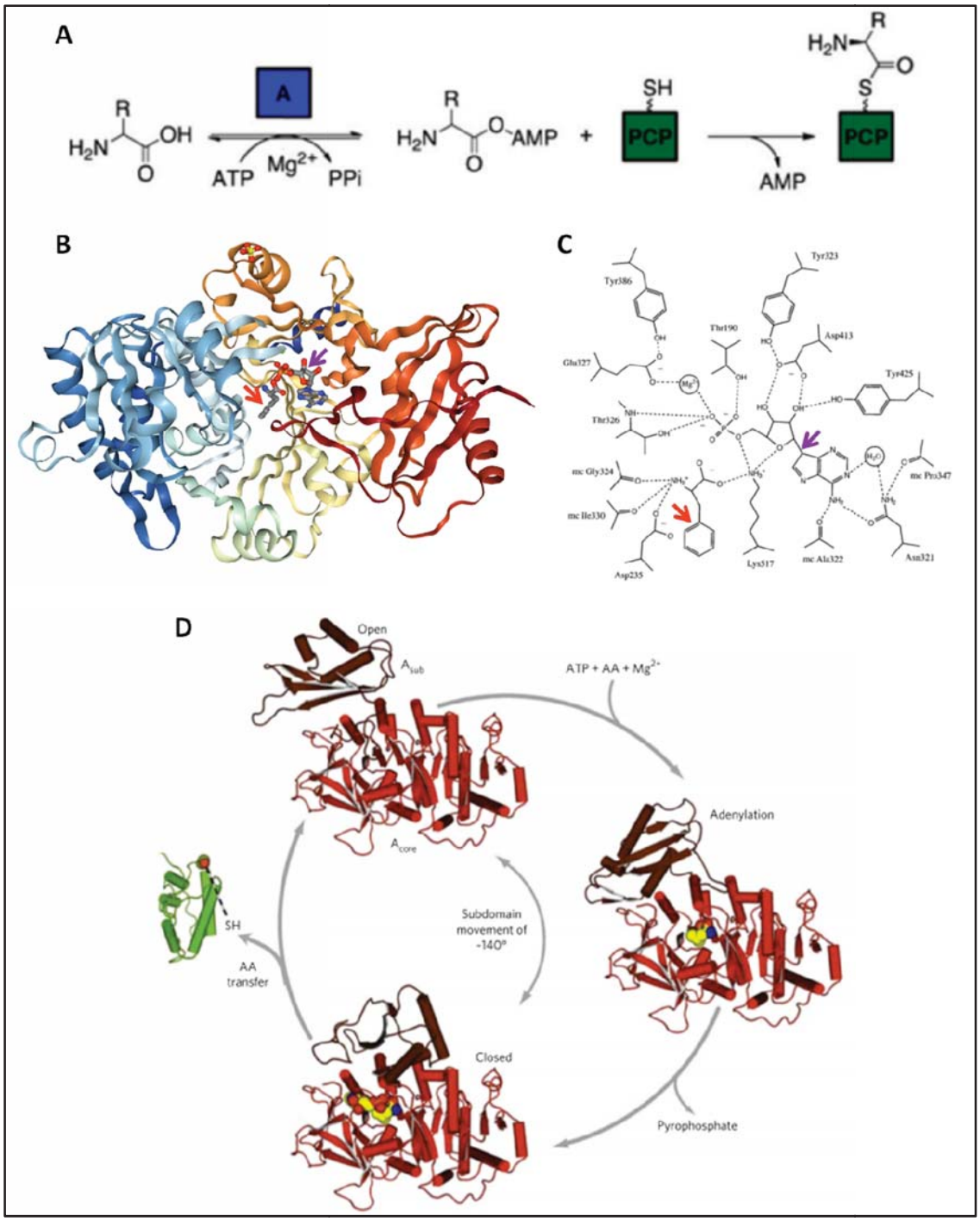

Figura 6. Estructura y función del dominio de adenilación. A, Reacción catalizada por el dominio de adenilación (adaptado de referencia 87); B, Estructura del primer dominio de adenilación de la sintetasa 1 de la Gramicidina A (PBD: 1AMU) en complejo con sus ligandos; C, esquema que representa las interacciones entre los ligandos y el bolsillo catalítico del dominio de adenilación (adaptado de referencia 45). Flecha roja, Fenilalanina. Flecha morada, AMP; D, Cambios conformacionales que sufre el dominio de adenilación durante las etapas de reacción. Rojo, dominio de adeniación. Verde, dominio PCP. Amarillo, substrato (adaptado de referencia 89).

Dentro del sitio activo (Figura 6C), se han identificado de 8 a 10 aminoácidos responsables de la especificidad del substrato, estas firmas se ha confirmado que se conservan para cada uno de los diferentes aminoácidos, incluso pudiendo distinguir entre la conformación estereoquímica de cada aminoácido (90). Esto se ha referido como el código no ribosomal, el cual ha sido explotado por varios algoritmos como 'NRPSpredictor2' (91). Por otro lado, se han llevado a cabo varios estudios para tratar de modificar este código con intención de cambiar la especificidad del sitio activo.

2.2 Dominio de condensación

Los dominios de condensación (C) son responsables de catalizar la unión de dos aminoácidos por un enlace peptídico. Este dominio comprende alrededor de 450 aa, y está ubicado en la sección N-terminal del módulo NRPS. Se cataliza la unión de dos substratos aminoacil monoméricos unidos a los dominios PCP adyacentes, donde se lleva a cabo un ataque nucleofílico por parte del grupo amino del substrato del módulo anterior hacia el grupo tioéster del siguiente substrato en la secuencia de síntesis (Figura 7A), formando un enlace amida y transfiriendo el péptido intermediario de un módulo al siguiente (18).

Basándose en estructuras cristalográficas del gen *vibH* compuesto únicamente por un dominio de C (92), como de un dominio C y un dominio PCP escindidos del gen original (93) se observa que estos dominios están compuestos por dos subdominios N- y C-terminal, formando una 'V' con el sitio catalítico localizado donde estos dos subdominios se juntan (Figura 7B). Aunque no se ha comprobado completamente, por varios estudios mutacionales y de análisis de pKA de los aminoácidos conservados dentro del dominio catalítico, se piensa que por interacciones electrostáticas, estos dos subdominios son capaces de llevar a cabo esta reacción, sufriendo cambios conformacionales "cierran" la 'V' y que acercan a los dos substratos, lo que fuerza la formación del enlace peptídico.

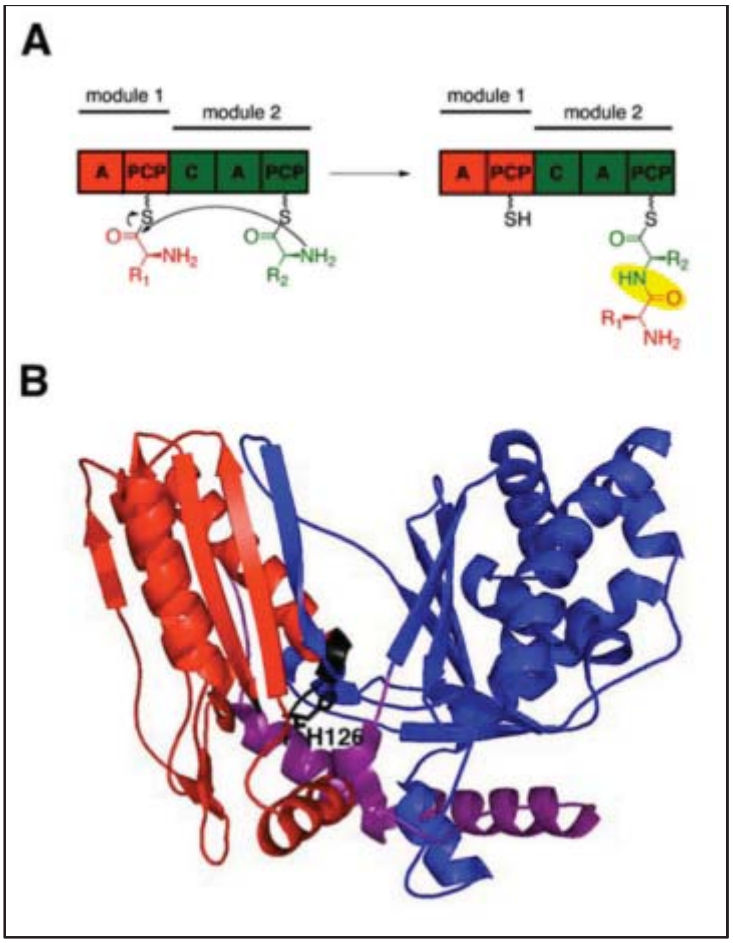

Figura 7. Estructura y función del dominio de condensación de las NRPS. A, Reacción catalizada por el dominio de condensación. B, Estructura la proteína VibH (PBD: 1L5A) con un único dominio de condensación. Rojo, N-terminal. Azul, C-terminal. Morado, región linker. Negro, motivo conservado catalítico (adaptado de referencia 13).

Este dominio ha sufrido una diversificación funcional dentro de las NRPS, habiendo al menos 6 tipos de dominios de condensación (Figura 8C). Estos son: el dominio $^{L}C_{L}$, que cataliza la unión entre dos L-aminoácidos; el dominio $^{D}C_{L}$, que une un D-aminoácido y un L-aminoácido; los dominios de Epimerización (E) ya sean únicos o duales (E/C); los dominios de heterociclización; y los dominios "C-starter", que solo se encuentran en el módulo de iniciación (94).

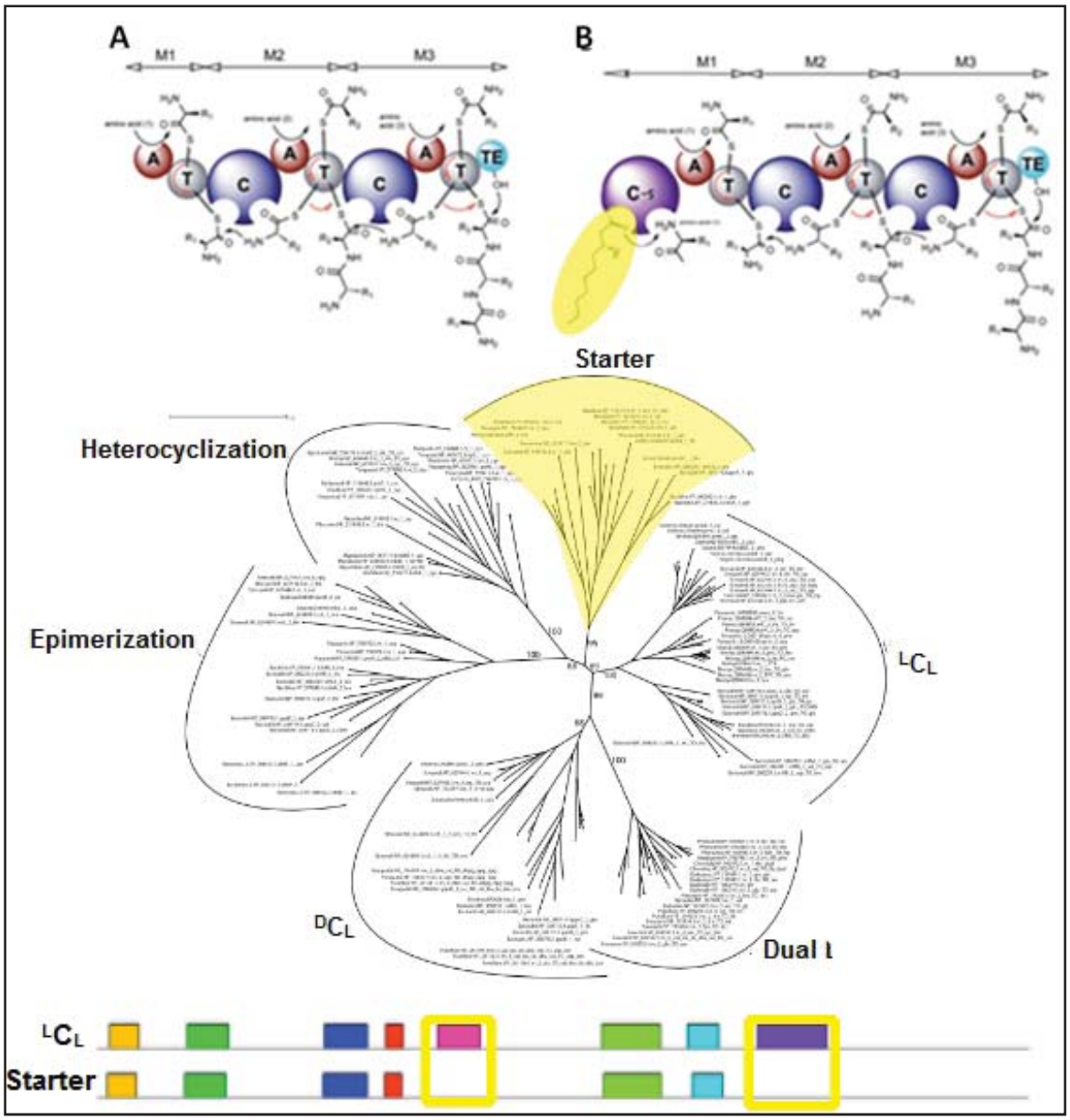

Figura 8. Diferentes dominios de condensación que aparecen en las NRPS, estos dominios tienen un origen en común, y se han diversificado las funciones. A-B, dos tipos de NRPS, dependiendo si contienen un dominio C-starter o no. M1, módulo de iniciación. M2, módulo(s) de elongación, repetidos n-2 veces. M3, módulo de terminación. El módulo de iniciación puede contener un dominio de condensación adicional (C-starter) que es el responsable de añadir la cadena lipídica (adaptada de referencia 78); C, los diferentes dominios de condensación comparten un ancestro en común, del cual se han diversificado. Sombreado amarillo, rama en la que se agrupan los dominios C-Starter (adaptado de referencia 93); D, esquema de la secuencia de los dominios de condensación $^{L}C_{L}$ y C-starter y sus motivos conservados. Se observan diferencias (e-value < 1e-10) principalmente entre los en los motivos señalados con un recuadro amarillo (Basado en referencia 93, comprobado en nuestros datos usando el programa MEME).

El dominio C-starter es responsable de llevar a cabo la unión de una cadena de acilos de cadena variable al primer aminoácido de la secuencia de ensamblaje (Figura 8B). Este dominio C-starter tiene variaciones en sus motivos conservados (94), lo cual lo diferencia de los demás tipos de dominios C en los motivos C4 y C6 (Figura 8D). Una hipótesis que ha surgido a lo largo del trabajo es que estos motivos de alguna manera están involucrados en la afinidad hacia una cadena lipídica en lugar de un aminoácido, probablemente en la estructura del bolsillo catalítico, permitiendo la entrada de una molécula de mayor longitud y con mayor hidrofobicidad comparada a un aminoácido, como lo es una cadena lipídica, aunque es necesario llevar a cabo estudios estructurales para determinar su papel.

2.3 Dominio acarreador de péptidos

Unos dominios pequeños de 80-100 aminoácidos llamados proteínas acarreadoras de peptidilos (PCP por sus siglas en inglés ó T) tienen la función de sostener y transferir a la cadena peptídica en crecimiento por todo el proceso de catálisis (Figura 6D). Éste dominio es modificado postraduccionalmente por una PPTasa, transfiriendo el brazo móvil 4'-fosfopanteteína de una CoA hacia un residuo serina en el sitio activo de la PCP, convirtiéndolo a su forma activa. Lo cual le permite unir aminoacilos vía un enlace tioéster catalizado por el dominio de adenilación al grupo sulfhidrilo terminal del brazo fosfopanteteína (93, 95).

2.4 Dominio tioesterasa

Los dominios tioesterasa (TE) se encuentran generalmente en el último módulo de la línea de ensamblaje de las NRPS, aunque pueden encontrarse como genes propios, como el gen srfTE. Estos pueden funcionar como hidrolasas, catalizando la liberación del péptido terminado al hidrolizar el grupo tiol del brazo PPT del último dominio PCP (Figura 9), o como ciclasas al llevar a cabo la ciclación del péptido.

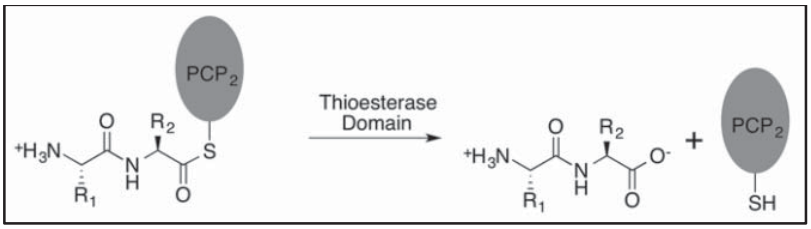

Figura 9. Reacción catalizada por el dominio tioesterasa (adaptado de referencia 13).

2.5 Dominios adicionales

Con dominios y enzimas accesorias las NRPS logran sintetizar péptidos que de otra forma nunca se podrían observar en la naturaleza. Estas proteínas decorativas pueden o no encontrarse dentro de los genes NRPS, pudiendo actuar como dominios enzimáticos dentro de la secuencia de síntesis del péptido, como enzimas que modifican a los aminoácidos precursores formando aminoácidos no proteinogénicos, o enzimas de modificación post-síntesis (haciendo alusión a la modificación post-traduccional).

Dentro de estos dominios adicionales están los dominios que han surgido por la diversificación del dominio de condensación: El dominio de epimerización es uno de los más comunes, el cual tiene la función de llevar a cabo una epimerización in-situ del carbono alfa del aminoácido acarreado por la PCP. Asimismo, los dominios duales E/C llevan a cabo la misma función (94), que les permite la síntesis de péptidos con D-aminoácidos, que produce estructuras únicas que influyen en la actividad biológica, también se ha observado que al incorporar un D-aminoácido se orienta el péptido en crecimiento de cierta forma que se puedan llevar a cabo modificaciones adicionales, como es el caso de la Vancomicina (11); Otro dominio accesorio homólogo a los dominios de condensación es el dominio de heterociclización (Cy) el cual se encarga de formar heterociclos dentro del péptido producto entre posiciones variables dentro del péptido (96, 97). Esto se logra al formar anillos oxazolinona o tiazolinona, que derivan de serina/treonina o cisteína respectivamente, en tres pasos enzimáticos: primero el aminoácido que contiene el grupo beta nucleófilo se condensa con el siguiente aminoácido en la secuencia, el enlace peptídico ahora sufre un ataque nucleofílico por el grupo hidroxilo de la serina/treonina o por el grupo sulfhidrilo de la cisteína, formando el

anillo. Finalmente sufre una deshidratación para dar lugar a anillos como los presentes en la Mycobactina o la Bacitracina (18).

Existen otras reacciones enzimáticas que añaden complejidad a las estructuras peptídicas sintetizadas por las NRPS, algunas de ellas son las metilaciones, las formilaciones o las halogenizaciones, etc (6). Estas son catalizadas por enzimas específicas, generalmente dentro del BGC, aunque también pueden presentarse dentro de las NRPS como parte de un módulo funcional.

Las metilaciones son catalizadas por las metiltransferasas (MT) los cuales llevan a cabo N- o C-metilaciones dependiendo de si la metilación ocurre durante la biosíntesis del péptido o como una modificación al aminoácido precursor (98), respectivamente, transfiriendo el grupo metilo del co-substrato S-adenosil metionina. Un ejemplo de una NRPS que contiene un dominio de metilación es: El BGC recientemente descubierto que sintetiza un antibiótico peptídico de 11 residuos llamado Teixobactina (Figura 10), el cual es un nuevo tipo de antibiótico con doble bioactividad en contra de varias cepas gram positivas multiresistentes como MRSA o VRE sin desarrollo detectable de resistencia. Esta NRPS presenta en el primer módulo del gen *txo1* un dominio de metilación el cual lleva a cabo una N-metilación al residuo [Phe_1] (99, 100).

Un dominio de formilación es responsable de catalizar la N-formilación de aminoácidos, usando el cofactor N-formiltetrahidrofolato como donador del grupo formil. Este dominio ha sido poco estudiado debido a que solo se han identificado pocos representantes, un ejemplo es, el perteneciente a la NRPS que sintetiza una gramicidina linear en una cepa de *Brevibacillus brevis*, el cual contiene en el primer módulo del gen *lgrA* un dominio de formilación, responsable de modificar el residuo [Val_1] (46);

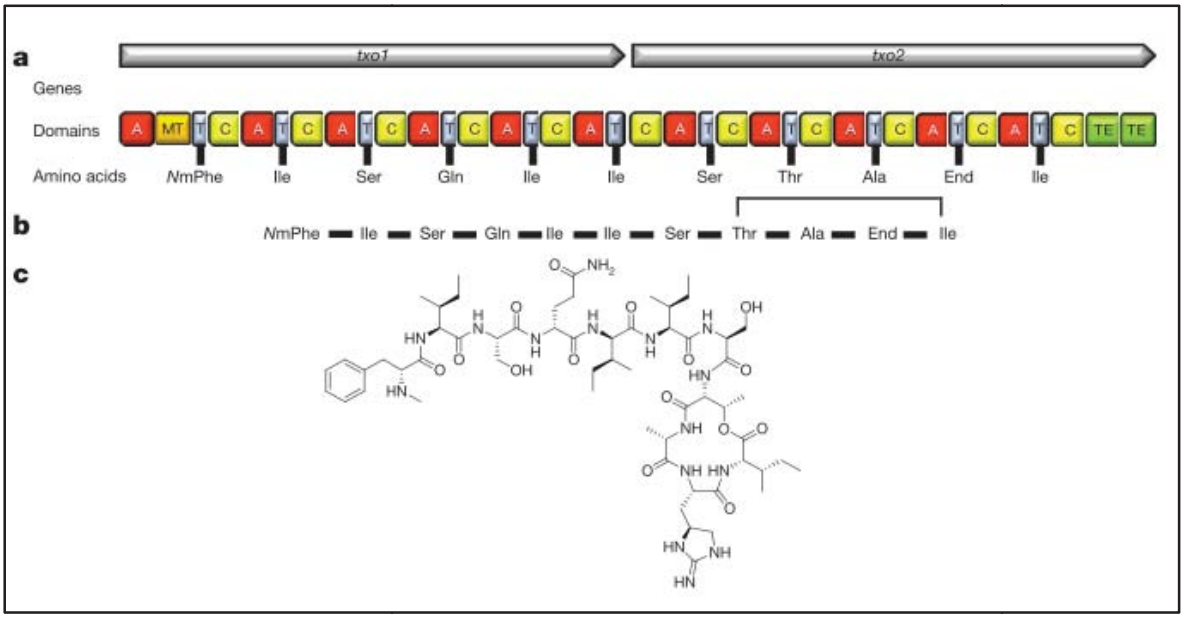

Figura 10. Arquitectura del BGC que sintetiza a Teixobactina. El primer módulo presenta un dominio metiltransferasa (adaptado de referencia 100).

Por último, una modificación a los NRP que tiene una gran importancia en la actividad farmacológica es el dominio enzimático responsable de halogenar residuos en las NRPS. Esta reacción es catalizada por un dominio de halogenación, que puede presentarse en dos tipos en las NRPS (18, 87). El primer tipo son las halogeneasas dependientes de flavina, los cuales unen el grupo halógeno a un anillo aromático, ejemplos de NRPS que contienen estos dominios son aquellas que sintetizan la Vancomicina (11) y la Balhimicina (46), donde los grupos cloro son los que confieren su actividad antimicrobiana (Figura 2, resaltado en rojo), comparado con sus contrapartes sin halogenaciones, se aumenta más de diez veces su antibiosis. El segundo tipo son las halogenasas dependientes de hierro, que son responsables de halogenar NRP en átomos de carbono centrales alifáticos no activados, como en la Syringomicina E, en donde su grupo cloro le confiere su actividad fitotóxica (71).

2.6 Híbridos NRPS/PKS

Existen BGC que contienen tanto módulos de NRPS como módulos característicos de Sintetasas de Policétidos (PKS por sus siglas en inglés), los cuales se cree que han aparecido por eventos de recombinación y 'barajeo' de los módulos biosintéticos entre

estas dos clases de megasintetasas. Las PKS utilizan unidades de malonil-CoA de dos carbonos derivados de tioésteres de acetato u otros ácidos carboxílicos de cadena corta como unidades de síntesis de cada módulo. Los módulos de las PKS están compuestas por tres dominios básicos, una aciltransferasa (AT) que selecciona y transfiere la unidad extensiva, una proteína acarreadora de acilos (ACP) con un brazo de fosfopanteteína móvil, y una sintasa de ketoacilos (KS) (101). Ejemplos de péptidos sintetizados por NRPS/PKS son el antibiótico iturina (38), o la bleomicina, que presenta propiedades anti-tumorales, además de ser utilizado para tratar una variedad de enfermedades cancerígenas (102, 103). Las NRPS/PKS han tenido especial atención en los últimos años, debido a que sus productos tienen una variabilidad en sus actividades y estructuras muy extensa. Estos BGC hibridados se presentan casi en las mismas proporciones que sus contrapartes únicas (Figura 11), evocando un mecanismo generalizado en las bacterias que provoca la combinación de estos dos tipos de megasintetasas multimodulares(80).

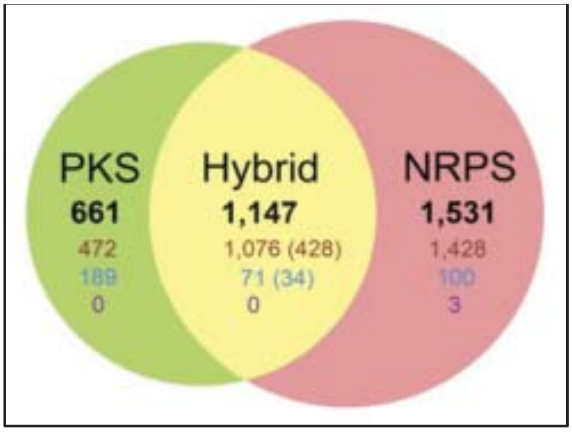

Figura 11. Diagrama de Venn expresando la cantidad de BGCs de PKS, Híbridos NRPS/PKS, y NRPS identificados en una búsqueda por palabras clave en genomas anotados. Números en rojo, azul y morado representan a los BGCs encontrados en Bacterias, Hongos Ascomicetos y Arqueas, respectivamente (tomado de referencia 80).

3. Minería Genómica, Búsqueda de Productos Naturales, Predicción y Caracterización Bioquímica del Producto.

3.1 Secuenciación masiva

Las nuevas tecnologías de secuenciación masiva nos han permitido dar un vistazo a la enorme diversidad de microorganismos que habitan en casi cualquier nicho ecológico: como organismos de vida libre en los océanos, en los suelos, e incluso en la atmosfera; como formas de vida simbiótica como en las ventilas hidrotermales en mares profundos, asociadas a esponjas, consorcios de bacterias, en el intestino de casi todos los organismos multicelulares, etc. Aunado a esto, las técnicas de cultivo actuales no nos permiten aislar a la gran mayoría de estos microorganismos; se ha estimado que alrededor del 99% de la diversidad de bacterias, no son posibles de cultivar con los métodos actuales, y se es sabido que no se tiene un representante aislado en más de la mitad de los taxa bacterianos (Figura 12). Con la revolución de la secuenciación masiva, y la creciente capacidad de computadoras y nuevos algoritmos de análisis de secuencias, se ha vuelto posible estudiar a esta gran mayoría de microorganismos anteriormente inaccesibles, aunque esto ha provocado un crecimiento exponencial en la cantidad de información en las bases de datos de secuencias genómicas y metagenómicas. Un ejemplo de esto, es la expedición “Tara Oceans” (104), donde se secuenciaron 210 muestras provenientes de todos los océanos alrededor del mundo, esto produjo una base de datos de alrededor de 7.2 billones de pares de bases, y una colección de 40 millones de genes no redundantes, así como una librería de millones de secuencias de 16S y 18S de organismos en un intervalo de tamaño entre 0.2 µm hasta 2 mm. Es imperativo el desarrollo de nuevas metodologías y algoritmos que nos permitan acceder a esta cada vez más inmensa cantidad de datos, para la identificación, clasificación, y priorización de secuencias genómicas y metagenómicas para llevar a cabo estudios de forma eficiente, así como para la predicción certera de la función de cada gen, y potenciar el descubrimiento de nuevos productos naturales.

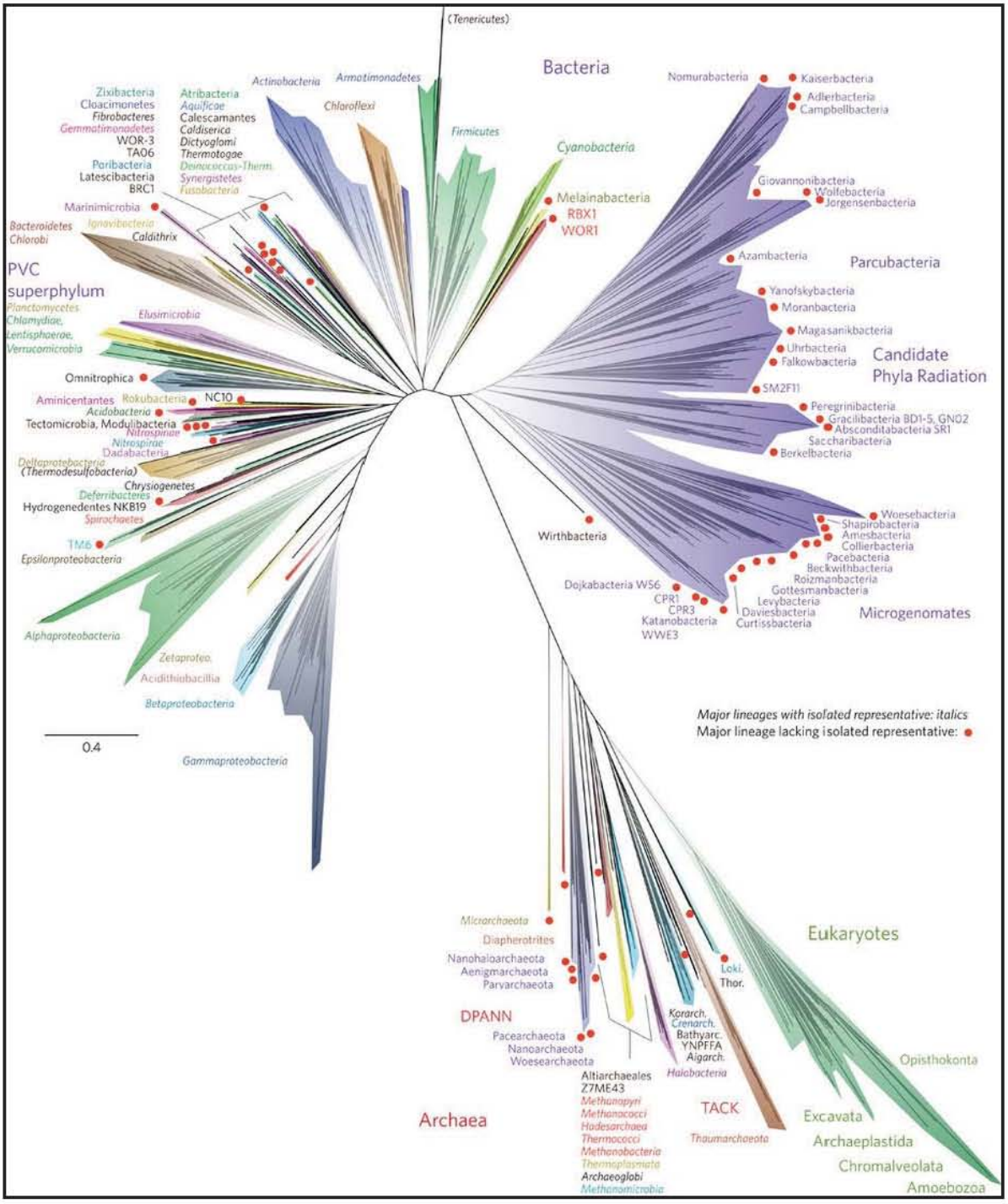

Figura 12. Árbol filogenético que incluye a los nuevos clados que han surgido a partir de la secuenciación masiva. Los puntos rojos indican que para ese linaje no se tiene un representante aislado (adaptado de referencia 102).

3.2 Herramientas Bioinformáticas

Se han hecho muchos estudios de detección por función de alto rendimiento de cepas productoras, microbiomas y metagenomas para descubrir nuevos tipos de NRPS (15), pero pocos desde un enfoque de análisis de secuencias con herramientas bioinformáticas, debido a la gran variedad de estructuras químicas, y la gran diversidad de genes involucrados en su biosíntesis.

El análisis computacional es cada vez más importante para inferir las funciones y estructuras de las proteínas debido a que la velocidad de la secuenciación del DNA ha superado desde hace mucho tiempo la velocidad a la que la función biológica de las secuencias se puede dilucidar experimentalmente. Los algoritmos establecidos de comparación de secuencias detectan similitudes significativas entre secuencias de bases de datos conocidas. Los métodos de comparación de secuencias como el BLAST, generalmente asumen que todas las posiciones de aminoácidos son igualmente importantes. Los alineamientos múltiples de familias de secuencias de proteínas indican qué residuos están más conservados que otros, y los puntos en los que las inserciones y deleciones son más frecuentes (106).

Las proteínas generalmente se componen de una o más regiones funcionales, comúnmente denominadas dominios. Diferentes combinaciones de dominios dan lugar a la amplia gama de proteínas que se encuentran en la naturaleza. La identificación de los dominios que ocurren dentro de las proteínas puede, por lo tanto, proporcionar información sobre su función.

El "perfil" de una proteína está definido como un modelo consenso de la estructura primaria que consiste en puntajes específicos de la posición del residuo y penalidades de inserción o deleción. Los Modelos ocultos de Markov (HMM por sus siglas en inglés) son una técnica general de modelado probabilístico para problemas lineales como secuencias o series temporales y han sido ampliamente utilizados en el análisis computacional de secuencias, incluyendo el modelado estructural de proteínas y el análisis a gran escala de secuencias genómicas y metagenómicas (107). La mayoría de los repositorios de

secuencias y bases de datos estructurales de proteínas, como la base de Datos de Dominios Conservados (CDD) (108), o la “Pfam” (109) usan este tipos de algoritmos basados en HMMs.

Al generar un perfil HMM de un alineamiento de secuencias homólogas (Figura 13), se pueden hacer búsquedas de secuencias distantes en grandes bases de datos (110). Los HMM permiten reconocer dominios conservados dentro de los modelos de proteínas, con lo cual se pueden reconocer genes homólogos distantes, filogenéticamente hablando.

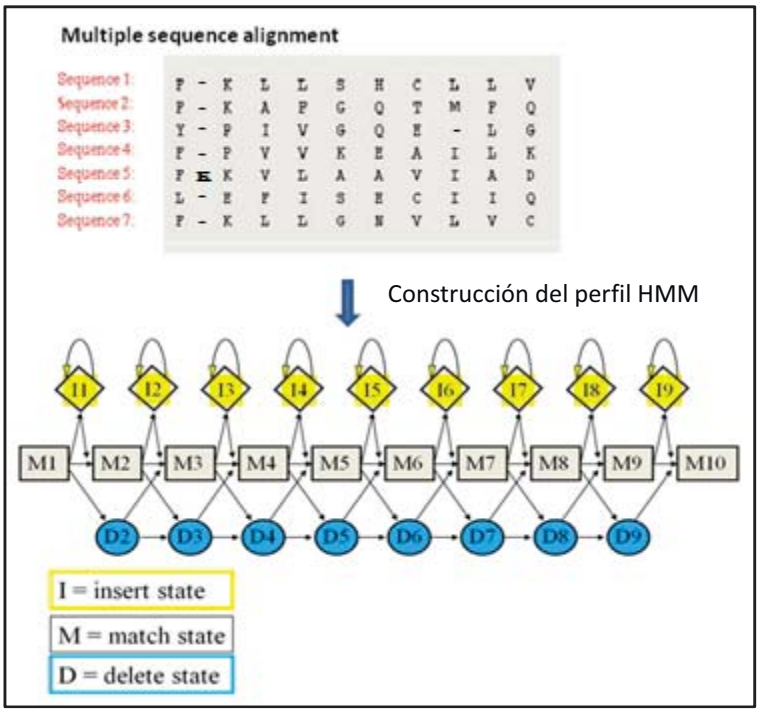

Figura 13. Representación de un modelo de Markov oculto basado en una alineación de secuencias múltiples. A los aminoácidos se les da una puntuación en cada posición en la alineación de secuencia de acuerdo con la frecuencia con la que se producen. Las estados de transición (es decir, las probabilidades de cada par de aminoácidos en una posición n, n+1), los estados de inserción y deleción también son modelados.

Los HMMs son modelos finitos que describen una distribución de probabilidad sobre un número infinito de posibles secuencias. Estos se componen de un número de estados, que corresponden a posiciones de un alineamiento múltiple. Cada estado 'emite' símbolos (residuos) de acuerdo a las probabilidades intrínsecas de cada posición, de la misma forma

los estados están interconectados por las probabilidades de transición. Partiendo de un estado inicial, se genera una secuencia moviéndose de estado a estado de acuerdo con las probabilidades de transición, hasta que se alcanza un estado final, creando una secuencia observable de símbolos (111). Usando los perfiles HMM se pueden reconocer las regiones conservadas dentro de una secuencia, y clasificarlos dentro de un dominio ya definido, asignándole un valor numérico a la similitud entre la secuencia blanco y el perfil HMM. Los dominios proteicos se han clasificado en 16712 familias para la última versión de Pfam (marzo 2018), las cuales abarcan todas las regiones o estructuras conservadas con una función específica (109).

La secuencia de estados es una cadena de Markov, porque la elección del siguiente estado a ocupar depende de la identidad del estado actual. Sin embargo, esta secuencia de estado no se observa: está oculta. Sólo se observa la secuencia de símbolos que generan estos estados ocultos. La secuencia de estado más probable debe inferirse de una alineación del HMM a la secuencia observada (107).

Cabe destacar que ningún programa disponible para la búsqueda de NRPS distingue entre NRPS que sintetizan NRP normales de aquellos que sintetizan lipopéptidos. Por ello, este estudio se enfoca en desarrollar un método para el reconocimiento y análisis de aquellas NRPS que tengan una alta probabilidad de sintetizar lipopéptidos, así como la búsqueda de este tipo de secuencias en bases de datos de secuencias genómicas y metagenómicas.

3.3 Predicción de BGCs, de la estructura y de la función del producto

Al ser co-lineal la secuencia del péptido producto a la organización genética de las NRPS, han surgido bastantes programas computacionales como AntiSMASH (112), SANDPUMA (113), NRPSPredictor2 (91), entre otros, que permiten predecir la secuencia y estructura del producto con bastante confianza (114). Los HMMs se han usado como base de varios algoritmos de análisis de secuencias, uno de los más reconocidos es AntiSMASH (112), que utiliza perfiles HMM para reconocer BGC en genomas bacterianos y predecir sus productos usando otro programa, NRPSpredictor2 (91) que igualmente incorpora al algoritmo HMM para reconocer la afinidad de los dominios de adenilación, analizando las

firmas conservadas y caracterizadas para hacer una predicción del aminoácido que utilizan como substrato. Estos programas se han utilizado extensivamente para la búsqueda de BGC en genomas de bacterias y de hongos, como en Nielsen et al., 2017 donde se analizaron 24 genomas de especies del género *Penicillium* (15 especies ya caracterizadas y 9 especies nuevas) donde se encontró un gran potencial para el descubrimiento de nuevos metabolitos secundarios. Esta aproximación no solo sirve para NRPS, en un estudio reciente se usaron perfiles HMM de unas enzimas de importancia biotecnológica llamadas Lacasas, para llevar a cabo un análisis bioinformático de la diversidad de este tipo de enzimas en genomas bacterianos y metagenomas (116). Asimismo, se han llevado a cabo búsquedas de NRPS en diferentes tipos de bases de datos; como ejemplo, en (117) se realizó minería de genomas que igualmente reveló un gran potencial para la producción de lipopéptidos y sideróforos en el género *Burkholderia*.

3.4 Ingeniería genética de NRPS

La organización modular sistemática de las NRPS permite la alteración estructural del producto intercambiando dominios o módulos de estas megasintetasas para el diseño de nuevos péptidos no ribosomales (118). En un estudio reciente (119), realizaron mutagénesis del sitio activo del dominio PheA de la sintetasa A de Gramicidina S para cambiar su especificidad a diferentes substratos, donde una mutación doble de los aminoácidos T278M/A301G y una mutación única en H322E, exitosamente cambiaron la afinidad del substrato WT, Phe, por una Leu y un Asn, respectivamente. Habiendo una cantidad importante de trabajos relacionados a la modificación de los péptidos producto, la mayoría tienen algunos contratiempos, como la reducción en la producción, el mal plegamiento de la holoenzima, o simplemente que el péptido producto pierde su función(120). Otro aspecto importante es que la biología sintética permite el rediseño de BGCs para la expresión heteróloga eficaz en hospederos prediseñados, que en última instancia potenciará la construcción de plataformas estandarizadas de alto rendimiento para el descubrimiento de productos naturales (121). La expresión heteróloga en cepas hospederas optimizadas es una alternativa práctica para identificar compuestos de

clústeres de genes biosintéticos novedosos en combinación con herramientas de detección por espectrometría de masas.

Recientemente ha habido un impulso para establecer plataformas optimizadas de expresión en cepas de *Streptomyces* y *Bacillus*, entre otros, en las que se han eliminado los genes NRPS endógenos (78). Estudios de éxito sobre la expresión heteróloga de NRPS en cepas de bacterias productoras, han elucidado aspectos esenciales para el diseño de cepas. Por ejemplo en un estudio reciente, se demostró que la expresión heteróloga del gen *swrW* de *Serratia marcescens* en *E. coli* produjo serrawettina W1 sin tener que expresar genes adicionales (17), pues se ha reportado que los genes de la fosfopanteteíl transferasa PswP y una ACP son esenciales para la producción de la serrawettina W1 (65, 122). Esto se cree que es posible debido a la interacción de las proteínas homólogas en *E. coli* con la NRPS, permitiendo así, la producción de este lipopéptido (17).

Otro ejemplo importante es el diseño de análogos del péptido Teixobactina recientemente descubierto, los cuales presentan una actividad altamente potente (principalmente un análogo resaltó, donde sustituyeron dos substratos: [D-Arg_4] y [Leu_{10}]) contra de cepas multiresistentes *in-vitro* e *in-vivo*. Ésta molécula es una de las pocas hasta la fecha que no presenta citotoxicidad, aparte de tener una gran eficacia antibacterial en modelos *in-vivo*.

4. Antecedentes

La minería de genes es una de las principales fuentes de nuevas biomoléculas con aplicaciones biotecnológicas, aunado al crecimiento exponencial de la secuenciación genomas y metagenomas, ha producido una carrera tecnológica para desarrollar nuevas y cada vez más eficientes formas de explotar esta *mina de oro*. El gran acervo de conocimiento genético se ha generado por esfuerzos mundiales de: la revolución en la secuenciación masiva de DNA; el análisis *in vitro* e *in vivo*, que genera conocimiento genético, funcional, estructural y bioquímico de proteínas, enzimas, productos naturales y metabolitos secundarios; así como, el análisis *in silico* de las secuencias de nucleótidos y aminoácidos para la predicción de las estructuras terciarias de una proteína. Toda esta información nos permite extrapolar los datos bioquímicos a secuencias homólogas y análogas para lograr predecir la función, la regulación, el plegamiento y la estructura de una nueva proteína. Las proteínas están compuestas por dominios funcionales, los cuales son responsables de su actividad, estos dominios están conservados a través de todas las ramas de la vida, lo cual hace posible el reconocimiento de estos dominios dentro de genes completamente nuevos, permitiendo así, su caracterización funcional y estructural.

Las bases de datos genéticas se caracterizan por contener la secuencia de nucleótidos del DNA de un organismo, lo que se conoce como un genoma. A partir de incontables estudios de expresión, regulación y caracterización de genes, se puede inferir y delimitar un gen, permitiendo la traducción a las secuencias de DNA a los aminoácidos de las proteínas de un organismo, generando su proteoma. Esto permite el estudio específico de un gen, de un grupo de genes homólogos, o de un grupo de genes que llevan a cabo una función específica en el organismo. Existen una gran cantidad de bases de datos de secuencias genéticas, así como de la información generada a partir de su estudio. De entre ellas destacan, por la comprensión de conocimiento y la calidad de curación de la información: la base de datos RefSeq de secuencias de proteínas no-redundates del Centro Nacional de Información Biotecnológica (NCBI); y la base de datos UniProtKB del Instituto Europeo de Bioinformática (EBI), que contiene al mes de enero del 2018, más de cien millones de secuencias de proteínas (100,043,962), pertenecientes a más de un cien

mil unidades taxonómicas únicas (108,633) que comprenden los tres dominios de la vida y los virus. La enorme cantidad de Información en estas bases de datos provoca que la minería genética sea prácticamente imposible de lograr con métodos convencionales, por lo que se han desarrollado herramientas informáticas para su clasificación, estudio y anotación funcional. Algunos ejemplos de herramientas ampliamente utilizadas por la comunidad científica para el análisis de secuencias: el famoso BLAST, que utiliza alineamientos locales de secuencias para encontrar la relación de parentesco entre dos o más secuencias; otro ejemplo, los perfiles generados a partir de modelos ocultos de Markov (HMMs), el cual analiza cada región dentro de un grupo de secuencias, permitiendo la caracterización de regiones conservadas dentro de un gen, de esto se derivan herramientas bioinformáticas como HMMER, la base de Datos de Dominios Conservados (CDD), la herramienta de búsqueda de BGC AntiSMASH, etc.

Otro tipo de base de datos de secuencias genéticas que surgió gracias a los avances en la secuenciación masiva de DNA, lo cual hizo posible analizar toda una comunidad microbiana, sin tener que recurrir al aislamiento de cepas, que es llamado un metagenoma. Los metagenomas puede ser de dos tipos: el primero se compone únicamente de las secuencias de la subunidad 16S del ribosoma, el cual es llamado un "Amplicón", con esto se puede saber rápidamente la composición de especies de una muestra ambiental, así como su abundancia; el otro tipo de metagenoma, es el generado a partir del secuenciamiento "whole metagenome shotgun" (WMS por sus siglas en inglés), el cual genera pequeños trozos de secuencias que posteriormente son reconstruidas en conjuntos de secuencias largas, que comprenden genes completos de los organismos presentes, esto permite un análisis genético y metabólico en la muestra. Aquí surge uno de los problemas en este tipo de aproximación, ya que, el WMS genera miles de millones de secuencias únicas en una muestra promedio, por lo tanto, los métodos convencionales para la anotación de secuencias como el BLAST, requerirían un gran poder computacional y cantidades de tiempo exorbitantes.

A partir de las secuencias en estas bases de datos, se han realizado estudios sobre la distribución de NRPS a través de los tres dominios de la vida en varias ocasiones, donde se

han llevado a cabo estudios genómicos extensivos que comprenden desde género hasta dominio. Los estudios más completos hasta la fecha representan la distribución de clústeres de NRPS que existen en cientos de cepas representativas de cada linaje mayor (Figura 14), dando un panorama general acerca de estos genes (80, 123).

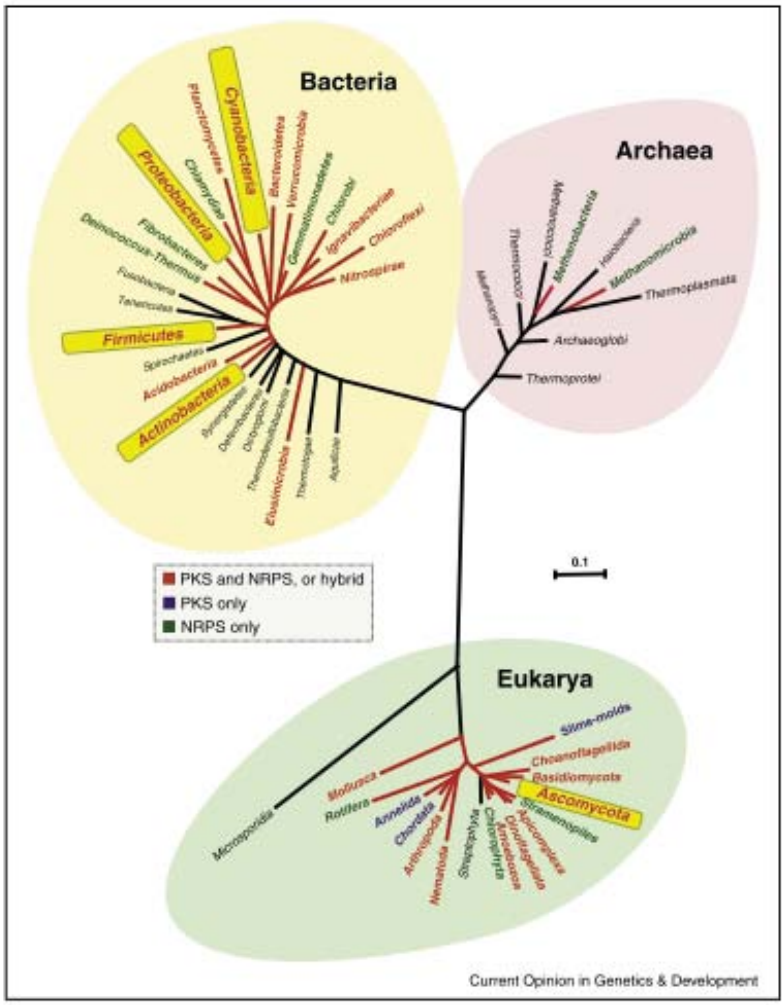

Figura 14. Árbol filogenético (16S y 18S) que ilustra los clados que contienen genes NRPS, PKS e híbridos. Los linajes que contienen PKSs y NRPSs, o enzimas PKS-NRPS híbridas se indican en rojo, los que contienen solo NRPS están en verde, y los que contienen solo PKS en azul. Aquellos Phyla con abundantes vías PKS/NRPS se enfatizan por un cuadro amarillo (tomado de referencia 120).

Existen varios estudios acerca de como los BGCs con arquitecturas modulares han evolucionado, y se han descrito algunos mecanismos que impulsan la variabilidad de este tipo de sintetasas (Figura 15), los cuales concuerdan en que existe más de un mecanismo que gobierna la evolución de estos genes. A pesar de todos estos estudios, poco se sabe acerca del intercambio de módulos, la generación de arquitecturas nuevas, o de la evolución del sitio de reconocimiento del substrato. No se ha hecho un análisis acerca de las relaciones entre módulos individuales, ya que, se han concentrado en genes completos o el BGC en su totalidad. Las relaciones estructurales y filogenéticas entre dominios aislados o módulos individuales, nos podrán acercar a entender estos mecanismos con mayor profundidad.

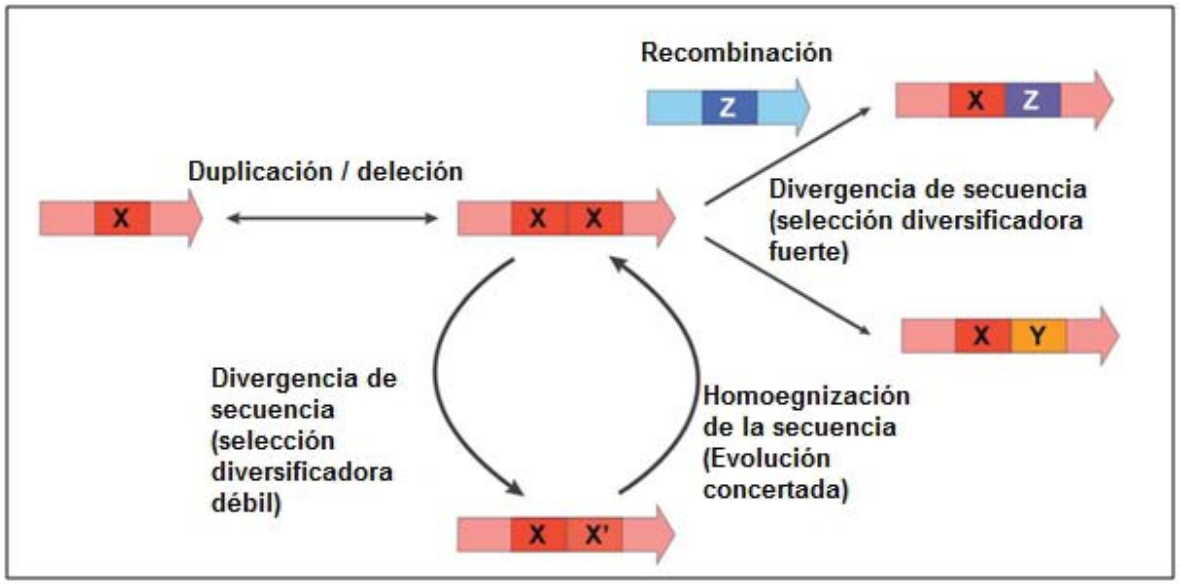

Figura 15. Modelo cualitativo de la evolución de dominios NRPS/PKS. Las secuencias quedan atrapadas en un ciclo donde la diversificación y la deriva génica son contrarrestadas por recombinaciones internas que homogenizan los dominios e impulsan la evolución concertada. Este ciclo puede romperse por recombinaciones externas o diversificación fuerte (adaptado de referencia 79).

Se han reportado en varias ocasiones que la transferencia horizontal de genes ha jugado un papel crucial para la distribución de las NRPS; un ejemplo que destaca es la transferencia de BGC a los hongos Ascomicetos a partir de las bacterias (80, 124), aunque existe incertidumbre al respecto (125). Inclusive dentro del dominio Bacteria se han identificado eventos de HGT entre ellas, un estudio genómico llevado a cabo en

cianobacterias reportó que se distinguen al menos 7% de genes sin homología directa dentro del Phylum (126).

Los estudios realizados hasta el momento se han concentrado mayormente en los genes o clústeres completos, donde pocos han tratado de encontrar un reloj molecular efectivo para el análisis de la historia evolutiva de los genes NRPS. En este estudio se analizarán las NRPS a partir del dominio C-starter, así como, por cada módulo individual. Lo cual nos dará las herramientas para poder reconstruir su historia evolutiva.

5. Justificación

Para la correcta detección de genes involucrados en vías metabólicas con relevancia biotecnológica o ecológica, de genes regulatorios, de genes con relevancia en la patogenicidad de las cepas, así como genes con posibilidades de ser blancos farmacéuticos, o de genes involucrados en la síntesis de productos naturales, etc., es necesario el uso y desarrollo de nuevos algoritmos tanto eficientes, como con gran sensibilidad. Dicho esto, este estudio pretende ampliar el conocimiento acerca de las NRPS, así como simplificar su descubrimiento y anotación funcional, diseñando un pipeline para su reconocimiento dentro de cualquier tipo de base de datos de secuencias, la predicción de su respectivo producto y si se es posible el análisis del BGC en el que se encuentra.

Al analizar los genes NRPS presentes en Bacterias y determinar sus relaciones filogenéticas y estructurales, será posible entender los mecanismos evolutivos por los cuales ha surgido la inmensa diversidad de afinidades a diferentes substratos y arquitecturas genéticas. Ésta información podrá ser utilizada para llevar a cabo ingeniería genética de NRPS, que a su vez, permitirá la investigación de la estructura y función de los péptidos no ribosomales, así como para la optimización del descubrimiento y predicción de nuevas NRPS.

6. Objetivos

Objetivo general

Realizar una búsqueda de proteínas NRPS en los proteomas de referencia no redundantes de la base de datos UniProt, para llevar a cabo un estudio filogenético y estructural de las NRPS a través del Dominio Bacteria.

Objetivos particulares

1. Reconocimiento de secuencias NRPS con mayor sensibilidad.
2. Construir una base de datos de proteínas NRPS.
3. Estudiar las relaciones evolutivas entre las NRPS.
4. Caracterizar bioinformáticamente la arquitectura modular de las NRPS y predecir los substratos de cada uno de los módulos dentro de cada proteína.
5. Analizar las relaciones estructurales y filogenéticas de proteínas NRPS.

7. Metodología

7.1 Búsqueda bibliográfica de NRP

Se buscaron lipopéptidos sintetizados por NRPS en la base de datos NORINE http://bioinfo.lifl.fr/norine, la cual recopila la información de más de 1,100 NRP (127). Se comprobó que existieran datos estructurales y que cada producto que sea efectivamente un lipopéptido. Como semillas, se eligieron las secuencias de las proteínas NRPS responsables de la síntesis de los lipopéptidos en la Tabla 1.

7.2 Obtención y alineamiento de secuencias

Las secuencias de las proteínas se obtuvieron en la base de datos de secuencias de proteínas UniProt. El programa MUSCLE v3.8.31 (128), el cual se puede obtener en http://www.drive5.com/muscle, fue utilizado alinear las secuencias de aminoácidos, usando los parámetros por defecto. Este programa fue utilizado para generar los alineamientos con los cuales se generaron los perfiles HMM. Para alineamientos de más de 500 secuencias, MUSCLE se vuelve inconsistente con los resultados, y no es recomendable usarlo. Por ésta razón, el programa Clustal Omega v1.2.4 con parámetros defecto (129), el cual es más apropiado para alineamientos grandes, fue utilizado para generar los alineamientos utilizados para construir los arboles filogenéticos, se puede obtener en http://www.clustal.org/omega.

7.3 Construcción de perfiles HMM

A partir de las secuencias obtenidas por la búsqueda bibliográfica, se escindieron las secuencias de cada módulo inicial (aproximadamente los primeros 1050 aminoácidos que comprenden los dominios C-starter, A_1 y T_1) del primer gen NRPS en la línea de ensamblaje. De la misma manera, se escindieron las secuencias únicamente del dominio C-starter de cada gen (~450 aa). Se alinearon con MUSCLE, descrito anteriormente. Para generar los perfiles HMM se usó el programa *hmmbuild* del paquete HMMER v3.1b2 con los valores predeterminados (130) al cual se le alimenta con alineamientos de secuencias homólogas, el paquete HMMER se puede obtener en http://hmmer.org/download.html.

Se generaron dos perfiles HMM, el perfil del módulo inicial (pMI) y el perfil del dominio C-starter (pCs).

7.4 Búsqueda de NRPS

Se obtuvieron los proteomas de referencia de Bacterias en UniProt versión de Noviembre 2017 (128). Se puede acceder en https://www.uniprot.org/proteomes haciendo una búsqueda con los filtros *redundant:no AND reference:yes AND taxonomy:"Bacteria [2]"*. Los 8,586 proteomas de referencia de bacterias han sido curados y anotados, además de ser no redundantes, cubren los Phyla bacterianos más representativos y las especies con mayor relevancia científica. Para la búsqueda de proteínas NRPS que puedan existir dentro de los proteomas se utilizó el programa *hmmsearch* del paquete HMMER v3.1b2, usando el pMI como perfil semilla para reconocer NRPS. Para este estudio se mantuvo un umbral de puntuación mínima de 900 para cada secuencia reconocida (opción *--incdomT*, "bit-score per-domain inclusion threshold"), el cual se impuso para evitar falsos positivos. Con los genes resultantes, se construyó una base de datos para su análisis posterior, para esto se usó la aplicación web de la UniProt para mapear la información relacionada a cada proteína: como la filogenia de la especie a la que pertenece; su secuencia completa; los dominios que contiene; longitud de la proteína, etc.

7.5 Construcción de árboles filogenéticos

Para generar los árboles filogenéticos por máxima verosimilitud, se utilizó el programa IQTree v1.3.3 (131) usando los valores predeterminados con un análisis "bootstrap" de 1000 iteraciones para construir la filogenia de los alineamientos arrojados por Clustal Omega, el programa se puede obtener en http://www.iqtree.org/#download. Se usó la dependencia ModelFinder incluida en el programa IQTree (132) para la búsqueda de modelos, la cual calcula el mejor modelo de substitución de aminoácidos para los genes en particular. El modelo recomendado para las NRPSs fue LG+F+I+G4, conforme al puntaje BIC (Criterio de Información Bayesiano). Para la visualización y edición de los árboles filogenéticos se utilizó la herramienta web interactiva iTol v3 (133), se puede acceder a esta herramienta en https://itol.embl.de.

7.6 Predicción de substratos

Para la predicción del substrato de cada dominio de adenilación, se utilizó el programa NRPSpredictor2 (91), el cual se puede obtener en https://github.com/roettig/NRPSpredictor2. NRPSPredictor2 utiliza perfiles HMM de una región conservada del dominio de adenilación, la cual se ha reportado que es responsable de la afinidad a cada substrato. Estas firmas se crearon a partir de las diferentes afinidades a substratos de genes ya caracterizados, las cuales son específicas de cada substrato. El resultado de cada módulo se añadió a la base de datos para su análisis cuantitativo, filogenético y estructural.

7.7 Análisis de BGC y de la arquitectura de los genes NRPS

El análisis del contexto genómico de los genes NRPS en cada especie de interés se realizó con el programa AntiSMASH v3.0.5.1 (112), el cual predice los BGCs dentro de cada genoma, se puede obtener en https://antismash.secondarymetabolites.org/#!/download. Por otro lado, se utilizó la herramienta web SMART (134) en donde se analiza cada proteína para observar su arquitectura de los dominios proteicos, se puede acceder en http://smart.embl-heidelberg.de/. Para analizar los genes adicionales dentro de un BGC se utilizó BLAST y la base de datos de dominios conservados de la NBCI (CDD) (108).

7.8 Cálculo del mapa de calor y las redes de similitud de secuencias

Para generar una matriz de similitud 'all-vs-all', se usó el programa Clustal Omega, con valores default y las opciones *--distmat-out* y *--full*, utilizando 1,000 secuencias de módulos individuales seleccionadas al azar. Con la matriz de similitud se generó un mapa de calor usando el paquete 'heatmap.plus' del programa R v3.4.1.

La construcción de la red de similitud de secuencias (SSN) se utilizó la herramienta web EFI-EST (135), que genera la red de similitud a partir del E-value dado calculado de la similitud entre cada par de secuencias, se puede utilizar esta herramienta en https://efi.igb.illinois.edu/efi-est/. Se utilizaron las secuencias de 10,323 módulos de 5,687 genes NRPS provenientes de 1,417 especies de bacterias, estas secuencias fueron aquellas

encontradas con el método de búsqueda con el perfil HMM del módulo inicial. Una conexión exitosa, que significa que el programa encuentra una similitud entre dos secuencias únicas, y procede a crear una conección o "Edge" entre ellas, estas conexiones son generadas a partir de un valor de probabilidad o "e-value" que es calculado por la probabilidad de que una secuencia aleatoria contenga la misma secuencia (entre más pequeño sea este valor, se puede asumir que las secuencias tienen mayor probabilidad de tener un ancestro común inmediato). Se filtran las conexiones (edges) a partir de un valor mínimo lo que genera clústeres de similitud: para este caso, se utilizó un límite de e-value de 1e-345 para definir las agrupaciones de secuencias, debido a su alta conservación (con un límite mayor no se lograba una buena distinción entre las redes de similitud). El SSN fue visualizado y personalizado en el programa Cytoscape 3.6.1 (136), el cual se puede descargar en http://www.cytoscape.org/download.php.

8. Resultados

8.1 Búsqueda de NRPS

Una búsqueda utilizando el programa hmmsearch del paquete HMMER, se compone de la base de datos y el perfil con el cual se llevará a cabo la búsqueda. El programa arroja los resultados con mayor significancia ('e-value') y le otorga un puntaje ('bit score'), que representa el grado de parentesco de la secuencia blanco al perfil HMM. Al comparar los resultados se distingue que las proteínas que efectivamente son NRPS, tienen un e-value de menos de 1e-200, al distinguir aquellas proteínas que contienen los tres dominios básicos (C-A-T). Cabe destacar que el e-value de los falsos positivos es de más de 1e-80, con lo cual se pueden descartar fácilmente. Algunas proteínas que se consideran falsos positivos, son las sintetasas de acetil CoA y las ligasas de ácidos grasos de cadena larga, que al igual que el dominio de adenilación de las NRPS pertenecen a la familia de ligasas/sintetasas dependientes de AMP, a lo que se debe que se obtengan esos valores de similitud, pero no contienen al dominio de condensación, por lo que no pueden ser clasificados como NRPS. Otra forma de distinguirlos es por su longitud, ya que, una proteína NRPS comprende al menos 1050 aa (un módulo promedio), y los falsos positivos varían entre los 650 aa, esto a su vez asegura que solo se incluyan las NRPS con al menos un módulo completo.

Para poder determinar la efectividad y sensibilidad de los perfiles HMM generados a partir de las secuencias de los Módulos Iniciales y los dominios C-starter. Se llevaron a cabo búsquedas a diferentes niveles, la primera comprende los proteomas de los tres Dominios de la Vida de UniProt, para observar el alcance de los perfiles HMM. La búsqueda con el pMI) dio como resultado más de 38,000 proteínas NRPS (cutoff: Bit Score = 700), donde se reconocieron NRPS pertenecientes a Bacterias (36,000), Hongos Ascomycetos (2,000) y unas pocas Arqueas.

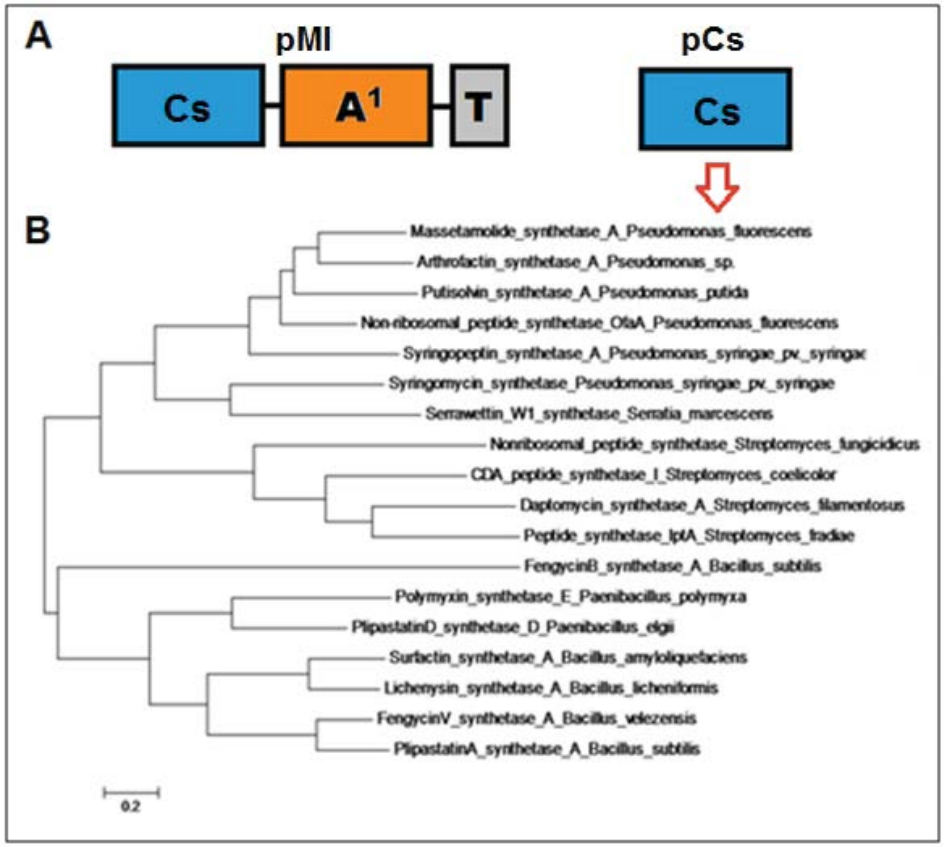

Figura 16. Construcción de los perfiles HMM. A, Esquema de las regiones de las NRPSs usadas para construir los perfiles HMM: Perfil Mi, perfil construido a partir del alineamiento de 230 secuencias del módulo inicial (~1050 aa) de NRPS que contienen un dominio C-starter con motivos característicos, curado manualmente; pCs, perfil construido a partir del alineamiento de 18 secuencias (~450 aa) de NRPSs que se tenga caracterizado su producto como un lipopéptido. B, cladograma generado a partir del alineamiento de secuencias seleccionadas como representantes comprobados de NRPS que sintetizan lipopéptidos usado para construir el pCs.

Para los análisis posteriores se utilizó la base de datos de Proteomas de Referencia 'RefProt' de UniProt. Esta ha sido revisada y curada para evitar secuencias repetidas, fragmentadas, mal anotadas o con baja calidad de secuenciación. Asimismo, se utilizaron únicamente los proteomas pertenecientes al Dominio Bacteria y se aumentó el 'cut off' a un Bit Score de 900 para reducir la cantidad de falsos positivos o dominios aislados. RefProt contiene 8,586 proteomas de bacterias, el perfil del módulo inicial (Figura 16), se utilizó para reconocer todas las posibles secuencias NRPS dentro estos proteomas. Se recuperaron las secuencias completas de los genes con un Bit Score significativo. Con los

genes resultantes, se conformó una base de datos de secuencias de NRPS, así como la información relevante de cada proteína, conteniendo: su identificador (UniProtID); la cepa donde proviene; su taxonomía a nivel de Dominio, Phylum, Clase, Orden, Familia y Género; longitud de la secuencia; número de módulos y su arquitectura; así como la predicción de la afinidad de cada Dominio de Adenilación dentro de la proteína.

8.2 Análisis cuantitativo de NRPS encontradas

Se encontraron una mayor cantidad de genes NRPS en los Phyla con más representantes secuenciados en contraste con Phyla que poco representados, aunque aún así, la ocurrencia promedio de estos genes biosintéticos es de 16% en los proteomas analizados (Figura 17). El Phylum con mayor proporción de especies que presentan NRPS son las cyanobacterias, por otro lado, el género con mayor cantidad de genes NRPS es *Streptomyces* (Tabla 2). Esto puede estar sesgado por la cantidad de cepas que se han secuenciado en cada Género. Como ejemplo, considerando solo los Proteomas de Referencia de UniProt: en el género *Streptomyces* se tienen 231 proteomas; comparado con el Género *Nocardia* en el cual solo se tienen 14 proteomas conteniendo 1030 y 127 genes NRPS respectivamente (Tabla 2).

Un problema que se identificó en las bases de datos es la incorrecta anotación de este tipo de genes, frecuentemente anotadas por el nombre de la familia de uno de los dominios que contienen, o simplemente como proteína predicha sin caracterizar. Al considerar un módulo como un dominio proteico funcional, el perfil HMM MI generado en este estudio, se puede proponer para la identificación de genes NRPS, al tener la capacidad de reconocer este tipo de combinación de dominios con gran fidelidad. De igual forma se podrán generar perfiles HMMs para la correcta identificación de genes híbridos NRPS/PKS y PKSs.

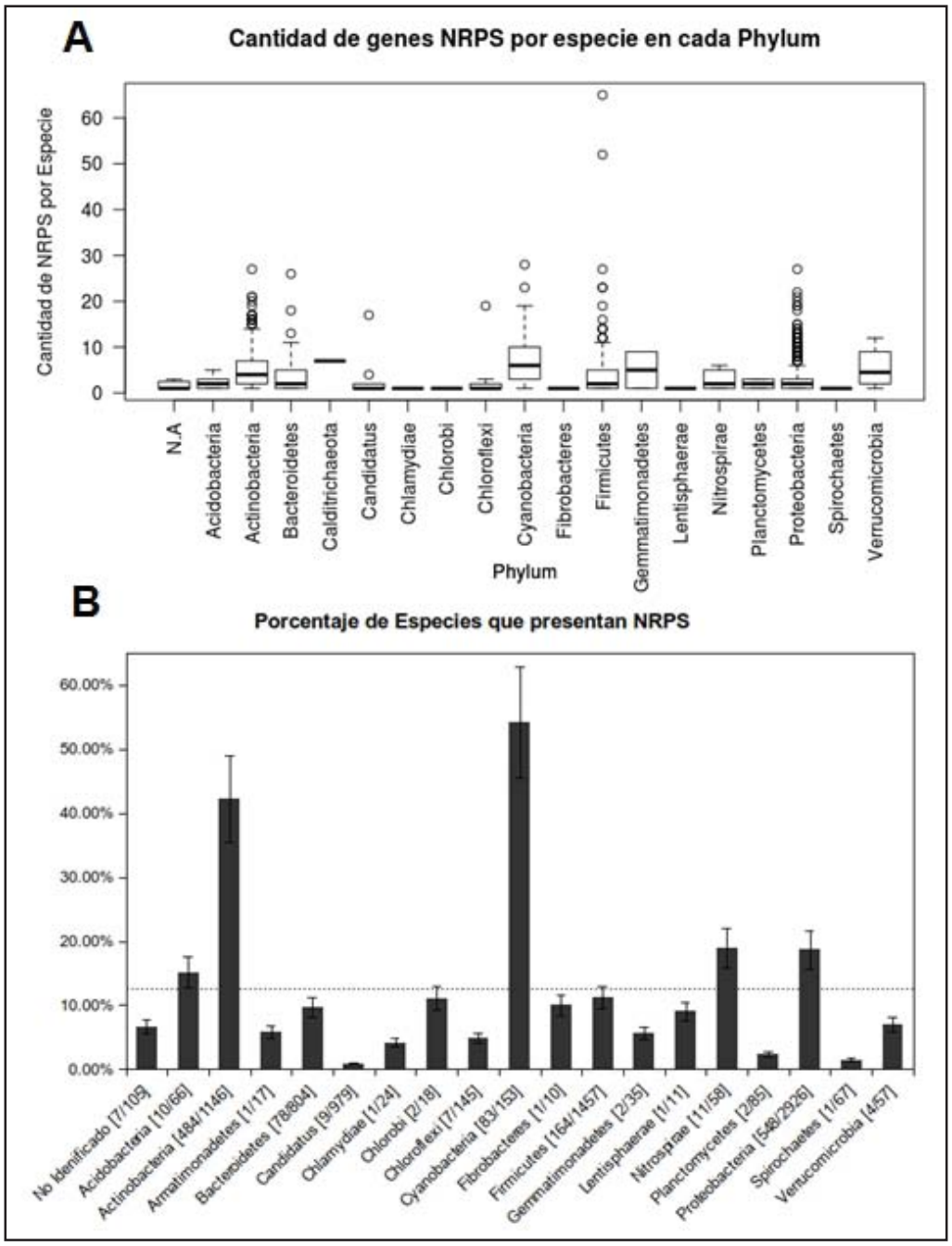

Figura 17. Análisis de las NRPS. A, gráfica de cajas que indica la cantidad de genes que contiene cada cepa de Bactetrias; B, gráfica que expone el porcentaje de genes NRPS encontrados en una búsqueda realizada en 8586 Proteomas de Referencia en UniProt, donde se identificaron 1417 proteomas que contienen genes NRPS. Barras de error calculadas con el error estándar. Línea punteada, proporción total dentro del Dominio Bacteria. [n Proteomas donde se encontraron NRPS/n total].

Tabla 2. Géneros con mayor abundancia de proteínas NRPS.

Género	Procentaje de proteomas donde se encontraron NRPS (%)	Proteomas con NRPS	Genes NRPS	Promedio de NRPS por cepa ± desv. est.
Tumebacillus	100	3	129	43 ± 22.5
Kutzneria	100	2	33	17 ± 1.7
Scytonema	100	3	46	15 ± 6.3
Tolypothrix	100	4	53	13 ± 2.9
Nostoc	90.9	10	123	12 ± 5.3
Actinokineospora	100	3	35	12 ± 2.7
Nocardia	100	11	127	12 ± 3
Xenorhabdus	100	11	117	11 ± 2.6
Brevibacillus	75	6	62	10 ± 7.6
Lysobacter	38.5	5	47	9 ± 1.6
Chitinophaga	66.7	8	72	9 ± 5
Rhodococcus	100	20	167	8 ± 2.9
Alloactinosynnema	100	3	24	8 ± 2.1
Pseudoalteromonas	35	7	55	8 ± 2.5
Amycolatopsis	90.9	10	74	7 ± 3.3
Lentzea	100	3	22	7 ± 1.3
Cyanothece	80	4	29	7 ± 2.2
Flavobacterium	11.9	7	49	7 ± 3.6
Saccharothrix	100	4	28	7 ± 1.9
Janthinobacterium	66.7	4	27	7 ± 5.7
Actinomadura	75	3	20	7 ± 1.5
Paenibacillus	54.5	36	232	6 ± 4.1
Micromonospora	94.7	18	114	6 ± 3.4
Planktothrix	100	3	19	6 ± 4.7
Kitasatospora	100	5	30	6 ± 3.7
Streptomyces	90	180	1030	6 ± 3
Nonomuraea	100	5	28	6 ± 1
Methylomonas	60	3	15	5 ± 2.7
Rheinheimera	80	4	20	5 ± 3.7
Saccharopolyspora	80	4	18	5 ± 1.3
Pseudomonas	62	44	194	4 ± 2.3
Actinoplanes	88.9	8	35	4 ± 1.9
Calothrix	100	4	17	4 ± 1.9
Niastella	80	4	17	4 ± 1
Bacillus	28.6	28	117	4 ± 3.1
Gordonia	100	16	65	4 ± 1.7
Pedobacter	25	6	24	4 ± 3.8
Rhizobacter	75	3	12	4 ± 0.7
Paraburkholderia	72.2	13	50	4 ± 4.2
Burkholderia	61.9	13	47	4 ± 3
Variovorax	100	12	41	3 ± 1.8
Mycobacterium	93.9	93	310	3 ± 1.8
Vibrio	29.3	12	35	3 ± 1.5

El género *Streptomyces* es el más secuenciado y a su vez es el que presenta más NRP caracterizados. En contraste, el género Tumebacillus (Tabla 2) recientemente descubierto en suelos de China, del cual se han secuenciado tres cepas, resalta por la cantidad exuberante de NRPS que presentan dos de ellas, con 66 y 53 genes, en cambio, la cepa restante solo tiene 10 NRPS, lo que puede señalar que estas cepas sufrieron una reciente expansión en la cantidad de genes NRPS. Al analizar estas cepas, y las relaciones evolutivas de sus genes, se podría identificar el mecanismo evolutivo por el cual adquirieron tan alta cantidad de genes, ya sea por varias duplicaciones, HGT, o una combinación de ambas. Otros Géneros con potencial para producir NRP que han sido poco estudiados son: *Nocardia, Rhodococcus, Paenibacillus, Brevibacillus, Tholypotrix, Xenorhabdus, Flavobacterium, Myxococcus*, Scytonema, *Actinokineospora, Chondromyces, Kitatospora,* entre otros. Dado su alta cantidad de genes NRPS y su conservación a través de sus especies, es probable que estos organismos dependan de los NRP para defensa, competencia por nutrientes u otras funciones biológicas.

8.3 Filogenia a partir del dominio C-starter

Con base a la literatura y en resultados observados con MEME sobre los motivos conservados de los Dominios C-starter, se propone la hipótesis que los dominios C-starter, al tener en común un solo substrato (un ácido graso), no han sido influenciados por la gran presión de selección que sufre cada dominio de elongación hacia sus respectivos substratos. Aunque es necesario un análisis más profundo para determinar el modelo de presión de selección que afecta a los dominios NRPS, ya sea por un modelo neutro o un modelo positivo de selección. Con esto en mente, se propone la hipótesis de que los dominios C-starter pueden ser usados como marcadores filogenéticos de las NRPS, para poder observar los eventos evolutivos que han ocurrido a lo largo de la historia de estas bacterias. Para esto se realizó un tratamiento de las secuencias NRPS encontradas para una nueva búsqueda con el pCs.

Los resultados de esta búsqueda son los dominios C-starter más probables dentro de la base de datos, descartando dominios de condensación intermedios o de epimerización.

De estos resultados se tomaron 1,600 secuencias con más alto puntaje, con un promedio de 450 aminoácidos de longitud. Se llevó a cabo una limpieza de secuencias redundantes con el programa CDHit (valores default) a 95% de identidad y una curación manual de aquellos que no presentaban los motivos característicos del dominio C-starter. A partir de una muestra de 1,000 secuencias resultantes, se realizó un alineamiento múltiple usando Clustal Omega, con el cual se construyó el árbol filogenético por máxima verosimilitud con el programa IQtree (Figura 18).

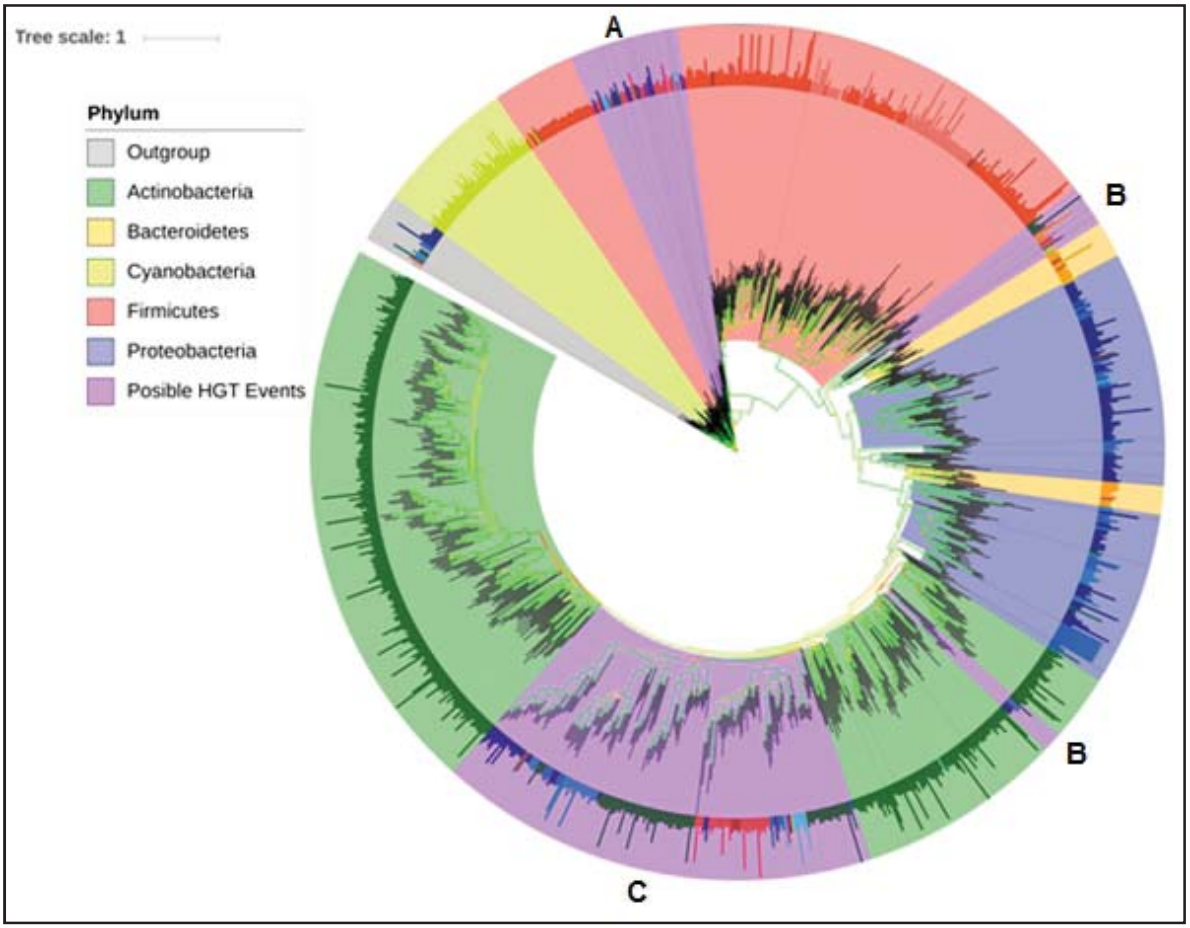

Figura 18. Árbol filogenético generado a partir del alineamiento de 1000 secuencias de dominios C-starter junto con 11 secuencias de dominios $_{L}C^{L}$ como outgroup (Análisis 'bootstrap' de 1000 iteraciones). Eventos de transferencia horizontal de genes: A, eventos tempranos; B, Eventos aislados; C, eventos tardíos. Árbol completo disponible en https://itol.embl.de/tree/13224820720819108152666275 7

En éste cladograma se observa que los dominios C-starter evolucionaron a partir de dominios $_{L}C^{L}$ en cianobacterias, al ser más parecidos al grupo externo (dominios $_{L}C^{L}$) que a los C-starter actuales. Posteriormente se especializaron como dominios C-starter. Por eventos de HGT se dispersaron por los demás taxones bacterianos. Estos eventos de HGT, se separan en tres tipos: El primero, eventos tempranos de HGT, donde genes NRPS pertenecientes a Cianobacterias fueron transferidos a ciertas bacterias de Firmicutes y Proteobacteria (Figura 18, A); el segundo, eventos aislados de HGT como aquellos donde cepas de Bacteroidetes adquirieron algunos NRPS, o se compartieron NRPS entre especies pertenecientes al Phylum Actinobacteria y cepas del género *Vibrio* (Figura 18, B); El tercero y más abundante, son eventos de intercambio masivo de genes NRPS entre especies de los Phyla Proteobacteria, Actinobacteria y Firmicutes (Figura 18, C). Ya se tenían indicios de la posibilidad de que las NRPS se habían dispersado a través de las bacterias y hacia los Hongos por HGT, aunque no ha sido posible comprobar y se ha debatido ampliamente si los resultados que se han obtenido en análisis filogenéticos de este estilo pueden tomarse como evidencia (125). Por lo que son necesarios análisis adicionales para corroborar los eventos de HGT, en un protocolo establecido en (137) en el que se utilizó no solo el porcentaje de GC, sino también el índice de adaptación de codones, así como la preferencia de codones para generar la evidencia suficiente para confirmar o refutar existe un evento de HGT.

Como un estudio de caso, se identificó un BGC caracterizado en *Pectobacterium atrosepticum*, al cual se realizó un análisis del contenido de G-C (138). Analizando el vecindario genómico con ± 15000 nucleótidos que flanquean al BGC (Figura 19A), lo cual puede indicar si una región de DNA es ajena a la especie y fue adquirido por un evento de HGT. Se encontró una diferencia significativa en el contenido de G-C con respecto al vecindario genómico de los genes NRPS (Figura 19B). Un análisis con AntiSMASH del clúster biosintético identificó un BGC semejante en la proteobacteria *Xenorhabdus doucetiae*, aislada del tracto digestivo de un nemátodo entomopatogénico. Se observa que hay diferencias en la arquitectura y cantidad de genes, donde se identifica una posible unión de genes en *P. atrosepticum*. Otro hecho que apoya esta hipótesis, es que existe un

tRNA para el aminoácido asparagina flanqueando al primer gen NRPS, este locus podría estar involucrado en la capacidad para intercambiar genes por transferencia horizontal. Lo cual se ha reportado que los loci para tRNA frecuentemente se encuentran flanqueando islas de patogenicidad y existe la teoría son usados como anclaje para la recombinación de elementos móviles, así como ciertos bacteriófagos que usan secuencias de tRNAs como sitios de inserción en el genoma hospedero (139).

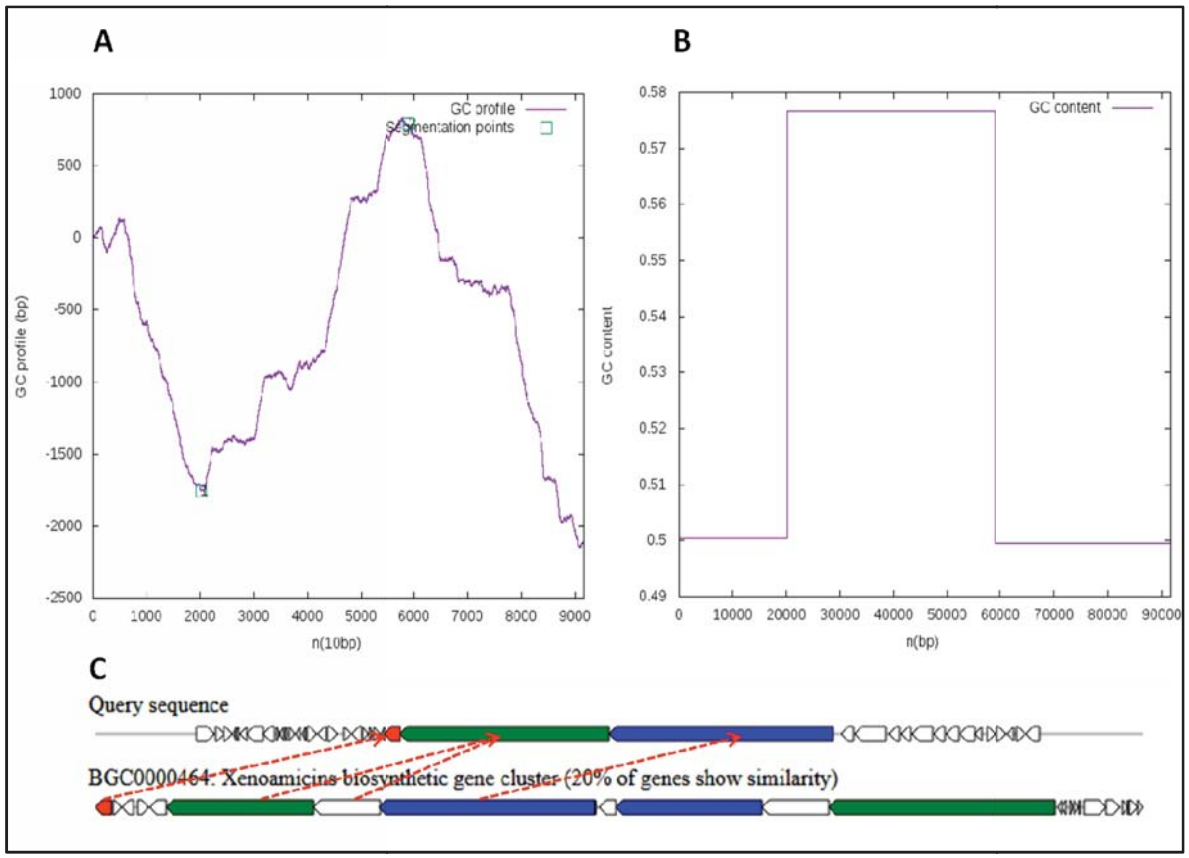

Figura 19. Posible evento de transferencia horizontal de un BGC caracterizado en la bacteria patógena de la papa, *Pectobacterium atrosepticum* SCRI1043. A-B, Análisis del contenido G-C de la región genómica entre los nucleótidos 1664805:1756396, donde el punto de inflexión del contenido de G-C cambia en la secuencia de los genes NRPS. C, un análisis con AntiSMASH reveló un BGC homólogo en la bacteria *Xenorhabdus doucetiae*. Flechas rojas, posible dirección de la transferencia.

Al compartirse estos genes entre diferentes especies, alude al hecho que existen mecanismos por los cuales estos genes llegan a intercambiarse entre organismos, incluso siendo filogenéticamente lejanos, proveyendo una ventaja evolutiva a la cepa receptora.

Aunado a el hecho de que estos genes se extienden por muchas de las especies de los Phyla más abundantes que se contrasta con Phyla donde hay poca ocurrencia de genes NRPS dentro de sus especies. Estos hallazgos sugieren que los NRPS de los últimos mencionados fueron adquiridos por HGT, siendo una hipótesis más plausible a que haya habido una pérdida masiva de genes en los taxones con poca ocurrencia de NRPS. Esto apoya la hipótesis de que existe un "pool" genético' de módulos NRPS, que sustenta la formación de nuevas arquitecturas modulares (Figura 20), de la misma forma que se fomenta la exploración de las estructuras de los péptidos producto por un mecanismo de intercambio de módulos y de remodelación de arquitecturas de proteínas NRPS (126). Otro tipo de genes y vías metabólicas que se ha reportado que tienen una alta frecuencia de HGT son las aclamadas islas de patogenicidad (140), las cuales pueden conferir a las cepas receptoras ventajas evolutivas como resistencia a antibióticos, capacidades metabólicas nuevas, genes de síntesis de metabolitos de defensa, producción de biofilm, genes relacionados con la adherencia o la matiz extracelular, etc.

8.4 Análisis de las arquitecturas modulares

Existe una ocurrencia común de ciertas arquitecturas modulares, analizando las NRPS pertenecientes al Phylum Firmicutes (Figura 20), se observa que esta ocurrencia común converge en una base genética de arquitecturas: [A-T-C], [C-A], [C-A-T-C], [C-A-TE] (1 módulo); [C-A-T_C-A-T], [C-A-T_C-A-T-C], [C-A-T_C-A-T-TE], [C-A-T_E-C-A-T] (2 módulos); [C-A-T_C-A-T_C-A-T], [C-A-T_C-A-T_C-A-T-C] (3 módulos); [C-A-T_C-A-T_C-A-T_C-A-T-C], [C-A-T_C-A-T_C-A-T_C-A-T-TE] (4 módulos). A partir de esta base genética, se han generado arquitecturas más complejas principalmente por mecanismos de combinación de arquitecturas por unión de genes. Mecanismos de remodelado e intercambio modular por recombinación de regiones homólogas y la deleción de dominios únicos por perdida de la función podrían haber jugado un papel en la expansión de la diversidad de arquitecturas modulares de las NRPS.

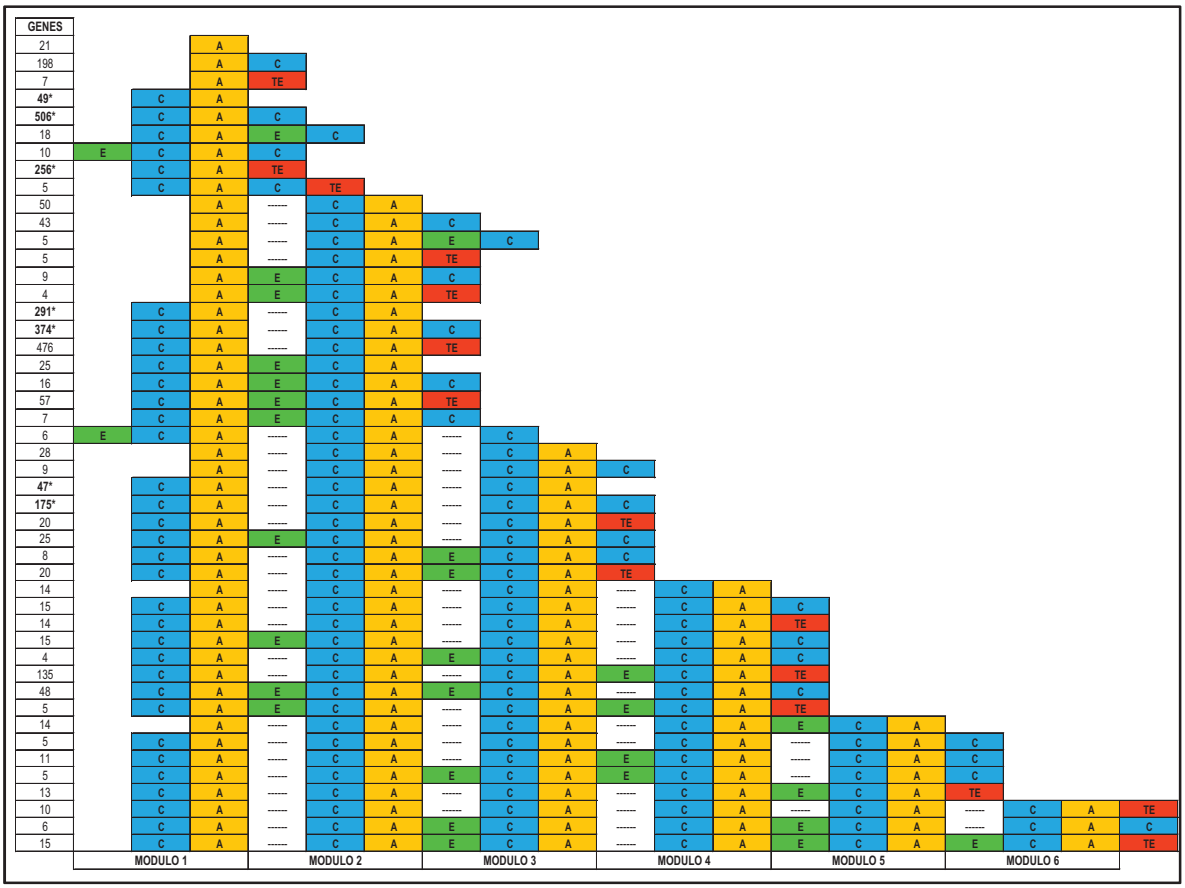

Figura 20. Variedad de arquitecturas modulares más comunes de NRPS en Firmicutes. Azul, dominio de condensación (C); Amarillo, dominio de adenilación y dominio T (A); Rojo, dominio tioesterasa (TE); Verde, dominio de epimerización (E). Asteriscos, arquitecturas más comunes.

Los arboles filogenéticos de las arquitecturas similares (Figura 21) sugieren que estas organizaciones modulares se han formado en distintas ocasiones, de la misma forma, las afinidades del dominio de adenilación han surgido como eventos de convergencia evolutiva, al tener diferencias en su secuencia y estructura, así como distintos caminos evolutivos, aunque sea afín al mismo substrato, por lo cual se asume que la afinidad tiene una gran plasticidad. Al combinarse en distintas ocasiones y con diferentes módulos, la historia evolutiva de las NRPS tiene una mayor frecuencia de intercambios de módulos o genes completos, que de dominios individuales o módulos incompletos. Lo cual podría deberse a que presentan secuencias conservadas flanqueando al módulo y/o al dominio, funcionando como locus para la recombinación homóloga, permitiendo así la movilidad entre módulos completos y dominios individuales.

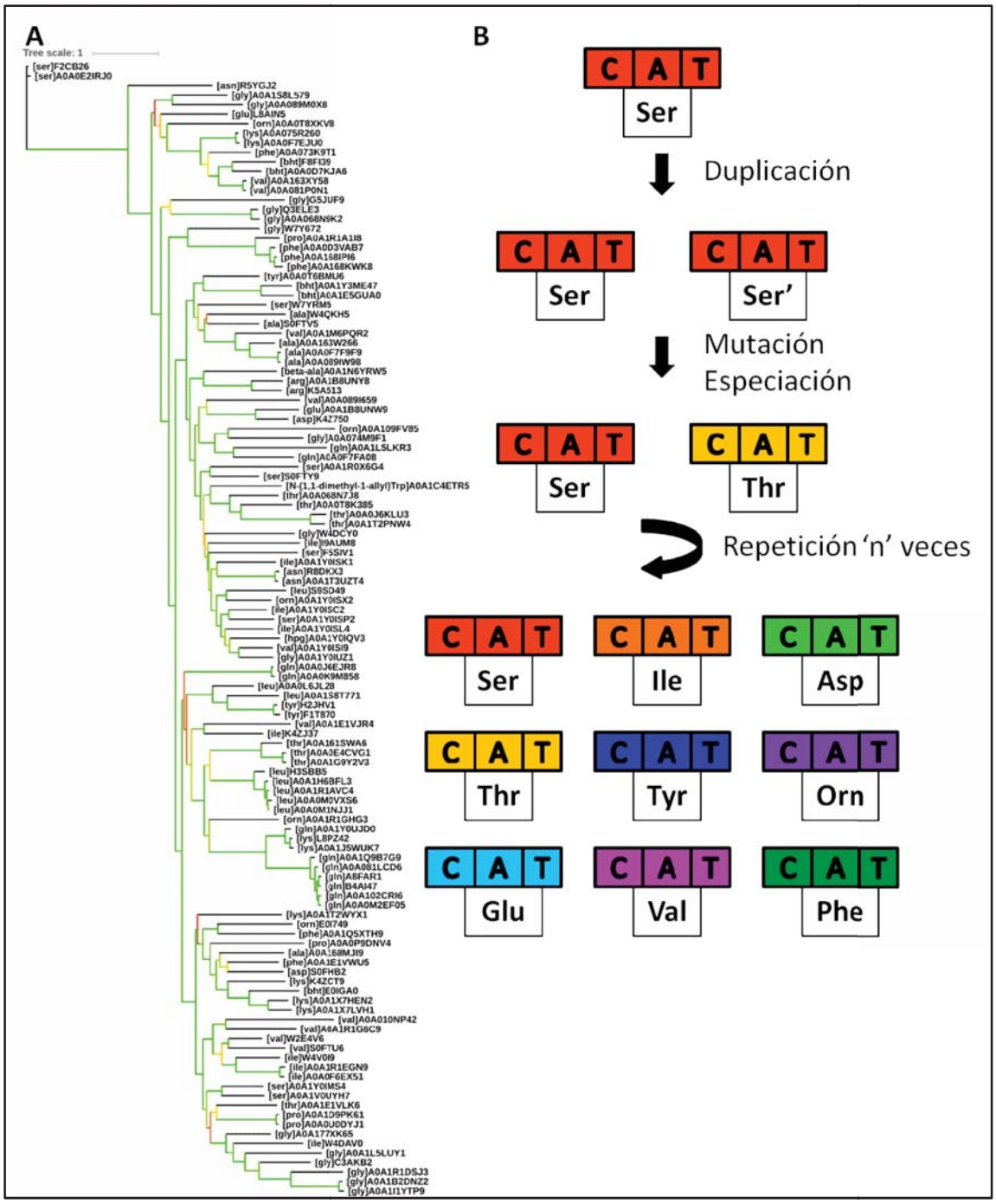

Figura 21. Mecanismo evolutivo por el cual ha surgido la diversidad de módulos NRPS con diferentes afinidades de substratos en NRPS. A, árbol filogenético generado a partir de las secuencias de NRPS con la arquitectura modular [C-A-T]. B, esquema de los mecanismos evolutivos por los cuales ha surgido la diversidad de afinidades a diferentes aminoácidos en NRPS unimodulares.

8.5 Análisis de módulos individuales y la predicción de sus substratos

Con base al análisis de la secuencia de firmas conservadas caracterizadas para cada afinidad de dominios de adenilación con substratos ya caracterizados, en lo que se concentra el algoritmo de NRPSpredictor2, se realizó la predicción de los substratos afines a cada módulo en la base de datos de NRPS. Teniendo como substratos más comunes a los aminoácidos serina, treonina, glicina, fenilalanina y valina (Figura 22).

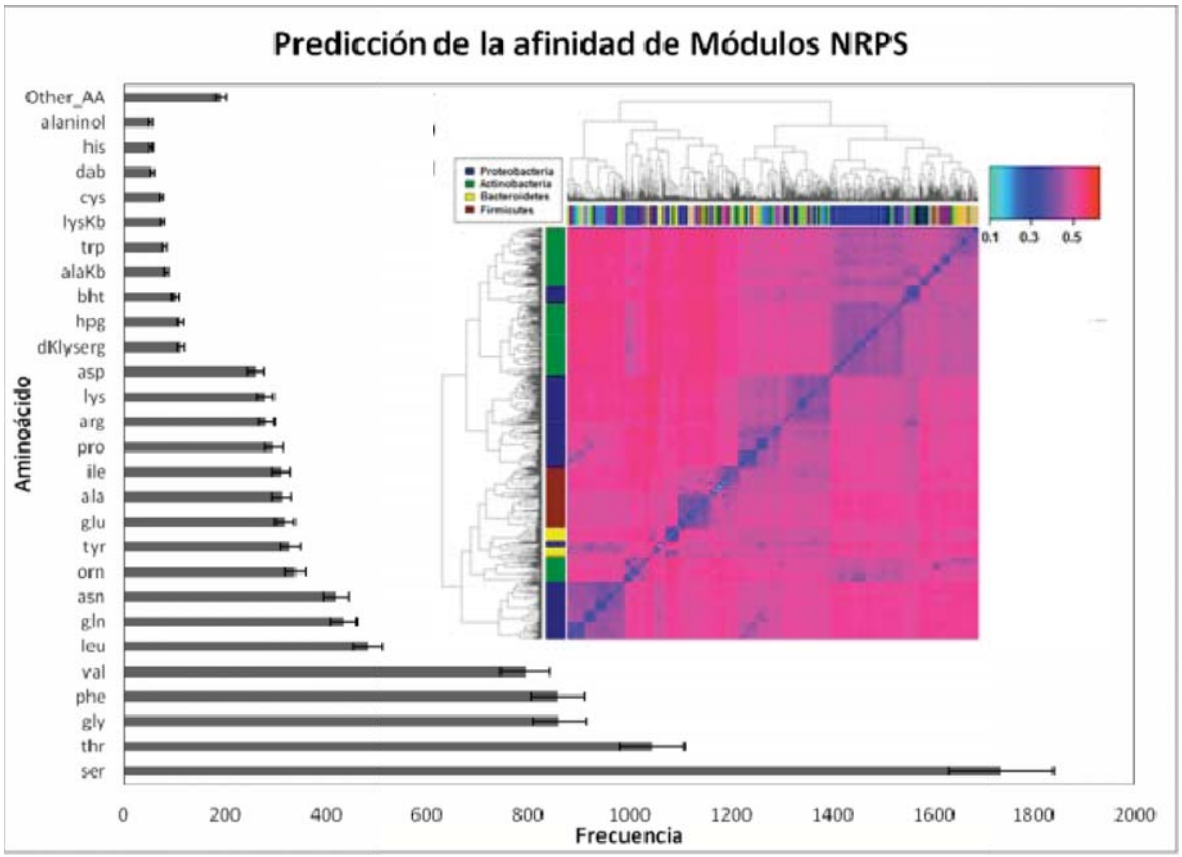

Figura 22. Análisis cuantitativo de los resultados de la predicción de aminoácidos de módulos NRPS (n = 10,323 módulos de 5,687 NRPS provenientes de 1,417 especies de bacterias) A, Gráfica de la frecuencia de los substratos afines a cada módulo analizado; barras de error calculadas con el error estándar. Heatmap construido a partir de una matriz de distancia realizada con un BLAST All-vs-All de 1000 módulos NRPS seleccionados al azar. Barra lateral, los colores indican el Phylum al que pertenece. Barra superior, los colores indican el substrato al que es áfin el dominio de adenilación.

Con base a la predicción de substratos de los dominios de Adenilación de cada gen NRPS, sin tomar en cuenta el BGC completo, se identificaron 1192 diferentes posibles productos con diferente secuencia de aminoácidos, donde la variedad de la secuencia del péptido producto es inversamente proporcional a la cantidad de módulos del gen. Es decir, que una secuencia de aminoácidos en particular se vuelve menos frecuente con respecto a la el número de residuos del péptido, sugiriendo que la formación de NRPS con más de 5 módulos o arquitecturas complejas han sido eventos aislados, y por lo tanto estas arquitecturas modulares son polifiléticas.

Comprender cómo evolucionan los NRPS proporcionará un conocimiento básico para el éxito de la ingeniería de estas proteínas y similares. Integrando el conocimiento que proviene de este estudio, el dominio de condensación también tiene especificidad de sustrato, esta podría ser la razón por la que los enfoques de bioingeniería combinatorial y sintética han sido insostenibles, con este estudio existe la posibilidad de seleccionar los dominios correctos para una combinación específica de AA_n y AA_{n+1}. En el ejemplo digamos que queremos cambiar el segundo aminoácido en la proteína SrfA-A [Glu-Leu-Leu] de Leu a Thr, necesitamos un módulo que tenga un dominio A con la especificidad a la Treonina, pero también ambos dominios C colindantes capaces de catalizar la condensación de Glu_1:Thr_2 y Thr_2:Leu_3. Con esta perspectiva, es imperativo generar una base de conocimiento, así como una base de datos con secuencias representativas o secuencias consenso de módulos para cada combinación de AA.

Los resultados de la predicción de substrato de los dominios de adenilación (Figura 22) indican que la afinidad a la serina es la más abundante y que se encuentra ampliamente relacionada con los otros módulos. Se pueden clasificar los módulos NRPS por su afinidad y por su grado de conservación: el primer tipo son aquellos donde se observa una gran conservación, así como una gran cantidad de representantes con afinidad hacia el mismo substrato; el segundo tipo es aquel que hay gran conservación de la secuencia, aunque la afinidad a su substrato sea variable; el tercer tipo de módulos NRPS, son aquellos que presentan baja conservación, así como diferentes afinidades. Cada tipo podría tener implicaciones evolutivas diferentes, siendo el primer tipo de módulos los más

especializados y en donde existe una presión selectiva mayor. Lo cual podría estar relacionado a que ese aminoácido está jugando un papel importante en la estructura y función específica del NRP producto. El segundo y tercer tipo representan a los módulos que están bajo menos presión selectiva. Su substrato, por lo tanto, no tendría un rol concreto en la función o la estructura del péptido producto. Más bien, este tipo de módulos tendrían un papel evolutivo, siendo los responsables de la gran variabilidad de NRP, al estar actuando como 'comodines', explorando diferentes estructuras e innovando la función biológica del péptido.

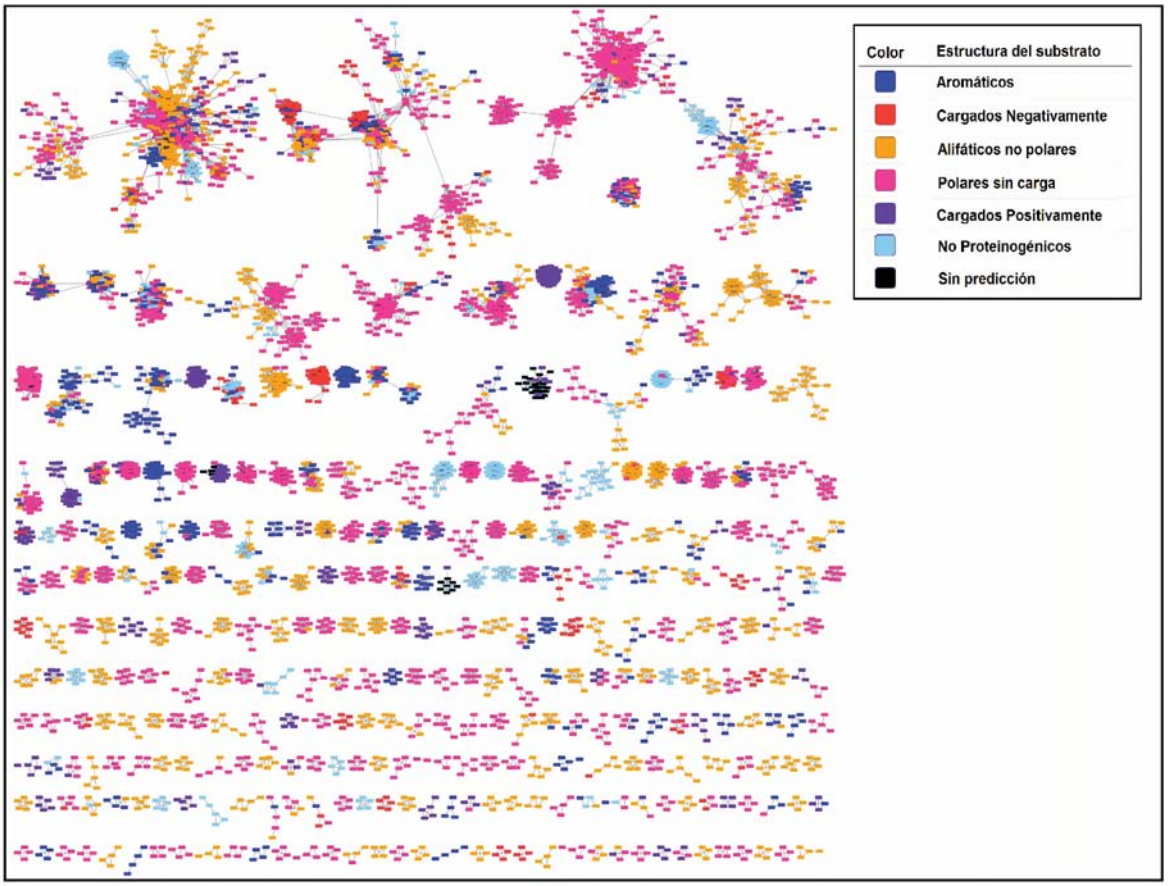

Figura 23. Red de similitud de secuencias (SSN) de 10,323 módulos individuales (C-A-T) analizados en este estudio, representando la relación secuencia/substrato de los módulos NRPS. Los colores indican la predicción de la afinidad hacia el respectivo substrato. Cada conexión tiene un e-value menor a 1e-345. Existen módulos donde no fue posible la predicción del substrato (negro).

Como ya se mencionó antes, se necesitan análisis más profundos para corroborar que modelo de selección está afectando a cada tipo de módulo, donde una selección purificadora podría estar actuando en el primer tipo de módulos, eliminando los alelos que sufran mutaciones de la población. En cambio, el segundo y tercer tipo de módulos NRPS tendrían una proporción de sitios bajo selección menor comparada con el primer tipo, aludiendo al hecho de que se encuentran bajo una presión de selección moderada o baja.

En el árbol filogenético del Dominio C-starter se observa que han sucedido bastantes eventos de intercambio genético entre taxones bacterianos, un resultado similar se observa en el SNN agrupado por Phylum (Figura 24), donde las conexiones entre los taxones se observan por la similitud de la secuencia de algunos módulos. Curiosamente, no se comparten módulos intermediarios entre Actinobacterias y Cianobacterias, ni se identificaron secuencias similares a aquellas pertenecientes a los Bacteroidetes.

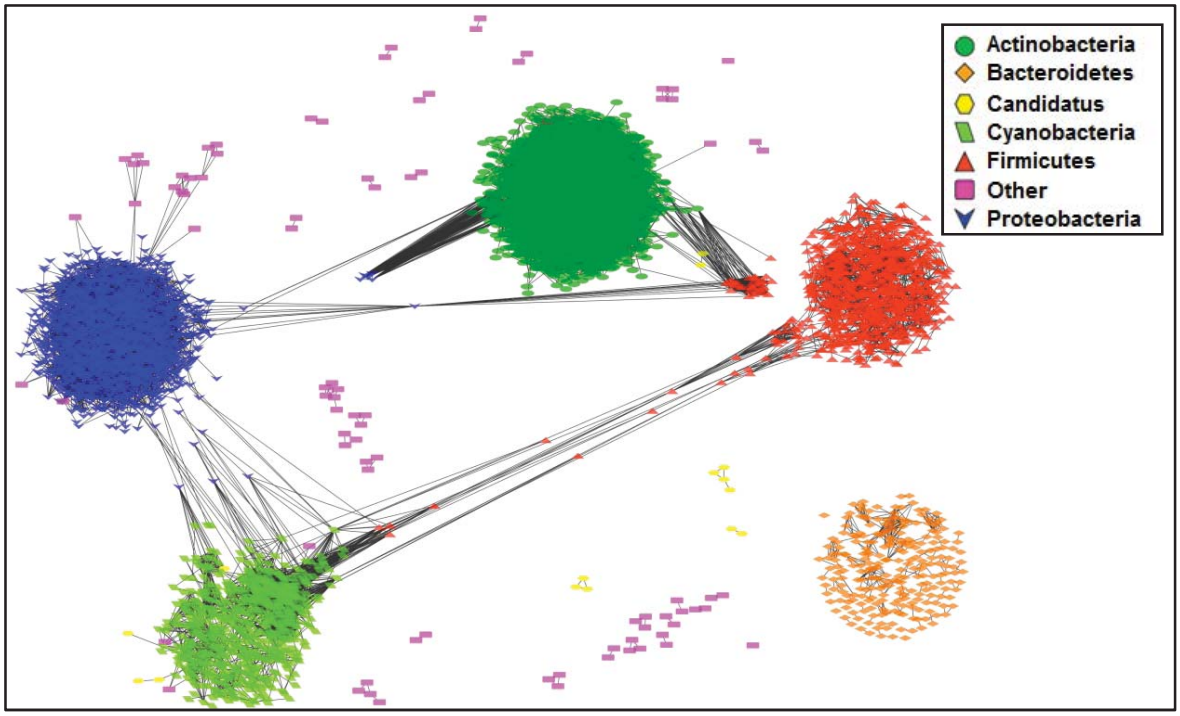

Figura 24. Red de similitud de secuencias de módulos NRPS agrupados por Phylum. Se observan los módulos que actúan como interconexiones entre Phyla bacterianos, los cuales conectan cuatro de los Phyla más abundantes, a excepción de los Bacteroidetes (Naranja). Realizado con setsApp (141) para Cytoscape 3.0.

Por otro lado, el SSN agrupado por Phylum indica que se conserva la estructura dependiendo de la afinidad que se tiene a cada substrato, con excepción de las NRPS pertenecientes a cianobacterias donde las secuencias se han conservado bastante, independientemente de su correspondiente afinidad, de la misma forma, al observar que las cianobacterias comparten conexiones con los otros Phyla bacterianos, así como el hecho de que presentan la mayor proporción de especies con genes NRPS, se presenta la hipótesis de que en determinado momento las cianobacterias fueron las que originalmente desarrollaron estas vías metabólicas, y consecuentes eventos de HGT las han esparcido por los demás lados bacterianos (Figura 24).

Al haber un pool genético de módulos, estos pueden intercambiarse dentro del BGC, con esta información podría generarse secuencias consenso para cada substrato. Es necesario un estudio más a fondo de las regiones conservadas al inicio y al final de los módulos NRPS, donde podría encontrarse una región especializada para reconocerse por la maquinaria de recombinación, funcionando de forma análoga a un alelo eucarionte, donde una pequeñas regiones entre cada módulo están conservadas para poder empalmarse entre ellas y llevar a cabo una recombinación precisa del módulo. Incluso, este intercambio horizontal de módulos podría ocurrir entre especies filogenéticamente distantes que estén ecológicamente relacionadas. De la misma forma, es necesario un estudio de las regiones Linker intramodulares, los cuales podrían jugar un papel importante en la compatibilidad entre diferentes módulos.

La predicción de los substratos puede ser errónea en ciertos casos, por lo que hace falta expandir el conocimiento acerca de cada afinidad, conglomerando datos bioquímicos, estructurales, y de la secuencia de los módulos y los genes dentro del BGC. Existen módulos que se desconoce el substrato al cual son afines (Figura 23, negro), esto puede deberse a que es afín a un substrato que no se ha identificado anteriormente, o a que no se ha incluido esta firma dentro de la base de datos del programa NRPSpredictor2. Al predecir la estructura de la región que reconoce al substrato, se podrá inferir su estructura.

9. Discusión y conclusiones

La meta principal de este estudio de secuencias genéticas de las NRPSs, es poder reconocer genes con la posibilidad de producir un nuevo NRP que tengan importancia biotecnológica, así como poder priorizar su estudio bioquímico, reduciendo así los costos de un análisis funcional de cepas productoras; asimismo, se espera poder identificar a los genes responsables de la biosíntesis de un péptido descubierto por métodos bioquímicos, a partir de un análisis metagenómico dentro de una comunidad bacteriana, en la cual exista la impotencia de aislar la cepa productora. Al comparar los resultados aquí presentados, con cualquier producto natural que se sospeche que es sintetizado por una NRPS, proveerá de pistas acerca de la posible bacteria productora, facilitando su identificación y aislamiento. Por otro lado, al duplicarse un gen NRPS podría conferir a esta bacteria ancestral, una ventaja evolutiva al aumentar la producción del NRP. Consecuentemente, la copia, al no estar sometida a una presión selectiva, ésta puede mutar la afinidad al substrato de sus dominios, evolucionando a un nuevo módulo con diferente afinidad. Consecuentes duplicaciones de estos genes permitirían la evolución de nuevos módulos con diferentes substratos, así como, la formación de BGC y la exploración de nuevas estructuras del producto. Sin embargo, es necesario entender por completo el proceso de formación de nuevas arquitecturas modulares, ya que, por sí solo, la duplicación de genes no es suficiente para poder explicar la gran variedad de arquitecturas (Figura 20). Eventos de intercambio de módulos por recombinación, duplicaciones de dominios únicos (como ej., la aparición de Dominios dobles de condensación, y su especialización hacia una reacción de epimerización) y mutaciones que inserten codones de paro para separar módulos o donde se elimine el codón original de paro del gen para combinar dos genes NRPs a uno solo, podrían haber tenido un papel en la reorganización de arquitecturas, separación o inserción de módulos, y en la aparición de nuevas arquitecturas únicas y diferentes arreglos en los BGC.

Se propone que las NRPS ancestrales debieron tener una función parecida a la NRPS productora de serrawettina W1, SwrW, ya que, es una NRPS iterativa de un solo módulo, es decir, se utiliza la misma proteína NRPS para sintetizar un dipéptido compuesto por dos

serinas y dos cadenas de acilos (Figura 25B). Asimismo, la SwrW contiene los cuatro dominios básicos de una NRPS (Figura 25A).Como se ha reportado antes (79, 80, 123, 142) los mecanismos evolutivos que han generado la variedad de NRPS han sucedido de forma variada. Estos mecanismos incluyen la duplicación, inserción/deleción, recombinación, mutación y rearreglo de genes (Figura 26). De la misma forma, se piensa que la evolución concertada juega un papel importante en la conservación de módulos exitosos, así como la transferencia horizontal de genes aumenta la variabilidad y la ocurrencia de NRPS en clados bacterianos distantes.

Al observar la conservación de ciertos BGCs dentro de géneros bacterianos, se nota que existe una gran presión de selección sobre estos genes, dado que al encontrar un péptido exitoso como la Surfactina, este va a conservarse a través de más cepas dentro de un taxón específico.

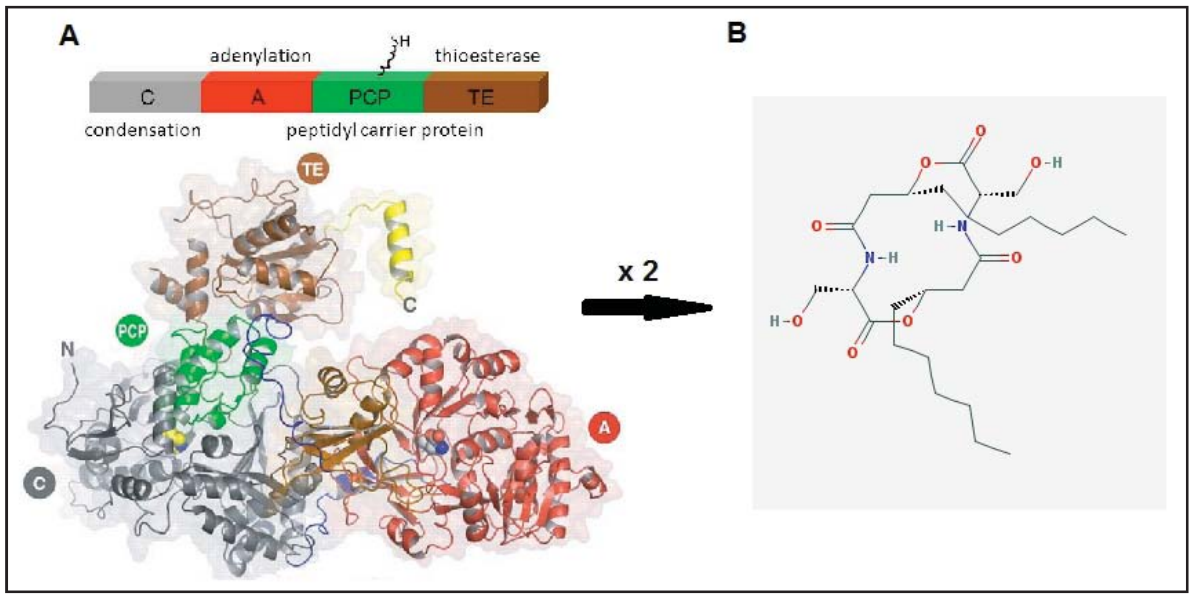

Figura 25. Esquema donde se ilustra una posible NRPS ancestral, su mecanismo de síntesis y su producto. Esta NRPS contiene un único módulo, el cual actúa de forma Iterativa, catalizando la formación de un dipéptido biológicamente activo compuesto por dos aminoácidos y dos cadenas lipídicas que le confiere una ventaja evolutiva. A, Dominios básicos de una NRPS y su estructura cristalográfica (adaptado de referencia 143). B, Estructura de la Serrawettina como ejemplo de un NRP ancestral (PubChem).

Es necesario un mayor conocimiento acerca de la estructura de las NRPS, ya que, hasta la fecha nunca se ha logrado caracterizar la estructura de una NRPS completa, tampoco de la interacción entre subunidades NRPS; únicamente de módulos individuales, de donde se ha inferido la estructura cuaternaria de la holo-enzima (5). Con las nuevas tecnologías de elucidación y modelamiento de estructuras proteicas como la Microscpia Cryo-Electrónica de partículas individuales (144, 145), donde se puede observar el estado original de una macromolécula a una alta resolución (~ 4 Å), será posible finalmente observar como las NRPS se arreglan en la holo-enzima. Esto podrá darnos las bases para diseñar protocolos en donde las NRPS puedan llevar a cabo la síntesis de péptidos *in-vitro*, aumentando la productividad y reduciendo los costos de producción a gran escala.

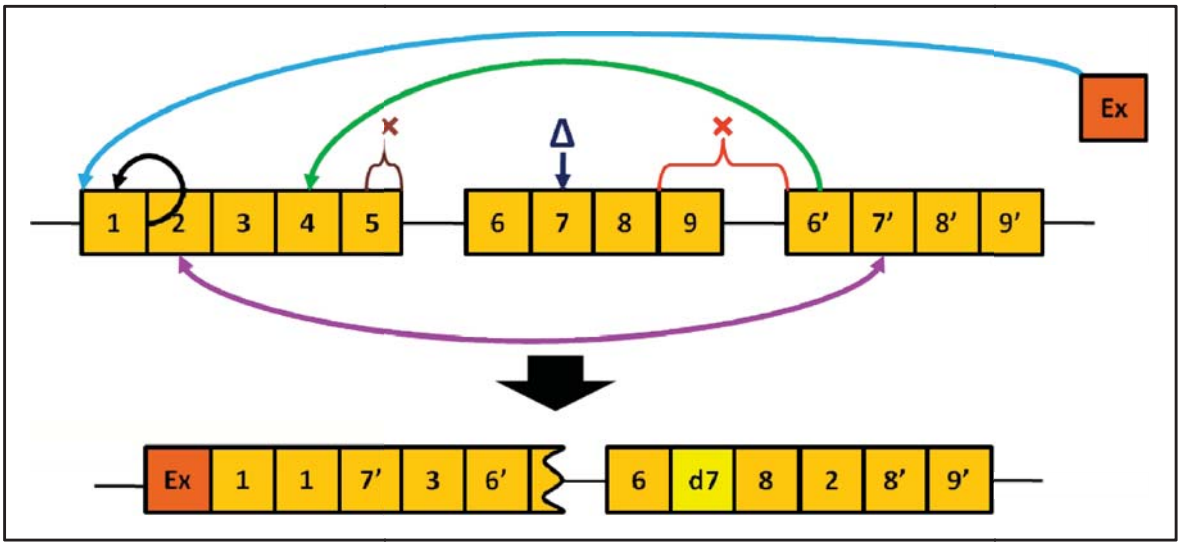

Figura 26. Diagrama ejemplificando los posibles eventos de reorganización dentro de un BGC NRPS por diferentes mecanismos en un clúster hipotético con un gen de 5 módulos, y dos genes idénticos duplicados de 4 módulos. Flechas: azul claro, inserción de una secuencia externa, ya sea, de otro BGC NRPS o PKS, un dominio accesorio como un dominio metiltransferasa, o de una secuencia adquirida por HGT; negra, duplicación de una región de la secuencia; morada, intercambio de módulos; verde, substitución de un módulo; café, perdida de la función o deleción de un dominio; azul obscuro, modificación de la afinidad al substrato por mutación del sitio activo; roja, deleción de una sección de la secuencia. La alta presión de selección, y la evolución concertada provocan que la mayoría de los módulos se conserven intactos.

La mayoría de las NRPS bacterianos están compuestos por un solo módulo, lo que indica que la evolución de NRPS multimodulares es un rasgo de una pequeña porción de especies. Esto sugiere la idea de que las NRPS multimodulares han surgido por un mecanismo especial que les permite duplicar, enlazar y explorar diferentes organizaciones de módulos, y en consecuencia, formar un BGC compuesto por varios genes de NRPS multimodulares. Éste mecanismo también podría estar implicado en el intercambio de módulos individuales o genes completos a través de la transferencia horizontal de genes, alimentando el pool genético de módulos y arquitecturas dentro un determinado grupo de bacterias, ya sea por estar estrechamente relacionados o porque habitan el mismo ecosistema.

Por otra parte, los resultados apuntan a que la estructura del sitio de reconocimiento del substrato es altamente variable, dando pie a la exploración de la afinidad a diferentes aminoácidos, sin que se pierda la función enzimática y que se mantenga intacta la estructura terciaria y cuaternaria de la proteína, así como del complejo enzimático. Convergencia evolutiva en la estructura del sitio de unión al substrato explicaría el porqué existen módulos sin homología inmediata pero que comparten la afinidad por el mismo substrato. Al analizar esta variedad de estructuras y comparando sus características, se podrán identificar las reglas que gobiernan está variabilidad de afinidades, dando pie a una predicción más precisa del substrato, su quiralidad, y posibles modificaciones pre-síntesis. Asimismo, permitirá el diseño racional de módulos NRPS afines a substratos completamente nuevos.

En conclusión, las NRPS ofrecen una oportunidad de explorar nuevos compuestos personalizados para una gran cantidad de problemas biotecnológicos, como la lucha contra las cepas de microbios con resistencia a múltiples antibióticos, las terapias personalizadas contra el cáncer, y el diseño de nuevas clases de medicamentos y biomoléculas comerciales.

10. Perspectivas

Estos hallazgos, aunados a varios estudios realizados anteriormente (146, 147) sugieren que los esfuerzos de bioingeniería de las NRPS deberían estar concentrados con la substitución de módulos completos en lugar de únicamente mutar la afinidad del dominio de la adenilación, lo que podría resolver el problema de la disminución en la eficiencia de síntesis (37, 148), provocado probablemente por la falta de afinidad al substrato en los otros dominios involucrados. Se propone que: la construcción de una biblioteca genética generada a partir de los tres tipos de módulos (Figura 23), compuesta de módulos consenso o representantes viables específicos para cada substrato, que a su vez sean compatibles e intercambiables, podría finalmente dar acceso a los investigadores a una plataforma de desarrollo de NRP sintéticos.

Una perspectiva clara es analizar a los genes NRPS para encontrar los sitios dentro de cada módulo que se encuentran afectados por una presión de selección, al buscar por evidencia de una selección purificadora o estabilizadora. Asimismo, sería interesante comparar las distancias Robinson-Foulds calculadas para módulos de NRPS que presenten una alta conservación en busca de coevolución (149). Por otro lado, es necesario investigar a fondo ciertos aspectos de las NRPS que no han recibido tanta atención por parte de los grupos de investigación dedicados al estudio de éstas megasintetasas: como el transporte de péptidos al medio extracelular; la caracterización de la secuencia 'linker' intermodular y su papel en la estructura y la organización modular de la holoenzima NRPS; así como, la regulación de la expresión y la autoresistencia de la cepa productora.

Con las nuevas tecnologías de secuenciación, como por ejemplo, la secuenciación de células únicas , se podría avanzar y potencializar aún más el descubrimiento de productos naturales y sus BGC respectivos, así como el reconocimiento y estudio de las cepas productoras, aún siendo no cultivables en laboratorio, usando algunas técnicas para el aislamiento y cultivo, como el 'iChip', desarrollado para el aislamiento de células únicas y su cultivo en condiciones in-situ (150). Éstas técnicas permitirían realizar proyectos genómicos y transcriptómicos extensivos y precisos en comunidades microbianas de difícil

acceso, pero con gran potencial metabólico donde se pueda reconocer la expresión de genes NRPS. Como ejemplos de posibles ecosistemas donde se cree que existe un gran potencial metabólico, están las comunidades degradadoras de hidrocarburos; microbiomas asociados tractos digestivos de animales, rizomas de plantas o microorganismos simbiontes de cnidarios, esponjas, etc.; microbiomas marinos, fluviales y de suelos contaminados; microbiomas de comunidades con importancia en la industria, como las de fermentación de alimentos y bebidas, productoras de biocombustibles, fitopatógenas o con relevancia médica.

Por último, la generación de un acervo científico sobre las NRPS permitirá el tener la capacidad de diseñar nuevos tipos de péptidos para fines específicos o blancos terapéuticos relevantes, con nuevas y/o mejoradas funciones, que combatan los desafíos que tenemos a la mano, como por ejemplo, el aumento de cepas resistentes a antibióticos, o el uso de nuevos medicamentos para combatir el cáncer. Ultimadamente este acervo científico servirá para el desarrollo de cepas que sinteticen metabolitos secundarios y productos naturales con usos en la farmacéutica e industria biotecnológica; así como, para la ingeniería de las NRPS. Utilizando una cepa con potencial para producir metabolitos, y con gran cantidad de genes NRPS (ej. *Tumebacillus sp*.), nos será posible tener un excelente modelo para la bioingeniería de NRPS, utilizando los módulos que ya existen dentro de la especie, para generar cepas recombinantes donde se exploren diferentes estructuras de NRP.

Esperamos que este trabajo de pie a subsecuentes estudios de estructura, función, evolución, rediseño y descubrimiento de las NRPS, así como proveer una base bibliográfica para la generación de nuevos proyectos de investigación y de divulgación de estas tan interesantes megaproteínas.

11. Bibliografía

1. **Gudiña EJ**, **Teixeira JA**, **Rodrigues LR**. 2016. Biosurfactants produced by marine microorganisms with therapeutic applications. Mar. Drugs **14**:38.
2. **Medema MH**, **Fischbach MA**. 2015. Computational approaches to natural product discovery. Nat. Chem. Biol. **11**:639–648.
3. **Kries H**, **Hilvert D**. 2011. Tailor-made peptide synthetases. Chem. Biol. **18**:1206–1207.
4. **Strieker M**, **Tanović A**, **Marahiel MA**. 2010. Nonribosomal peptide synthetases: Structures and dynamics. Curr. Opin. Struct. Biol. **20**:234–240.
5. **Marahiel MA**. 2016. A structural model for multimodular NRPS assembly lines. Nat. Prod. Rep. **33**:136–140.
6. **Miller BR**, **Gulick AM**. 2016. Structural Biology of Nonribosomal Peptide Synthetases. Methods Mol. Biol. **1401**:3–29.
7. **Schwarzer D**, **Finking R**, **Marahiel MA**. 2003. Nonribosomal peptides: from genes to products. Nat. Prod. Rep. **20**:275.
8. **Desriac F**, **Jégou C**, **Balnois E**, **Brillet B**, **Le Chevalier P**, **Fleury Y**. 2013. Antimicrobial peptides from marine proteobacteria. Mar. Drugs.
9. **Felnagle EA**, **Jackson EE**, **Chan YA**, **Podevels AM**, **Berti D**, **Mcmahon MD**, **Thomas MG**. 2011. Nonribosomal Peptide Synthetases Involved in the Production of Medically Relevant Natural Products **5**:191–211.
10. **Liaimer A**, **Helfrich EJN**, **Hinrichs K**, **Guljamow A**, **Ishida K**, **Hertweck C**, **Dittmann E**. 2015. Nostopeptolide plays a governing role during cellular differentiation of the symbiotic cyanobacterium *Nostoc punctiforme*. Proc. Natl. Acad. Sci. **112**:1862–1867.
11. **Recktenwald J**, **Shawky R**, **Puk O**, **Pfenning F**, **Keller U**, **Wohlleben W**, **Pelzer S**. 2002. Nonribosomal biosynthesis of vancomycin-type antibiotics: A heptapeptide backbone and eight peptide synthetase modules. Microbiology **148**:1105–1118.
12. **Shaligram NS**, **Singhal RS**. 2010. Surfactin -a review on biosynthesis, fermentation, purification and applications. Food Technol. Biotechnol.
13. **Abderrahmani A**, **Tapi A**, **Nateche F**, **Chollet M**, **Leclère V**, **Wathelet B**, **Hacene H**, **Jacques P**. 2011. Bioinformatics and molecular approaches to detect NRPS genes involved in the biosynthesis of kurstakin from Bacillus thuringiensis. Appl. Microbiol. Biotechnol. **92**:571–581.
14. **Hamley IW**. 2015. Lipopeptides: from self-assembly to bioactivity. Chem. Commun. **51**:8574–8583.
15. **Jackson SA**, **Borchert E**, **O'Gara F**, **Dobson ADW**. 2015. Metagenomics for the discovery of novel biosurfactants of environmental interest from marine ecosystems. Curr. Opin. Biotechnol. **33**:176–182.
16. **Abdel-Mawgoud AM**, **Rudolf Hausmann**, **Lépine F**, **Müller MM**, **Déziel E**. 2011. BiosurfactantsBiosurfactants.
17. **Thies S**, **Santiago-Schübel B**, **Kovačić F**, **Rosenau F**, **Hausmann R**, **Jaeger KE**. 2014. Heterologous production of the lipopeptide biosurfactant serrawettin W1 in Escherichia coli. J. Biotechnol. **181**:27–30.
18. **Hur GH**, **Vickery CR**, **Burkart MD**. 2012. Explorations of catalytic domains in non-ribosomal peptide synthetase enzymology. Nat. Prod. Rep. **29**:1074.
19. **Gu JQ**, **Alexander DC**, **Rock J**, **Brian P**, **Chu M**, **Baltz RH**. 2011. Structural characterization of a lipopeptide antibiotic A54145E(Asn3Asp9) produced by a genetically engineered strain of Streptomyces fradiae. J Antibiot **64**:111–116.
20. **Miao V**, **Brost R**, **Chapple J**, **She K**, **Coëffet-Le Gal MF**, **Baltz RH**. 2006. The lipopeptide antibiotic A54145 biosynthetic gene cluster from Streptomyces fradiae, p. 129–140. *In* Journal of Industrial Microbiology and Biotechnology.
21. **Nguyen KT**, **He X**, **Alexander DC**, **Li C**, **Gu JQ**, **Mascio C**, **Van Praagh A**, **Mortin L**, **Chu M**, **Silverman JA**, **Brian P**, **Baltz RH**. 2010. Genetically engineered lipopeptide antibiotics related to A54145 and daptomycin with improved properties. Antimicrob. Agents Chemother. **54**:1404–1413.
22. **Morikawa M**, **Daido H**, **Takao T**, **Murata S**, **Shimonishi Y**, **Imanaka T**. 1993. A new lipopeptide biosurfactant produced by Arthrobacter sp. strain MIS38. J. Bacteriol. **175**:6459–6466.

23. **Roongsawang N**, **Hase K**, **Haruki M**, **Imanaka T**, **Morikawa M**, **Kanaya S**. 2003. Cloning and Characterization of the Gene Cluster Encoding Arthrofactin Synthetase from Pseudomonas sp. MIS38. Chem. Biol. **10**:869–880.
24. **Washio K**, **Lim SP**, **Roongsawang N**, **Morikawa M**. 2010. Identification and characterization of the genes responsible for the production of the cyclic lipopeptide arthrofactin by Pseudomonas sp. MIS38. Biosci. Biotechnol. Biochem. **74**:992–999.
25. **Lange A**, **Sun H**, **Pilger J**, **Reinscheid UM**, **Gross H**. 2012. Predicting the Structure of Cyclic Lipopeptides by Bioinformatics: Structure Revision of Arthrofactin. ChemBioChem **13**:2671–2675.
26. **Konz D**, **Klens A**, **Schörgendorfer K**, **Marahiel MA**. 1997. The bacitracin biosynthesis operon of Bacillus licheniformis ATCC 10716: Molecular characterization of three multi-modular peptide synthetaseKonz, D., Klens, A., Schörgendorfer, K., and Marahiel, M.A. (1997) The bacitracin biosynthesis operon of Bacillus . Chem. Biol. **4**:927–937.
27. **Suleiman SA**, **Song F**, **Su M**, **Hang T**, **Song M**. 2017. Analysis of bacitracin and its related substances by liquid chromatography tandem mass spectrometry. J. Pharm. Anal. **7**:48–55.
28. **Hojati Z**, **Milne C**, **Harvey B**, **Gordon L**, **Borg M**, **Flett F**, **Wilkinson B**, **Sidebottom PJ**, **Rudd BAM**, **Hayes MA**, **Smith CP**, **Micklefield J**. 2002. Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from Streptomyces coelicolor. Chem. Biol. **9**:1175–1187.
29. **Bum Kim H**, **Smith CP**, **Micklefield J**, **Mavituna F**. 2004. Metabolic flux analysis for calcium dependent antibiotic (CDA) production in Streptomyces coelicolor. Metab. Eng. **6**:313–325.
30. **Kraas FI**, **Giessen TW**, **Marahiel MA**. 2012. Exploring the mechanism of lipid transfer during biosynthesis of the acidic lipopeptide antibiotic CDA. FEBS Lett. **586**:283–288.
31. **Bloudoff K**, **Rodionov D**, **Schmeing TM**. 2013. Crystal structures of the first condensation domain of CDA synthetase suggest conformational changes during the synthetic cycle of nonribosomal peptide synthetases. J. Mol. Biol. **425**:3137–3150.
32. **Miao V**, **Coëffet-LeGal MF**, **Brian P**, **Brost R**, **Penn J**, **Whiting A**, **Martin S**, **Ford R**, **Parr I**, **Bouchard M**, **Silva CJ**, **Wrigley SK**, **Baltz RH**. 2005. Daptomycin biosynthesis in Streptomyces roseosporus: Cloning and analysis of the gene cluster and revision of peptide stereochemistry. Microbiology **151**:1507–1523.
33. **Robbel L**, **Marahiel MA**. 2010. Daptomycin, a bacterial lipopeptide synthesized by a nonribosomal machinery. J. Biol. Chem. **285**:27501–27508.
34. **Baltz RH**, **Brian P**, **Miao V**, **Wrigley SK**. 2006. Combinatorial biosynthesis of lipopeptide antibiotics in Streptomyces roseosporus, p. 66–74. *In* Journal of Industrial Microbiology and Biotechnology.
35. **Taylor SD**, **Palmer M**. 2016. The action mechanism of daptomycin. Bioorganic Med. Chem. **24**:6253–6268.
36. **Hatano K**, **Nogami I**, **Higashide E**, **Kishi T**. 1984. Biosynthesis of enduracidin: Origin of enduracididine and other amino acids. Agric. Biol. Chem. **48**:1503–1508.
37. **Wu MC**, **Styles MQ**, **Law BJC**, **Struck AW**, **Nunns L**, **Micklefield J**. 2015. Engineered biosynthesis of enduracidin lipoglycopeptide antibiotics using the ramoplanin mannosyltransferase Ram29. Microbiology **161**:1338–1347.
38. **Kim P Il**, **Ryu J**, **Kim YH**, **Chi YT**. 2010. Production of biosurfactant lipopeptides iturin A, fengycin, and surfactin A from Bacillus subtilis CMB32 for control of Colletotrichum gloeosporioides. J. Microbiol. Biotechnol. **20**:138–145.
39. **Wu CY**, **Chen CL**, **Lee YH**, **Cheng YC**, **Wu YC**, **Shu HY**, **Gotz F**, **Liu ST**. 2007. Nonribosomal synthesis of fengycin on an enzyme complex formed by fengycin synthetases. J. Biol. Chem. **282**:5608–5616.
40. **Steller S**, **Vollenbroich D**, **Leenders F**, **Stein T**, **Conrad B**, **Hofemeister J**, **Jacques P**, **Thonart P**, **Vater J**. 1999. Structural and functional organization of the fengycin synthetasemultienzyme system from Bacillus subtilis b213 and A1/3. Chem. Biol. **6**:R156–R156.
41. **Choi SK**, **Park SY**, **Kim R**, **Lee CH**, **Kim JF**, **Park SH**. 2008. Identification and functional analysis of the fusaricidin biosynthetic gene of Paenibacillus polymyxa E681. Biochem. Biophys. Res. Commun. **365**:89–95.
42. **Kajimura Y**, **Kaneda M**. 1996. Fusaricidin A, a new depsipeptide antibiotic produced by Bacillus polymyxa KT-8. Taxonomy, fermentation, isolation, structure elucidation and biological activity. J. Antibiot. (Tokyo). **49**:129–35.

43. **Li J**, **Jensen SE**. 2008. Nonribosomal Biosynthesis of Fusaricidins by Paenibacillus polymyxa PKB1 Involves Direct Activation of a d-Amino Acid. Chem. Biol. **15**:118–127.
44. **Kratzschmar J**, **Krause M**, **Marahiel M a**. 1989. Gramicidin-S Biosynthesis Operon Containing the Structural Genes Grsa and Grsb Has an Open Reading Frame Encoding a Protein Homologous to Fatty-Acid Thioesterases. J. Bacteriol. **171**:5422–5429.
45. **Conti E**, **Stachelhaus T**, **Marahiel M a**, **Brick P**. 1997. Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. Embo J **16**:4174–4183.
46. **Kessler N**, **Schuhmann H**, **Morneweg S**, **Linne U**, **Marahiel MA**. 2004. The Linear Pentadecapeptide Gramicidin Is Assembled by Four Multimodular Nonribosomal Peptide Synthetases That Comprise 16 Modules with 56 Catalytic Domains. J. Biol. Chem. **279**:7413–7419.
47. **Konz D**, **Doekel S**, **Marahiel MA**. 1999. Molecular and biochemical characterization of the protein template controlling biosynthesis of the lipopeptide lichenysin. J. Bacteriol. **181**:133–140.
48. **Yakimov MM**, **Timmis KN**, **Wray V**, **Fredrickson HL**. 1995. Characterization of a new lipopeptide surfactant produced by thermotolerant and halotolerant subsurface Bacillus licheniformis BAS50. Appl. Environ. Microbiol. **61**:1706–1713.
49. **Madslien EH**, **Rønning HT**, **Lindbäck T**, **Hassel B**, **Andersson MA**, **Granum PE**. 2013. Lichenysin is produced by most Bacillus licheniformis strains. J. Appl. Microbiol. **115**:1068–1080.
50. **Tran H**, **Ficke A**, **Asiimwe T**, **Höfte M**, **Raaijmakers JM**. 2007. Role of the cyclic lipopeptide massetolide a in biological control of Phytophthora infestans and in colonization of tomato plants by Pseudomonas fluorescens. New Phytol. **175**:731–742.
51. **De Bruijn I**, **De Kock MJD**, **De Waard P**, **Van Beek TA**, **Raaijmakers JM**. 2008. Massetolide A biosynthesis in Pseudomonas fluorescens. J. Bacteriol. **190**:2777–2789.
52. **Song C**, **Sundqvist G**, **Malm E**, **de Bruijn I**, **Kumar A**, **van de Mortel J**, **Bulone V**, **Raaijmakers JM**. 2015. Lipopeptide biosynthesis in *Pseudomonas fluorescens* is regulated by the protease complex ClpAP. BMC Microbiol. **15**:367.
53. **Hoffmann D**, **Hevel JM**, **Moore RE**, **Moore BS**. 2003. Sequence analysis and biochemical characterization of the nostopeptolide A biosynthetic gene cluster from Nostoc sp. GSV224. Gene **311**:171–180.
54. **Ma Z**, **Geudens N**, **Kieu NP**, **Sinnaeve D**, **Ongena M**, **Martins JC**, **Höfte M**. 2016. Biosynthesis, chemical structure, and structure-activity relationship of orfamide lipopeptides produced by Pseudomonas protegens and related species. Front. Microbiol. **7**:1–16.
55. **Tsuge K**, **Matsui K**, **Itaya M**. 2007. Production of the non-ribosomal peptide plipastatin in Bacillus subtilis regulated by three relevant gene blocks assembled in a single movable DNA segment. J. Biotechnol. **129**:592–603.
56. **Batool M**, **Khalid MH**, **Hassan MN**, **Hafeez FY**. 2011. Homology modeling of an antifungal metabolite plipastatin synthase from the bacillus subtilis 168. Bioinformation **7**:384–387.
57. **Choi SK**, **Park SYH**, **Kim R**, **Kim SB**, **Lee CH**, **Kim JF**, **Park SYH**. 2009. Identification of a polymyxin synthetase gene cluster of Paenibacillus polymyxa and heterologous expression of the gene in Bacillus subtilis. J Bacteriol **191**:3350–3358.
58. **Deng Y**, **Lu Z**, **Bi H**, **Lu F**, **Zhang C**, **Bie X**. 2011. Isolation and characterization of peptide antibiotics LI-F04 and polymyxin B6produced by Paenibacillus polymyxa strain JSa-9. Peptides **32**:1917–1923.
59. **Trapet P**, **Avoscan L**, **Klinguer A**, **Pateyron S**, **Citerne S**, **Chervin C**, **Mazurier S**, **Lemanceau P**, **Wendehenne D**, **Besson-Bard A**. 2016. The *Pseudomonas fluorescens* Siderophore Pyoverdine Weakens *Arabidopsis thaliana* Defense in Favor of Growth in Iron-Deficient Conditions. Plant Physiol. **171**:675–693.
60. **Schalk IJ**, **Guillon L**. 2013. Pyoverdine biosynthesis and secretion in Pseudomonas aeruginosa: Implications for metal homeostasis. Environ. Microbiol. **15**:1661–1673.
61. **Vandenende CS**, **Vlasschaert M**, **Seah SYK**. 2004. Functional characterization of an aminotransferase required for pyoverdine siderophore biosynthesis in Pseudomonas aeruginosa PAO1. J. Bacteriol. **186**:5596–5602.
62. **Apao MMN**, **Teves FG**, **Madamba MRSB**. 2012. Sequence analysis of putative swrW gene required for surfactant serrawettin W1 production from Serratia marcescens. African J. Biotechnol. **11**:12040-044.

63. **Li H**, **Tanikawa T**, **Sato Y**, **Nakagawa Y**, **Matsuyama T**. 2005. Serratia marcescens gene required for surfactant serrawettin W1 production encodes putative aminolipid synthetase belonging to nonribosomal peptide synthetase family. Microbiol. Immunol. **49**:303–310.
64. **Tanikawa T**, **Nakagawa Y**, **Matsuyama T**. 2006. Prodigiosin and Serrawettin W1 Biosynthesis in Serratia marcescens **50**:587–596.
65. **Matsuyama T**, **Tanikawa T**, **Nakagawa Y**. 2011. Serrawettins and Other Surfactants Produced by Serratia, p. 93–120. *In* Soberón-Chávez, G (ed.), . Springer Berlin Heidelberg, Berlin, Heidelberg.
66. **Kluge B**, **Vater J**, **Salnikow J**, **Eckart K**. 1988. Studies on the biosynthesis of surfactin, a lipopeptide antibiotic from Bacillus subtilis ATCC 21332. FEBS Lett. **231**:107–110.
67. **Menkhaus M**, **Ullrich C**, **Kluge B**, **Vater J**, **Vollenbroich D**, **Kamp RM**. 1993. Structural and functional organization of the surfactin synthetase multienzyme system. J. Biol. Chem. **268**:7678–7684.
68. **Ongena M**, **Jacques P**. 2008. Bacillus lipopeptides: versatile weapons for plant disease biocontrol. Trends Microbiol.
69. **Bender CL**, **Scholz-Schroeder BK**. 2004. New Insights Into the Biosynthesis, Mode of Action, and Regulation of Syringomycin, syringopeptin, and Coronatine, p. 125–158. *In* Pseudomonas. Springer US, Boston, MA.
70. **Scholz-Schroeder BK**, **Soule JD**, **Lu S-E**, **Grgurina I**, **Gross DC**. 2001. A Physical Map of the Syringomycin and Syringopeptin Gene Clusters Localized to an Approximately 145-kb DNA Region of *Pseudomonas syringae* pv. *syringae* Strain B301D. Mol. Plant-Microbe Interact. **14**:1426–1435.
71. **Anselmi M**, **Eliseo T**, **Zanetti-Polzi L**, **Fullone MR**, **Fogliano V**, **Di Nola A**, **Paci M**, **Grgurina I**. 2011. Structure of the lipodepsipeptide syringomycin e in phospholipids and sodium dodecylsulphate micelle studied by circular dichroism, NMR spectroscopy and molecular dynamics. Biochim. Biophys. Acta - Biomembr. **1808**:2102–2110.
72. **Becucci L**, **Tramonti V**, **Fiore A**, **Fogliano V**, **Scaloni A**, **Guidelli R**. 2015. Channel-forming activity of syringomycin e in two mercury-supported biomimetic membranes. Biochim. Biophys. Acta - Biomembr. **1848**:932–941.
73. **Carpaneto A**, **Dalla Serra M**, **Menestrina G**, **Fogliano V**, **Gambale F**. 2002. The phytotoxic lipodepsipeptide syringopeptin 25A from Pseudomonas syringae pv syringae forms ion channels in sugar beet vacuoles. J. Membr. Biol. **188**:237–248.
74. **Okano A**, **Isley NA**, **Boger DL**. 2017. Peripheral modifications of [Ψ[CH $_2$ NH]Tpg ^{4}]vancomycin with added synergistic mechanisms of action provide durable and potent antibiotics. Proc. Natl. Acad. Sci. 201704125.
75. **Horsman ME**, **Hari TPA**, **Boddy CN**. 2016. Polyketide synthase and non-ribosomal peptide synthetase thioesterase selectivity: logic gate or a victim of fate? Nat. Prod. Rep. **33**:183–202.
76. **Roongsawang N**, **Washio K**, **Morikawa M**. 2011. Diversity of nonribosomal peptide synthetases involved in the biosynthesis of lipopeptide biosurfactants. Int. J. Mol. Sci. **12**:141–172.
77. **Meyer S**, **Kehr JC**, **Mainz A**, **Dehm D**, **Petras D**, **Süssmuth RD**, **Dittmann E**. 2016. Biochemical Dissection of the Natural Diversification of Microcystin Provides Lessons for Synthetic Biology of NRPS. Cell Chem. Biol. **23**:462–471.
78. **Winn M**, **Fyans JK**, **Zhuo Y**, **Micklefield J**. 2016. Recent advances in engineering nonribosomal peptide assembly lines. Nat. Prod. Rep. **33**:317–347.
79. **Medema MH**, **Cimermancic P**, **Sali A**, **Takano E**, **Fischbach MA**. 2014. A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. PLoS Comput. Biol. **10**.
80. **Wang H**, **Fewer DP**, **Holm L**, **Rouhiainen L**, **Sivonen K**. 2014. Atlas of nonribosomal peptide and polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes. Proc. Natl. Acad. Sci. **111**:9259–9264.
81. **Kroken S**, **Glass NL**, **Taylor JW**, **Yoder OC**, **Turgeon BG**. 2003. Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proc. Natl. Acad. Sci. **100**:15670–15675.
82. **Brocchieri L**, **Karlin S**. 2005. Protein length in eukaryotic and prokaryotic proteomes. Nucleic Acids Res. **33**:3390–3400.
83. **Jackisch-matsuura AB**, **Santos LS**, **Eberlin MN**, **De AF**. 2014. Production and Characterization

Compounds from Gordonia amicalis of **57**:138–144.
84. **Luo C**, **Liu X**, **Zhou X**, **Guo J**, **Truong J**, **Wang X**, **Zhou H**, **Li X**, **Chen Z**. 2015. Unusual biosynthesis and structure of locillomycins from Bacillus subtilis 916. Appl. Environ. Microbiol. **81**:6601–6609.
85. **Bruner SD**, **Weber T**, **Kohli RM**, **Schwarzer D**, **Marahiel MA**, **Walsh CT**, **Stubbs MT**. 2002. Structural basis for the cyclization of the lipopeptide antibiotic surfactin by the thioesterase domain SrfTE. Structure **10**:301–310.
86. **Stachelhaus T**, **Mootz HD**, **Bergendahl V**, **Marahiel M a**. 1998. Peptide Bond Formation in Nonribosomal Peptide Biosynthesis. J. Biol. Chem. **273**:22773–22781.
87. **Till M**, **Race PR**. 2016. Nonribosomal Peptide and Polyketide BiosynthesisMethods in Molecular Biology. Springer New York, New York, NY.
88. **Vergnolle O**, **Xu H**, **Blanchard JS**. 2013. Mechanism and regulation of mycobactin fatty acyl-AMP ligase FadD33. J. Biol. Chem. **288**:28116–28125.
89. **Weissman KJ**. 2015. The structural biology of biosynthetic megaenzymes. Nat. Chem. Biol.
90. **Ackerley DF**. 2016. Cracking the Nonribosomal Code. Cell Chem. Biol. **23**:535–537.
91. **Röttig M**, **Medema MH**, **Blin K**, **Weber T**, **Rausch C**, **Kohlbacher O**. 2011. NRPSpredictor2 - A web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. **39**:362–367.
92. **Keating TA**, **Marshall CG**, **Walsh CT**, **Keating AE**. 2002. The structure of VibH represents nonribosomal peptide synthetase condensation, cyclization and epimerization domains. Nat. Struct. Biol. **9**:522.
93. **Samel SA**, **Schoenafinger G**, **Knappe TA**, **Marahiel MA**, **Essen LO**. 2007. Structural and Functional Insights into a Peptide Bond-Forming Bidomain from a Nonribosomal Peptide Synthetase. Structure **15**:781–792.
94. **Rausch C**, **Hoof I**, **Weber T**, **Wohlleben W**, **Huson DH**. 2007. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evol. Biol. **7**:78.
95. **Coulthurst SJ**. 2014. Role of the phosphopantetheinyltransferase enzyme, PswP, in the biosynthesis of antimicrobial secondary metabolites by Serratia marcescens Db10. Microbiology **160**:1609–1617.
96. **Bloudoff K**, **Fage CD**, **Marahiel MA**, **Schmeing TM**. 2016. Structural and mutational analysis of the nonribosomal peptide synthetase heterocyclization domain provides insight into catalysis.
97. **Du FY**, **Li XM**, **Zhang P**, **Li CS**, **Wang BG**. 2014. Cyclodepsipeptides and Other O-containing heterocyclic metabolites from Beauveria felina EN-135, a marine-derived entomopathogenic fungus. Mar. Drugs **12**:2816–2826.
98. **Milne C**, **Powell A**, **Jim J**, **Al Nakeeb M**, **Smith CP**, **Micklefield J**. 2006. Biosynthesis of the (2S,3R)-3-methyl glutamate residue of nonribosomal lipopeptides. J. Am. Chem. Soc. **128**:11250–11259.
99. **Giltrap AM**, **Dowman LJ**, **Nagalingam G**, **Ochoa JL**, **Linington RG**, **Britton WJ**, **Payne RJ**. 2016. Total Synthesis of Teixobactin. Org. Lett. **18**:2788–2791.
100. **Ling LL**, **Schneider T**, **Peoples AJ**, **Spoering AL**, **Engels I**, **Conlon BP**, **Mueller A**, **Schäberle TF**, **Hughes DE**, **Epstein S**, **Jones M**, **Lazarides L**, **Steadman VA**, **Cohen DR**, **Felix CR**, **Fetterman KA**, **Millett WP**, **Nitti AG**, **Zullo AM**, **Chen C**, **Lewis K**. 2015. A new antibiotic kills pathogens without detectable resistance. Nature **517**:455–459.
101. **Ansari MZ**, **Yadav G**, **Gokhale RS**, **Mohanty D**. 2004. NRPS-PKS: A knowledge-based resource for analysis of NRPS-PKS megasynthases. Nucleic Acids Res. **32**:405–413.
102. **Du L**, **Sánchez C**, **Chen M**, **Edwards DJ**, **Shen B**. 2000. The biosynthetic gene cluster for the antitumor drug bleomycin from Streptomyces verticillus ATCC15003 supporting functional interactions between nonribosomal peptide synthetases and a polyketide synthase. Chem. Biol. **7**:623–642.
103. **Shen B**, **Du L**, **Sanchez C**, **Edwards DJ**, **Chen M**, **Murrell JM**. 2002. Cloning and characterization of the bleomycin biosynthetic gene cluster from Streptomyces verticillus ATCC15003. J. Nat. Prod. **65**:422–431.
104. **Armbrust E V.**, **Palumbi SR**. 2015. Uncovering hidden worlds of ocean biodiversity. Science (80-.). **348**:865–867.
105. **Hug LA**, **Baker BJ**, **Anantharaman K**, **Brown CT**, **Probst AJ**, **Castelle CJ**, **Butterfield CN**, **Hernsdorf AW**, **Amano Y**, **Ise K**, **Suzuki Y**, **Dudek N**, **Relman DA**, **Finstad KM**, **Amundson R**, **Thomas BC**, **Banfield JF**. 2016. A new view of the tree of life. Nat. Microbiol. **1**:1–6.
106. **Eddy SR**. 2009. HMMER3 beta test : User's Guide. Biol. Seq. Anal. using profile hidden Markov

Model. 0–32.
107. **Eddy S**. 1998. Profile hidden Markov models. Bioinformatics **14**:755–763.
108. **Marchler-Bauer A**, **Derbyshire MK**, **Gonzales NR**, **Lu S**, **Chitsaz F**, **Geer LY**, **Geer RC**, **He J**, **Gwadz M**, **Hurwitz DI**, **Lanczycki CJ**, **Lu F**, **Marchler GH**, **Song JS**, **Thanki N**, **Wang Z**, **Yamashita RA**, **Zhang D**, **Zheng C**, **Bryant SH**. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. **43**:D222–D226.
109. **Finn RD**, **Coggill P**, **Eberhardt RY**, **Eddy SR**, **Mistry J**, **Mitchell AL**, **Potter SC**, **Punta M**, **Qureshi M**, **Sangrador-Vegas A**, **Salazar GA**, **Tate J**, **Bateman A**. 2016. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res. **44**:D279–D285.
110. **Durbin S**, **Eddy S**, **Krogh A**, **Mitchison G**. 1998. Biological sequence analysis: probabilistic models of proteins and nucleic acidsCambridge University Press, London. Cambridge university press.
111. **Rabiner LR**. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE **77**:257–286.
112. **Weber T**, **Blin K**, **Duddela S**, **Krug D**, **Kim HU**, **Bruccoleri R**, **Lee SY**, **Fischbach MA**, **Müller R**, **Wohlleben W**, **Breitling R**, **Takano E**, **Medema MH**. 2015. AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. **43**:W237–W243.
113. **Chevrette MG**, **Aicheler F**, **Kohlbacher O**, **Currie CR**, **Medema MH**. 2017. SANDPUMA: ensemble predictions of nonribosomal peptide chemistry reveal biosynthetic diversity across Actinobacteria. Bioinformatics 1–9.
114. **Singh M**, **Chaudhary S**, **Sareen D**. 2017. Non-ribosomal peptide synthetases: Identifying the cryptic gene clusters and decoding the natural product. J. Biosci. **42**:175–187.
115. **Nielsen JC**, **Grijseels S**, **Prigent S**, **Ji B**, **Dainat J**, **Nielsen KF**, **Frisvad JC**, **Workman M**, **Nielsen JC**. 2017. reveals vast potential of secondary metabolite production in Penicillium species **17044**.
116. **Ausec L**, **Zakrzewski M**, **Goesmann A**, **Schlüter A**, **Mandic-Mulec I**. 2011. Bioinformatic analysis reveals high diversity of bacterial genes for laccase-like enzymes. PLoS One **6**.
117. **Esmaeel Q**, **Pupin M**, **Kieu NP**, **Chataigné G**, **Béchet M**, **Deravel J**, **Krier F**, **Höfte M**, **Jacques P**, **Leclère V**. 2016. Burkholderia genome mining for nonribosomal peptide synthetases reveals a great potential for novel siderophores and lipopeptides synthesis. Microbiologyopen **5**:512–526.
118. **Kries H**, **Niquille DL**, **Hilvert D**. 2015. A subdomain swap strategy for reengineering nonribosomal peptides. Chem. Biol. **22**:640–648.
119. **Chen C-Y**, **Georgiev I**, **Anderson AC**, **Donald BR**. 2009. Computational structure-based redesign of enzyme activity. Proc. Natl. Acad. Sci. **106**:3764–3769.
120. **Williams G**. 2013. Engineering polyketide synthases and nonribosomal peptide synthetases. Curr. Opin. Struct. Biol. **23**:603–612.
121. **Medema MH**, **Kottmann R**, **Yilmaz P**, **Cummings M**, **Biggins JB**, **Blin K**, **de Bruijn I**, **Chooi YH**, **Claesen J**, **Coates RC**, **Cruz-Morales P**, **Duddela S**, **Dusterhus S**, **Edwards DJ**, **Fewer DP**, **Garg N**, **Geiger C**, **Gomez-Escribano JP**, **Greule A**, **Hadjithomas M**, **Haines AS**, **Helfrich EJN**, **Hillwig ML**, **Ishida K**, **Jones AC**, **Jones CS**, **Jungmann K**, **Kegler C**, **Kim HU**, **Kotter P**, **Krug D**, **Masschelein J**, **Melnik A V**, **Mantovani SM**, **Monroe EA**, **Moore M**, **Moss N**, **Nutzmann H-W**, **Pan G**, **Pati A**, **Petras D**, **Reen FJ**, **Rosconi F**, **Rui Z**, **Tian Z**, **Tobias NJ**, **Tsunematsu Y**, **Wiemann P**, **Wyckoff E**, **Yan X**, **Yim G**, **Yu F**, **Xie Y**, **Aigle B**, **Apel AK**, **Balibar CJ**, **Balskus EP**, **Barona-Gomez F**, **Bechthold A**, **Bode HB**, **Borriss R**, **Brady SF**, **Brakhage AA**, **Caffrey P**, **Cheng Y-Q**, **Clardy J**, **Cox RJ**, **De Mot R**, **Donadio S**, **Donia MS**, **van der Donk WA**, **Dorrestein PC**, **Doyle S**, **Driessen AJM**, **Ehling-Schulz M**, **Entian K-D**, **Fischbach MA**, **Gerwick L**, **Gerwick WH**, **Gross H**, **Gust B**, **Hertweck C**, **Hofte M**, **Jensen SE**, **Ju J**, **Katz L**, **Kaysser L**, **Klassen JL**, **Keller NP**, **Kormanec J**, **Kuipers OP**, **Kuzuyama T**, **Kyrpides NC**, **Kwon H-J**, **Lautru S**, **Lavigne R**, **Lee CY**, **Linquan B**, **Liu X**, **Liu W**, **Luzhetskyy A**, **Mahmud T**, **Mast Y**, **Mendez C**, **Metsa-Ketela M**, **Micklefield J**, **Mitchell DA**, **Moore BS**, **Moreira LM**, **Muller R**, **Neilan BA**, **Nett M**, **Nielsen J**, **O'Gara F**, **Oikawa H**, **Osbourn A**, **Osburne MS**, **Ostash B**, **Payne SM**, **Pernodet J-L**, **Petricek M**, **Piel J**, **Ploux O**, **Raaijmakers JM**, **Salas JA**, **Schmitt EK**, **Scott B**, **Seipke RF**, **Shen B**, **Sherman DH**, **Sivonen K**, **Smanski MJ**, **Sosio M**, **Stegmann E**, **Sussmuth RD**, **Tahlan K**, **Thomas CM**, **Tang Y**, **Truman AW**, **Viaud M**, **Walton JD**, **Walsh CT**, **Weber T**, **van Wezel GP**, **Wilkinson B**, **Willey JM**, **Wohlleben W**, **Wright GD**, **Ziemert N**, **Zhang C**, **Zotchev SB**, **Breitling R**, **Takano E**, **Glockner FO**. 2015. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol **11**:625–631.

122. **Sunaga S**, **Li H**, **Sato Y**, **Nakagawa Y**, **Matsuyama T**. 2004. Identification and characterization of the pswP gene required for the parallel production of prodigiosin and serrawettin W1 in Serratia marcescens. Microbiol. Immunol. **48**:723–728.
123. **Wang H**, **Sivonen K**, **Fewer DP**. 2015. Genomic insights into the distribution, genetic diversity and evolution of polyketide synthases and nonribosomal peptide synthetases. Curr. Opin. Genet. Dev. Elsevier Ltd.
124. **Bushley KE**, **Turgeon BG**. 2010. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. BMC Evol. Biol. **10**:26.
125. **Sasso S**, **Shelest E**, **Hoffmeister D**. 2014. Comments on the distribution and phylogeny of type I polyketide synthases and nonribosomal peptide synthetases in eukaryotes. Proc. Natl. Acad. Sci. **111**:E3946–E3946.
126. **Calteau A**, **Fewer DP**, **Latifi A**, **Coursin T**, **Laurent T**, **Jokela J**, **Kerfeld CA**, **Sivonen K**, **Piel J**, **Gugger M**. 2014. Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. BMC Genomics **15**:1–14.
127. **Caboche S**, **Pupin M**, **Leclère V**, **Jacques P**, **Kucherov G**. 2009. Structural pattern matching of nonribosomal peptides. BMC Struct. Biol. **9**:15.
128. **Bateman A**, **Martin MJ**, **O'Donovan C**, **Magrane M**, **Apweiler R**, **Alpi E**, **Antunes R**, **Arganiska J**, **Bely B**, **Bingley M**, **Bonilla C**, **Britto R**, **Bursteinas B**, **Chavali G**, **Cibrian-Uhalte E**, **Da Silva A**, **De Giorgi M**, **Dogan T**, **Fazzini F**, **Gane P**, **Castro LG**, **Garmiri P**, **Hatton-Ellis E**, **Hieta R**, **Huntley R**, **Legge D**, **Liu W**, **Luo J**, **Macdougall A**, **Mutowo P**, **Nightingale A**, **Orchard S**, **Pichler K**, **Poggioli D**, **Pundir S**, **Pureza L**, **Qi G**, **Rosanoff S**, **Saidi R**, **Sawford T**, **Shypitsyna A**, **Turner E**, **Volynkin V**, **Wardell T**, **Watkins X**, **Zellner H**, **Cowley A**, **Figueira L**, **Li W**, **McWilliam H**, **Lopez R**, **Xenarios I**, **Bougueleret L**, **Bridge A**, **Poux S**, **Redaschi N**, **Aimo L**, **Argoud-Puy G**, **Auchincloss A**, **Axelsen K**, **Bansal P**, **Baratin D**, **Blatter MC**, **Boeckmann B**, **Bolleman J**, **Boutet E**, **Breuza L**, **Casal-Casas C**, **De Castro E**, **Coudert E**, **Cuche B**, **Doche M**, **Dornevil D**, **Duvaud S**, **Estreicher A**, **Famiglietti L**, **Feuermann M**, **Gasteiger E**, **Gehant S**, **Gerritsen V**, **Gos A**, **Gruaz-Gumowski N**, **Hinz U**, **Hulo C**, **Jungo F**, **Keller G**, **Lara V**, **Lemercier P**, **Lieberherr D**, **Lombardot T**, **Martin X**, **Masson P**, **Morgat A**, **Neto T**, **Nouspikel N**, **Paesano S**, **Pedruzzi I**, **Pilbout S**, **Pozzato M**, **Pruess M**, **Rivoire C**, **Roechert B**, **Schneider M**, **Sigrist C**, **Sonesson K**, **Staehli S**, **Stutz A**, **Sundaram S**, **Tognolli M**, **Verbregue L**, **Veuthey AL**, **Wu CH**, **Arighi CN**, **Arminski L**, **Chen C**, **Chen Y**, **Garavelli JS**, **Huang H**, **Laiho K**, **McGarvey P**, **Natale DA**, **Suzek BE**, **Vinayaka CR**, **Wang Q**, **Wang Y**, **Yeh LS**, **Yerramalla MS**, **Zhang J**. 2015. UniProt: A hub for protein information. Nucleic Acids Res. **43**:D204–D212.
129. **Sievers F**, **Wilm A**, **Dineen D**, **Gibson TJ**, **Karplus K**, **Li W**, **Lopez R**, **McWilliam H**, **Remmert M**, **Söding J**, **Thompson JD**, **Higgins DG**. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. **7**.
130. **Eddy SR**. 2009. A new generation of homology search tools based on probabilistic inference. Genome Inform. **23**:205–211.
131. **Nguyen LT**, **Schmidt HA**, **Von Haeseler A**, **Minh BQ**. 2015. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. **32**:268–274.
132. **Kalyaanamoorthy S**, **Minh BQ**, **Wong TKF**, **von Haeseler A**, **Jermiin LS**. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods **14**:587.
133. **Letunic I**, **Bork P**. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. **44**:W242–W245.
134. **Letunic I**, **Bork P**. 2017. 20 years of the SMART protein domain annotation resource. Nucleic Acids Res. **46**:493–496.
135. **Gerlt JA**, **Bouvier JT**, **Davidson DB**, **Imker HJ**, **Sadkhin B**, **Slater DR**, **Whalen KL**. 2015. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. Biochim. Biophys. Acta - Proteins Proteomics **1854**:1019–1037.
136. **Shannon P**, **Markiel A**, **Owen Ozier 2**, **Baliga NS**, **Wang JT**, **Ramage D**, **Amin N**, **Schwikowski B**, **Ideker T**. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2498–2504.
137. **Cortez D**, **Delaye L**, **Lazcano A**, **Becerra A**. 2009. Composition-based methods to identify horizontal gene transfer. Methods Mol. Biol. **532**:215–225.

138. **Gao F**, **Zhang CT**. 2006. GC-Profile: A web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. Nucleic Acids Res. **34**:686–691.
139. **Schmidt H**, **Hensel M**. 2004. Pathogenicity islands in bacterial pathogenesis. Clin. Microbiol. Rev. **17**:14–56.
140. **Juhas M**, **Van Der Meer JR**, **Gaillard M**, **Harding RM**, **Hood DW**, **Crook DW**. 2009. Genomic islands: Tools of bacterial horizontal gene transfer and evolution. FEMS Microbiol. Rev. **33**:376–393.
141. **Morris JH**, **Lotia S**, **Wu A**, **Doncheva NT**, **Albrecht M**, **Pico AR**, **Ferrin TE**. 2015. setsApp for Cytoscape: Set operations for Cytoscape Nodes and Edges. F1000Research 1–10.
142. **Cimermancic P**, **Medema MH**, **Claesen J**, **Kurita K**, **Wieland Brown LC**, **Mavrommatis K**, **Pati A**, **Godfrey PA**, **Koehrsen M**, **Clardy J**, **Birren BW**, **Takano E**, **Sali A**, **Linington RG**, **Fischbach MA**. 2014. Insights into Secondary Metabolism from a Global Analysis of Prokaryotic Biosynthetic Gene Clusters. Cell **158**:412–421.
143. **Tarry MJ**, **Haque AS**, **Bui KH**, **Schmeing TM**. 2017. X-Ray Crystallography and Electron Microscopy of Cross- and Multi-Module Nonribosomal Peptide Synthetase Proteins Reveal a Flexible Architecture. Structure **25**:783–793.e4.
144. **Carroni M**, **Saibil HR**. 2016. Cryo electron microscopy to determine the structure of macromolecular complexes. Methods **95**:78–85.
145. **Skiniotis G**, **Southworth DR**. 2016. Single-particle cryo-electron microscopy of macromolecular complexes. Reprod. Syst. Sex. Disord. **65**:9–22.
146. **Doekel S**, **Gal MFC Le**, **Gu JQ**, **Chu M**, **Baltz RH**, **Brian P**. 2008. Non-ribosomal peptide synthetase module fusions to produce derivatives of daptomycin in Streptomyces roseosporus. Microbiology **154**:2872–2880.
147. **Owen JG**, **Calcott MJ**, **Robins KJ**, **Ackerley DF**. 2016. Generating Functional Recombinant NRPS Enzymes in the Laboratory Setting via Peptidyl Carrier Protein Engineering. Cell Chem. Biol. **23**:1395–1406.
148. **Kim SY**, **Park SY**, **Choi SK**, **Park SH**. 2015. Biosynthesis of polymyxins B, E, and P using genetically engineered polymyxin synthetases in the surrogate Host Bacillus subtilis. J. Microbiol. Biotechnol. **25**:1015–1025.
149. **Lauren E. Brooks,a Sabah Ul-Hasan,a Benjamin K. Chan b MJS**. 2018. Quantifying the evolutionary conservation of genes encoding multidrug efflux pumps in the ESKAPE pathogens to identify antimicrobial drug targets. mSystems **3**:e00024-18.
150. **Nichols D**, **Cahoon N**, **Trakhtenberg EM**, **Pham L**, **Mehta A**, **Belanger A**, **Kanigan T**, **Lewis K**, **Epstein SS**. 2010. Use of ichip for high-throughput in situ cultivation of "uncultivable microbial species". Appl. Environ. Microbiol. **76**:2445–2450.