

La ocurrencia y variación de la promiscuidad en rutas del metabolismo especializado son reveladas por herramientas bioinformáticas evolutivas.

---

Una tesis

presentada a

LANGEBIO-CINVESTAV

---

Para cumplir con los

requerimientos de la obtención del grado

de Doctor en Ciencias

---

Nelly Selém

Enero 2019



Aprobado por el laboratorio  
Evolución de la diversidad metabólica

---

Francisco Barona Gomez



# Agradecimientos

Quiero agradecer por todas las enseñanzas a mis compañeros de trabajo Pablo Cruz, Hilda Ramos, Ernesto Verduzco, Christian Eduardo, César Aguilar, Karina Verdel, Lianet Noda, Adriana Espinosa y Abraham Avelar. A Manu y Vero por su interés en emprender proyectos. A Dany mi veranito, a Ernesto de keri, Langebio, Secretaría inovación, CONABIO, BetterLab, Strain y DNAbits. A Abraham Avelar por sus enseñanzas, revisión y aportaciones a este trabajo.



# Prefacio

This is an example of a thesis setup to use the reed thesis document class.



# Índice general

<b>Introducción</b>	<b>1</b>
0.0.1. Funcion biologica de la promiscuidad enzimatica	2
0.0.2. Relacion del pangenoma con la promiscuidad enzimatica	3
0.0.3. Modelos bioinformaticos de promiscuidad	4
0.0.4. Promiscuidad in vitro y promiscuidad in vivo	5
0.0.5. El papel de la dinamica molecular en la promiscuidad	5
0.0.6. Modelo biológico diversidad de Actinobacteria	6
0.0.7. Modelo metabolico biosintesis de aminoacidos.	6
0.0.8. Modelo biológico	7
0.0.9. Subsistemas metabolicos	7
0.1. Modelos computacionales	10
<b>Antecedentes</b>	<b>11</b>
0.2. La promiscuidad puede abordarse a distintos niveles incluyendo enzima, familia y ruta de metabólica.	12
0.3. Antecedentes conceptuales	12
0.4. El establecimiento de un marco de conservación permite distinguir cambios	14
0.5. La genómica comparativa como herramienta en la distinción de familias y enzimas promiscuas que participan en el metabolismo especializado.	15
0.6. La genómica comparativa como herramienta en la priorización de clusters promiscuos	16
0.6.1. Expansion y contextos genomicos como herramienta de anotacion funcional	16
0.6.2. Contexto y vecindades genomicas	18
0.7. Estudio de una familia PriA	18
0.7.1. Caracterizacion in vivo	18
0.7.2. Caracterizacion bioquimica in vitro.	19
0.7.3. Modelado de dinamica molecular	19
<b>Pregunta biológica</b>	<b>21</b>
<b>Objetivos</b>	<b>23</b>
0.8. Objetivo General	23
0.9. Objetivos particulares	23

<b>Estrategias</b>	25
Obtener informacion genomica del phylum Actinobacteria.	25
0.9.1. Anotar consistentemente las secuencias codificantes de estos genomas.	25
Establecer las relaciones filogeneticas de los genomas colectados.	25
0.9.2. La promiscuidad en familias enzimaticas.	25
Promiscuidad in vitro dentro de miembros de una familia promiscua de enzimas.	26
Sistematizar Evomining para convertirla una plataforma descargable y utilizable en cualquier set de datos bacterianos relacionados taxonomicamente proporcionados por el usuario.	26
Seleccionar miembros homologos de la familia de enzimas.	26
Medir cineticas enzimaticas, contexto genomico, vecindad genomicica, flexibilidad y numero de conformeros.	26
Determinar posibles correlaciones entre los datos producidos.	26
0.9.3. Desarrollar una metodologia para la deteccion in vivo de promiscuidad enzimatica.	27
<b>Metodología</b>	29
0.9.4. La promiscuidad en familias enzimaticas.	29
Actinobacteria genomica	29
Annotation	29
Genomic DB phylogeny	29
0.9.5. Identificar cambios en la vecindad genomica en familias selectas de enzimas de metabolismo central.	30
Organizar y presentar los datos en una plataforma.	31
0.9.6. Promiscuidad in vitro	32
Datos cineticos:	32
Dinamica molecular	32
0.9.7. Promiscuidad in vivo	33
0.9.8. Consideraciones	33
<b>Capítulo 1: Desarrollo de Orthocore e implementación de otras herramientas computacionales para entender el pangenoma de un linaje genómico.</b>	37
1.1. La distribución de la función metabólica de las familias del pangenoma depende de la variabilidad del linaje seleccionado.	38
1.2. El core conservado permite la reconstrucción de filogenias complicadas	42
1.3. El algoritmo de Orthocore	42
1.3.1. Componentes técnicos de Orthocore	43
1.3.2. Aplicaciones de Orthocore, identificación del core conservado y de familias de genes marcadores.	44
1.3.3. Genes de islas de patogenicidad de <i>Salmonella</i> en México están conservados en la mayoría de los genomas.	45

1.3.4. Nostoc provenientes del metagenoma de cíadas se agrupan Cyanobacteria . . . . .	45
1.3.5. Identificación de genes marcadores de <i>Clavibacter michiganensis</i>	47
1.4. Clavisual: Identificación de genes marcadores a un cierto porcentaje de grupos seleccionados . . . . .	50
1.4.1. El pangenoma de <i>Clavibacter Michiganensis</i> es abierto . . . . .	50
1.5. Relación entre genes marcadores, Orthocore y la promiscuidad enzimática.	51
<b>Capítulo 2: EvoMining como herramienta para identificar el origen y el destino metabólico de familias enzimáticas . . . . .</b>	<b>55</b>
2.1. Introducción . . . . .	55
2.1.1. Las copias extra de familias enzimáticas que son reclutadas para una nueva función están relacionadas con la promiscuidad. . . . .	55
2.2. EvoMining es un método para encontrar BGC no tradicionales . . . . .	56
2.2.1. EvoMining es un paradigma que permite ubicar copias extra de familias enzimáticas y organizarlas visualmente acorde a eventos evolutivos. . . . .	56
2.3. Algoritmos y bases de datos de EvoMining 2.0 . . . . .	57
2.3.1. Algoritmo de expansión y clasificación de grupos de homólogos.	58
2.3.2. Algoritmo de reconstrucción filogenética y visualización . . . . .	61
2.3.3. Actualizaciones de las bases de datos de EvoMining . . . . .	63
2.4. Detectamos patrones de neofuncionalización a BGC en distintos linajes y en familias enzimáticas con distintos orígenes metabólicos OR EvoMining detecta distintas dinámicas evolutivas de las enzimas metabólicas que dependen de la familia enzimática y del linaje taxonómico. . . . .	65
2.4.1. Los perfiles de expansión de las proteínas dependen del linaje.	65
2.4.2. El shell genome posee expansiones en sus familias enzimáticas	67
2.4.3. GDH y ALS en el cluster escitonemina ejemplifican como familias pertenecientes a un mismo BGC pueden tener distintos patrones de expansión. . . . .	70
2.4.4. El cluster de escitonemina es un cluster promiscuo. . . . .	74
2.4.5. EvoMining en RNA transferasas . . . . .	78
2.4.6. La familia <i>tauD</i> tiene expansión y reclutamiento tanto en Actinobacteria como en Pseudomonas . . . . .	78
2.4.7. Rimosamide y Detoxin tienen en común la enzima TauD, . . . . .	79
2.5. Consideraciones finales sobre el uso de EvoMining . . . . .	80
<b>Capítulo 3: Desarrollo de CORASON como herramienta para organizar clusters biosintéticos y otras vecindades genómicas conservadas. . . . .</b>	<b>83</b>
3.1. Algoritmo y características de CORASON . . . . .	85
3.2. Las familias de BGCs están formadas por variantes del BGC de referencia.	86
3.3. Aplicaciones de CORASON en Actinobacteria y Pseudomonas . . . . .	87
3.4. Estas familias pueden clasificarse . . . . .	87
3.5. Un BGC reportado puede tener muchas variantes que conforman una familia de BGCs . . . . .	88

3.6. CORASON analiza la conservación del contexto genómico de los EvoMining hits . . . . .	91
3.7. El contexto de una rama divergente de <i>tauD</i> está conservado en Pseudomonas . . . . .	92
3.8. CORASON se adaptó a la herramienta BiG-SCAPE de clasificación de familias de BGCs . . . . .	92
3.9. BiG SCAPE y CORASON identificaron nuevos productos variantes de la familia de BGCs Rimosamide - Detoxin en Actinobacteria . . . . .	92
3.10. CORASON sugiere que las familias detoxin y rimosamide pertenecen a un amplio clan de familias dedicadas a la síntesis de péptidos. . . . .	96
3.11. Clados selectos del árbol de CORASON que contiene a las familias rimosamide/detoxin contienen diversidad génica que correlaciona con novedad química. . . . .	96
3.11.1. El clado P450/enoyl agrega una heptanamida al core molecular detoxin/rimosamide . . . . .	96
3.11.2. El superclado spectinomycin/ detoxin-rimosamide-clan produce cinco variantes de detoxin. . . . .	97
3.11.3. El clado <i>Amycolatopsis P450</i> produce cinco variantes de detoxin.	98
3.12. CORASON permitió expandir las capacidades de EvoMining y explorar la promiscuidad de familias de BGCs de enzimas divergentes de metabolismo central. . . . .	98
Discusión . . . . .	99
<b>Capítulo 4: La familia PriA . . . . .</b>	<b>101</b>
4.1. PriA como modelo de familia enzimática donde las expansiones no son condición necesaria para la promiscuidad. . . . .	102
PriA en EvoMining . . . . .	102
4.2. Análisis de contextos genómicos de PriA/HisA en distintos linajes utilizando CORASON como herramienta de visualización. . . . .	109
4.2.1. Contextos genómicos de HisA en Actinobacteria . . . . .	109
4.2.2. Contextos genómicos de HisA en Archaea . . . . .	110
4.2.3. Contextos genómicos de HisA en saxitoxin . . . . .	111
4.3. Evolución molecular y estructural de PriA . . . . .	112
4.3.1. Al transformar una subHisA en una PriA mediante mutaciones no se observó ninguna trayectoria creciente para ambos sustratos (darwiniana) . . . . .	112
4.3.2. Los residuos que covarián en el registro evolutivo de PriA permiten una reconstrucción aproximada de su estructura tridimensional	116
4.4. Afinidad de enzimas selectas por sustratos químicamente parecidos a PRA y PROFAR . . . . .	122
4.4.1. Selección de secuencias de PriA o familias relacionadas para docking con sustratos similares a lo nativos . . . . .	127
4.4.2. El análisis de PriA a nivel estructural sugiere que GTP es el sustrato más afín . . . . .	129
4.4.3. GTP es el sustrato por el que PriA muestra mayor afinidad .	130

4.4.4. Dinámica molecular vs datos experimentales . . . . .	133
4.5. PriA en cinéticas enzimáticas no tradicionales. . . . .	136
4.5.1. PriA puede tener actividad en GTP . . . . .	137
4.5.2. Cinéticas simultáneas para PRA y ProFAR . . . . .	139
4.6. Cinéticas . . . . .	139
4.7. Gráfico de pendientes . . . . .	143
4.8. Interpolación con Michaelis Menden . . . . .	143
<b>Metodología . . . . .</b>	<b>145</b>
4.1. Protocolos para usar Orthocore, myRAST, fastOrtho, Clavigenomics, y BPGA . . . . .	145
4.2. Orthocore . . . . .	146
<b>Capítulo 5: Conclusión . . . . .</b>	<b>147</b>
5.1. Discusion . . . . .	147
More info . . . . .	147
<b>Apéndice A: The First Appendix . . . . .</b>	<b>149</b>
In the main Rmd file: . . . . .	149
In : . . . . .	149
<b>Apéndice B: The Second Appendix, Open source code on this document</b>	<b>151</b>
B.1. R markdown . . . . .	151
B.2. Docker . . . . .	151
B.3. Git . . . . .	152
B.4. Connect GitHub and DockerHub . . . . .	152
B.5. Additional resources . . . . .	152
<b>Apéndice C: El tercer Apéndice, Protocolos experimentales utilizados en el estudio de PriA . . . . .</b>	<b>155</b>
C.1. Materiales . . . . .	155
C.2. Purification . . . . .	155
C.2.1. Centrifugation . . . . .	155
C.2.2. Sonication . . . . .	156
C.3. Protein Quantification . . . . .	158
C.3.1. Bradford calibration curve . . . . .	158
C.3.2. Enzyme quantification . . . . .	160
C.4. Concentration on microMolar . . . . .	161
<b>References . . . . .</b>	<b>163</b>



# Índice de cuadros

2.1. Homólogos de <i>tauD</i> en BGC reportados en MIBiG. . . . .	79
4.1. Árboles EvoMining de PriA en MicroReact . . . . .	105
4.2. Árboles EvoMining de PriA en MicroReact . . . . .	119
4.3. Secuencias selectas de PriA y familias relacionadas para el análisis de docking . . . . .	128
4.4. Estructuras cristalográficas disponibles de PriA y familias relacionadas . . . . .	129
4.5. Enzymes docking . . . . .	135
4.6. Formatos de descarga disponibles en myRAST . . . . .	146



# Índice de figuras

1.	PriA isomeriza los sustratos PRA y ProFAR . . . . .	8
2.	Antecedentes Conceptuales . . . . .	11
3.	Antecedentes Conceptuales . . . . .	13
1.1.	Orthocore calcula el <i>core genome</i> de un linaje y construye con él una filogenia . . . . .	37
1.2.	El metabolismo en el Pangenoma . . . . .	39
1.3.	Core y genes Marcadores . . . . .	41
1.4.	Los mejores hits n-direccionales generalizan a los <i>Bidirectional Best Hits</i> . . . . .	43
1.5.	Arbol filogenético de <i>Nostoc</i> construido utilizando la selección de genes del <i>core conservado</i> . . . . .	46
1.6.	Los genes del cluster biosintético clv son marcadores de <i>Clavibacter</i> . . . . .	47
1.7.	El Cluster <i>clv</i> es un marcador de la especie <i>Clavibacter Michiganensis</i> en el orden <i>Micrococcales</i> . . . . .	48
1.8.	El <i>core genome</i> de <i>Micrococcales</i> ayuda a identificar genes marcadores . . . . .	49
1.9.	Clavisual organiza cepas de <i>Clavibacter</i> con sus metadatos. . . . .	51
1.10.	El pangenoma de <i>Clavibacter</i> es abierto según el análisis de BPGA. . . . .	52
1.11.	Diagrama de Venn del pangenoma de genomas selectos del género <i>Clavibacter</i> . . . . .	53
2.1.	Existen cluster biosintéticos no clasificados (other) entre los reportados en MIBiG . . . . .	58
2.2.	EvoMining Algorithm . . . . .	59
2.3.	Seed genomes . . . . .	66
2.4.	Expansion patterns in 42 conserved families . . . . .	69
2.5.	EvoMining en cuatro linajes genómicos . . . . .	71
2.6.	EvoMining aplicado a la rama de escitoneminaa . . . . .	72
2.7.	EvoMining Algorithm . . . . .	75
2.8.	EvoMining Algorithm . . . . .	77
2.9.	AroB en cianobacteria . . . . .	78
3.1.	La herramienta corason localiza familias de clusters biosintéticos en un linaje genómico partiendo de un cluster y un gen de referencia. . . . .	84
3.2.	CORASON es una herramienta en línea de comandos para organizar filogenéticamente la variación de una familia de clusters. . . . .	85
3.3.	EvoMining Algorithm . . . . .	88

3.4. EvoMining Algorithm . . . . .	90
3.5. EvoMining Algorithm . . . . .	91
3.6. EvoMining Algorithm . . . . .	93
3.7. Algoritmo de CORASON Algorithm . . . . .	94
3.8. Rimosamide . . . . .	95
4.1. La reacción que cataliza PriA es una isomerización donde abre un anillo de cinco carbonos. . . . .	101
4.2. Expansiones de PriA en Actinobacteria, Cyanobacteria, Pseudomonas y Archaea . . . . .	103
4.3. Copias extras de PriA en Actinobacteria, Cyanobacteria, Pseudomonas y Archaea . . . . .	104
4.4. Árboles de destino metabólico de PriA en Actinobacteria, Cyanobacteria, Pseudomonas y Archaea según EvoMining . . . . .	106
4.5. PriA participa en la síntesis del antibiótico ada en Actinomadura . . . . .	108
4.6. Contextos de PriA en Actinobacteria . . . . .	109
4.7. Contextos de PriA en Archaea . . . . .	110
4.8. HisA en saxitoxin, un cluster de Cyanobacteria . . . . .	112
4.9. Rutas desde una subHisA hasta una variante con 9 mutaciones . . . . .	113
4.10. Positive increments on PRA . . . . .	115
4.11. Superposición de estructura de PriA generado por Folding con la estructura cristalográfica . . . . .	118
4.12. Visualización de los acoplamientos en la familia PriA . . . . .	120
4.13. Miembros de PriA que poseen variantes los residuos catalíticos D11, D130121	
4.14. Substatos químicamente similares a los sustratos de PriA (parte 1) . .	123
4.15. Substatos químicamente similares a los sustratos de PriA (parte 2) . .	124
4.16. Substatos químicamente similares a los sustratos de PriA (parte 3) . .	125
4.17. Substatos químicamente similares a los sustratos de PriA (parte 4) . .	126
4.18. Heatplot del docking de enzimas relacionadas a PriA vs posibles substratos	131
4.19. gel . . . . .	136
4.20. Scoe and non functional Scoe PriA acting on dGTP . . . . .	137
4.21. Scoe and non functional Scoe PriA acting on dGTP . . . . .	137

# **Abstract**

The preface pretty much says it all.

Second paragraph of abstract starts here.



# Dedicatoria

A mis seres queridos.



# Introducción

Las enzimas catalizan reacciones químicas transformando sustratos en productos. Durante el siglo XX, las enzimas fueron percibidas como catalizadores altamente específicos, sin embargo, esta percepción cambió con el descubrimiento de que pueden. Esta capacidad de catalizar varias funciones químicas se conoce como la promoción de la enzima. Escherichia coli contiene al menos 404 enzimas promiscuas. La relevancia de la promiscuidad radica tanto en su papel como mecanismo de evolución de la función enzimática, así como en la necesidad de su detección para la corrección de modelos de flujo metabólico y la determinación de efectos secundarios en drogas farmacológicas. A pesar de su frecuencia e importancia aún se está en el proceso de entender las causas y las características observables de la promiscuidad enzimática.

Para estudiar la promiscuidad es necesario contar con una definición, algunos autores emplean el término promiscuidad para describir actividades enzimáticas distintas a la función principal [1] otros lo ven como una actividad secundaria fortuita [2] que pudo aparecer de forma accidental o inducida artificialmente [3]. Otros más, cuando una enzima puede operar sobre un amplio rango de sustratos, prefieren llamarla multiespecífica [1]. A la acción de realizar distintas funciones catalíticas, ya sea al catalizar varias reacciones químicas o bien una misma reacción en sustratos diferentes se le conoce como promiscuidad enzimática [4]. Existen varios tipos de promiscuidad enzimática.

Por sustrato cuando la reacción es la misma pero se lleva a cabo en distintos sustratos ejemplo la familia PriA [5] y la familia de betalactamasas [6]

Catalítica cuando la enzima utiliza diferentes mecanismos de reacción y/o residuos catalíticos, e.g. la quimotripsina puede catalizar reacciones de amidasa y fosfotriesterasa en un mismo sitio activo. [4] Por condiciones del entorno, cuando la enzima cambia su conformación dependiendo de las condiciones químicas y físicas presentes como pH, temperatura, solventes orgánicos y salinidad e.g. algunas lipasas pueden actuar como sintetizadoras de ésteres en lugar de hidrolasas en presencia de solventes orgánicos [3].

Este trabajo se enfocará a la promiscuidad por sustrato, entendiendo así que la enzima es capaz de catalizar la misma reacción química en al menos dos sustratos. La promiscuidad por sustrato es importante en términos evolutivos, por ejemplo la enzyme commission number (EC) separa las enzimas en clases, a cada enzima se

asignan 4 digitos, los tres primeros corresponden a la reaccion y el ultimo al sustrato; el mayor numero de sustratos (4306 clases) que de reacciones quimicas (234 en el tercer nivel) sugiere que la mayor variacion evolutiva se da a nivel de sustrato y no de reaccion [8]. Otra evidencia de la importancia de la multiespecificidad por sustrato esta en el descubrimiento de las superfamilias, enzimas mecanistica y estructuralmente relacionadas que divergen en su afinidad por sustrato [9]

Si bien existen familias de enzimas con alta especificidad por sustrato, otras familias como el citocromo P450 [11] y las beta lactamasas [13] son promiscuas. Es posible que la vision previa de alta especificidad se deba a que las primeras rutas metabolicas estudiadas pertenecen al metabolismo central, donde la especificidad puede haber sido favorecida por presiones de seleccion [14]. Esta vision ha cambiado debido al conocimiento de mas enzimas con multifuncionalidad [15], sin afectar la eficiencia catalitica por la funcion primaria [16]. En 1976 el interes por la promiscuidad comenzó por su influencia en la evolucion de la funcion enzimatica[17], las aproximaciones variaron desde la aparicion de la sintesis funcional [19], cuando la disponibilidad de genomas permitio la combinacion de analisis filogeneticos con tecnicas de biologia molecular, bioquimica y biofisica (Fig 1). En 2003 la biofisica de las proteinas entra en escena al postularse que la diversidad conformacional durante la dinamica molecular debe incidir en la aceptacion de distintos sustratos. Recientemente se ha investigado su papel en efectos secundarios en drogas farmacologicas [20]. Entre 2005 y 2010 se avanza del estudio de una sola familia enzimatica hacia el interes por propiedades globales, por ejemplo dado un genoma se investiga la distribucion de familias promiscuas en subsistemas metabolicos. En estos años, surge el desarrollo de indices que reflejen las caracteristicas bioquimicas de enzimas promiscuas. En 2010, comienzan los intentos por desarrollar un metodo computacional de prediccion de promiscuidad. Desde 2012 a la fecha, a la par que las aproximaciones bioinformaticas se multiplican, se desarrollan investigaciones de aspectos biofisicos, bioquimicos y evolutivos de enzimas promiscuas reafirmando que todos estos aspectos estan relacionados al fenomeno. En las siguientes secciones se describiran trabajos importantes sobre la relacion que guarda la promiscuidad con expansiones genomicas y flexibilidad molecular. Ademas se hablara sobre analisis bioquimicos y metabolicos para la descripcion del fenomeno.

### 0.0.1. Funcion biologica de la promiscuidad enzimatica

¿Por que existe la promiscuidad enzimatica? Se tiene evidencia de dos papeles biologicos: el primero proporcionar robustez a la red metabolica de un organismo mediante redundancia de reacciones de otras enzimas; el segundo permitir plasticidad evolutiva, es decir materia prima para la adaptacion a variaciones ambientales [16] mediante la adquisicion de nuevas funciones quimicas. Respecto a la robustez, se probó que sobreexpresar enzimas promiscuas puede rescatar perdidas genicas [27]. De 104 knockout sencillos de genes esenciales para *E. coli* K-12, 20 % de las auxotrofias pudieron ser suprimidas por la sobreexpresión de plasmidos que contenian enzimas promiscuas. Otro ejemplo que aporta a la robustez es PriA, enzima de la ruta de histidina que

realiza en la ruta del triptofano la reaccion E.C. 5.3.1.24 [5]. En cuanto a la plasticidad se propone que para que la promiscuidad pueda dar origen a la aparicion de nuevas funciones la actividad promiscua debe proveer una ventaja fisiologica inmediata para poder ser seleccionada positivamente, ademas una vez que una funcion promiscua se vuelva relevante se debe poder mejorar mediante pocas mutaciones derivando en el intercambio entre la actividad promiscua y la principal[1].

Aun cuando el producto de la promiscuidad genera metabolitos que no se integran al metabolismo central de la celula, su efecto es positivo ya que estos metabolitos podrian colaborar a la adaptacion al entorno participando por ejemplo en una relacion de simbiosis o de competencia con otros organismos. Este tipo de metabolitos, por lo general, no son dañinos [28] y pueden servir como bloques de construccion para vias metabolicas nuevas [31]. La respuesta inmediata de adaptacion de un organismo podria ser una consecuencia de su grado de promiscuidad.

### **0.0.2. Relacion del pangenoma con la promiscuidad enzimática**

El pangenoma es el contenido génico total de un linaje taxonómico. Las familias génicas de un pangenoma pueden clasificarse según su frecuencia de presencia/ausencia en cada genoma del linaje. De acuerdo a esta clasificación los principales grupos de familias génicas en un pangenoma son el *core*, el *shell* y el *cloud* también conocido como *dispensable* o *accessory* genome. El *core genome* es el conjunto de familias con presencia en todos los genomas del linaje. Por ejemplo, tanto la secuencia de la subunidad 16s de rRNA así como los diversos genes ribosomales suelen estar en el *core* de la gran mayoría de linajes bacterianos. El *shell genome* es el grupo de familias presentes en la mayoría de los genomas pero no en todos. En el *shell* se ubican por ejemplo familias que estaban en el *core genome* pero que algunas bacterias del linaje sufrieron una dinámica de pérdida génica. Mientras que el *cloud genome* o *dispensable genome* es aquel grupo de familias que sólo ocurre en unos cuantos genomas del linaje.

En el Dominio Bacteria se estima que entre al rededor de 200 secuencias están altamente conservadas. Si bien no están en el *core*, estas secuencias están compartidas por 90 % de los genomas [34]. El *core* depende de los genomas seleccionados, entre menos amplio sea el rango filogenético electo mayor será el tamaño del *core*. Dada su conservacion el *core genome* puede utilizarse para trazar mejores relaciones filogeneticas que las obtenidas con el uso exclusivo de marcadores como la subunidad 16s del RNA ribosomal o el gen *rpoB*. El pangenoma es el conjunto complemento del core genome, es decir todas aquellas secuencias que estan ausentes de uno o mas organismos del grupo y por lo tanto no son necesarias para todos, sino solo posiblemente para el organismo que las posee. Como en el pangenoma la presion de seleccion esta relajada respecto al core-genome [14] es el conjunto donde la plasticidad genomica tiene facilidades para desarrollarse.

Esta idea puede restringirse a subsistemas metabolicos para identificar genes cuyas enzimas estan en proceso de cambio de funcion quimica, por ejemplo, en este trabajo se encontro que el gen *trpF* esta presente en solo 49 de 290 genomas analizados del genero *Streptomyces* por lo que se encuentra en el pangenoma de triptofano de este genero taxonomico, posiblemente adquiriendo una nueva funcion [31]. Para evitar problemas tecnicos del calculo del pangenoma existen otros modelos de medicion de variabilidad del genomica entre especies bacterianas [35].

### 0.0.3. Modelos bioinformaticos de promiscuidad

Con el fin de reducir la inversion en el proceso de experimentacion, se han implementado en los ultimos años algoritmos computacionales para predecir promiscuidad enzimatica [36]. Estos procedimientos cuentan con un conjunto de aprendizaje, unos descriptores del conjunto, una fase de ajuste de parametros y finalmente una prediccion. En 2010, Carbonell propone un algoritmo de soporte vectorial basado en subsecuencias de distinto tamaño que llama huellas moleculares. En este trabajo aplicado sobre 500,000 proteinas reportadas en la enciclopedia de Kyoto de genes y genomas (KEGG) se reporta 85 % de exito en deteccion de enzimas promiscuas anotadas en KEGG. En 2012, Cheng compara los metodos de random forest y soporte vectorial en 6799 proteinas provenientes de la base de datos Universal Protein Resource (UniProt). Las enzimas son descritas con subsecuencias de aminoacidos incorporando ademas caracteristicas biofisicas como polaridad. Se utiliza como grupo de control a familias de enzimas donde nunca se ha reportado una enzima promiscua.

Un aspecto no considerado en estos metodos es que hay familias de enzimas con alta identidad de secuencia entre sus miembros, con cambios bruscos en promiscuidad, debidos por ejemplo a la dinamica genomica [41], lo que dificulta que considerar solo la secuencia conlleve a buenos predictores de promiscuidad. Cuando se obtiene una prediccion positiva utilizando los modelos existentes, lo que significa es que dada esa secuencia, en su familia se conoce previamente un elemento promiscuo y que ademas sus subsecuencias de cierto tamaño son suficientemente similares. Estos enfoques no pueden predecir de novo, en familias donde la promiscuidad no ha sido previamente detectada experimentalmente, pues no consideran aspectos evolutivos ni mecanisticos de las enzimas.

Otra limitante a los enfoques descritos es que mezclan en su conjunto de entrenamiento fenomenos distintos de promiscuidad. Cheng p. g. incluye enzimas moonlight que si bien poseen funciones adicionales a la catalizacion, son distantes a las enzimas promiscuas [2]. Ademas en ambos casos mezclan en el mismo conjunto enzimas bacterianas y eucariotas, con lo que si existia una huella basada en secuencia entonces esta puede diluirse por la gran distancia taxonomica entre estos grupos (Tabla 1).

#### 0.0.4. Promiscuidad in vitro y promiscuidad in vivo

La ganancia de promiscuidad no solo puede entenderse como la capacidad de convertir mas sustratos [36], sino tambien como la mejora de la capacidad catalitica respecto a ellos. El I-index [12], esta definido como un rango de valores entre 0 y 1 que tiende a 1 entre mas parecida sea la actividad de la enzima sobre distintos sustratos, la capacidad catalitica es medida en terminos del cociente de Michaelis - Menten  $\frac{K_{cat}}{K_m}$ . El indice ha sido utilizado para predecir la afinidad por sustrato del citocromo P450 [22]. Una limitante del indice  $I$  es que se deben conocer los sustratos a los que la enzima es afin; sin embargo se puede sospechar que una enzima ha ganado promiscuidad aun sin conocer sus potenciales sustratos. Otro punto a señalar es que las variables K<sub>cat</sub>,K<sub>m</sub> son mediciones realizadas in vitro y no se consideran todos los sustratos presentes in vivo. Para solventar esta dificultad e investigar variaciones de sustratos nativos se pueden buscar productos similares a los ya conocidos por medio de analisis metabolicos [43] como los empleados en la detección de rutas no conservadas en la biosintesis de productos naturales [44]. En particular para este fin se ha utilizado espectrometria de masas MS/MS, [43] combinada con molecular networking para identificar productos similares [46]

#### 0.0.5. El papel de la dinamica molecular en la promiscuidad

La estructura tridimensional de una proteina es obtenida mediante previa purificacion y cristalizacion. Aunque mucho se ha hablado de la relacion estructura funcion, al cristalizar se obtienen estados conformacionales homogeneos, que bien pueden no ser la unica conformacion que adopta la proteina en solucion. [48]. En particular en el problema de promiscuidad, se ha observado que la variacion funcional no queda obviamente reflejada en la variacion estructural, lo que sugiere un rol significativo para la dinamica molecular [49]. Se postula que un aspecto de la dinamica molecular relevante para la diversificacion de especificidad por sustrato es el numero de conformeros [50]. Por ejemplo, en la actinobacteria *Corynebacterium diphtheriae* parece que el contexto genomico correlaciona con perdida de promiscuidad de PriA ya que al poseer el genoma una copia de trpF, la enzima perdió esta función química conservando solo la función EC 5.3.1.16 correspondiente a la ruta de histidina. Esta sub-funcionalización se refleja en la perdida de estados conformacionales cambiando desde 1 estado en *C. diphtheriae* hasta 4 presentes en la dinámica de PriA de *M. tuberculosis* [41].

Las regiones rígidas de una enzima proporcionan orientación adecuada con respecto a los grupos catalíticos, mientras que las regiones flexibles permiten al sitio activo adaptarse a los sustratos con diferentes formas y tamaños [2]. Esta consideración sugiere que la flexibilidad del sitio activo es otra característica de la dinámica molecular a considerar para obtener información de la capacidad de ligación de una enzima a distintos sustratos [51]. Recientemente el índice de flexibilidad dinámica (dfi) se utilizó como una medida cuantitativa basada en la respuesta a perturbaciones de aminoácidos

(PRS). Este indice se incremento en regiones cercanas al sitio activo de beta lactamasas promiscuas respecto al correspondiente dfi de  $\beta$  -lactamasas especialistas existentes [13].

#### **0.0.6. Modelo biologico diversidad de Actinobacteria**

Al escoger un conjunto acotado para investigar familias de enzimas promiscuas se debe recordar que la funcionalidad es jerarquica por lo que para mejorar la anotacion, es deseable reflejar el proceso evolutivo y restringirse a un grupo de organismos taxonomicamente relacionados [52]. Actinobacteria es un phylum que posee promiscuidad tanto en el metabolismo periferico como en el core metabolico. Entre datos publicos (NCBI) y privados estan disponibles alrededor de 1200 genomas no redundantes de especies de Actinobacteria. Como punto de partida, se han estudiado las relaciones filogeneticas y grupos de ortologia [53], en particular en Actinobacteria para identificar relaciones entre las familias del phylum, se obtuvieron arboles multilocus de entre 100 y 157 genomas [55]. Estos estudios sugieren como separar los genomas disponibles para hacer el calculo de grupos de ortologia. Finalmente, se han realizado estudios de plasticidad genomica en Streptomyces considerando 5 y 17 organismos de los 300 genomas disponibles en la actualidad [57] donde reportan 2,018 familias en el core genome y 32,574 en el pangenoma.

#### **0.0.7. Modelo metabolico biosintesis de aminoacidos.**

Al hacer el calculo vemos que Streptomyces, un genero del phylum Actinobacteria cuenta en su genoma con un promedio de 8316 secuencias codificantes segun la especie. Gran parte de estas secuencias pueden ser agrupadas en subsistemas metabolicos como metabolismo de carbohidratos o de lipidos; de estos subsistemas uno de los mas amplios es el metabolismo de aminoacidos con entre 429 y 910 secuencias segun el organismo. La sintesis de aminoacidos es un subsistema presente en todas las especies pero con suficientes variaciones que permiten hacer observaciones evolutivas. En un gran numero de Actinobacterias las rutas de histidina y triptofano de 7 y 11 pasos respectivamente convergen en una enzima bifuncional llamada PriA, que realiza tanto la funcion de HisA como la de TrpF [5]. La cantidad de familias en el subsistema de metabolismo de aminoacidos, su variabilidad, su conservacion entre distintos grupos taxonomicos y la existencia de estos ejemplos en Actinobacteria lo posicionan como un buen punto de partida para la busqueda de promiscuidad tanto de familias promiscuas como de miembros promiscuos de las mismas.

En las cuatro decadas de estudio de la promiscuidad enzimatica, hemos aprendido que es un fenomeno distribuido en distintos subsistemas metabolicos [59] y que su existencia puede deberse tanto al desarrollo de nuevas funciones para fines adaptativos [17], como al rescate de una funcion perdida [27]. Por ello la dinamica de perdida y ganancia de genes asociada al contexto genomico en bacterias se relaciona con

---

cambio en la funcion enzimatica [61]. Precisando, respecto a la ganancia de genes, se postula que la bifuncionalidad precede la duplicacion [62]. Lo que implica que dada una duplicacion muy posiblemente previamente la promiscuidad estuvo presente [63].

Se han desarrollado tecnicas bioquimicas y metabolicas de medicion [12], asi como algoritmos computacionales de prediccion de promiscuidad [36]. Un aspecto a mejorar dentro del modelado es la restriccion del conjunto de estudio a un grupo taxonomico tan reducido que exista congruencia en las familias de ortologia y a la vez tan amplio que permita observar efectos evolutivos; el phylum Actinobacteria ha probado tener ejemplos de promiscuidad. Si bien la secuencia no ha sido suficiente para la correcta prediccion de promiscuidad [42], es posible que dentro de las tecnicas computacionales la flexibilidad durante la dinamica molecular este correlacionada con la promiscuidad de los miembros de una familia [48].

### 0.0.8. Modelo biológico

De los mas de mil genomas actualmente disponibles de Actinobacterias, se seleccionaron 888 (correspondientes a 49 familias), que no estan excesivamente fragmentados; es decir con un estimado de al menos 5 genes por contig (Tabla 2). Estos genomas fueron divididos en tres grupos ([http://pubseed.theseed.org/wc.cgi?request=show\\_ots&base=/homes/nselem/Data/CS](http://pubseed.theseed.org/wc.cgi?request=show_ots&base=/homes/nselem/Data/CS)), uno de ellos correspondiente a Streptomycetaceae, la familia con la mayor cantidad de genomas disponibles; los otros dos grupos siguieron la taxonomia propuesta por Gao & Gupta en 2012. En el grupo de 290 genomas de Streptomycetaceae 2,126,832 ORFS fueron clasificados en 288,390 familias; de las 919,292 ORF del grupo I de Actinobacteria resultaron 269,406 familias. Las relaciones taxonomicas fueron corroboradas con algoritmos propios basados en best bidirectional hits (BBH).

### 0.0.9. Subsistemas metabolicos

Los operones his y trp de histidina [65] y triptofano [66] respectivamente, participantes del metabolismo de aminoacidos estan ampliamente distribuidos en los organismos bacterianos. En Actinobacteria la familia promiscua PriA participa en ambas rutas biosinteticas Figure 1, para su estudio se han generado datos bioquímicos Figure 0.0.9, genómicos y estructurales. En bacterias gram negativas estan presentes los operones his y trp y en lugar de PriA su familia homologa HisA. PriA comprende un conjunto de subfamilias en Actinobacteria. En *Streptomyces*, el gen *trpF* se desplaza de la vecindad genomica de *trp*, con lo que el homologo de *hisA* gana promiscuidad aunque con baja actividad de *TrpF*, a esta subfamilia se le llama PriB [67]. En otras Actinobacterias *trpF* se pierde totalmente y la familia homologa de *HisA*, se vuelve promiscua [5] realizando tanto la funcion quimica correspondiente a *HisA* como la de *TrpF*. Finalmente en la familia sub*HisA* se pierde la funcion *TrpF* debido posiblemente a la ganancia del

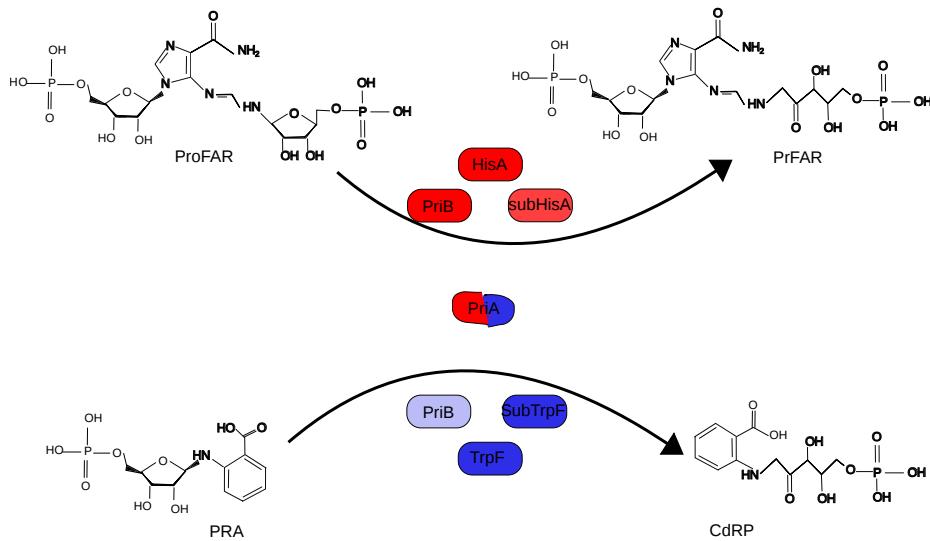


Figura 1: PriA cataliza la isomerización de los sustratos nativos ProFAR y PRA. Convertir ProFAR es un paso de la síntesis de histidina catalizado por hisA en otros linajes bacterianos. Analogamente la isomerización de PRA es catalizada por trpF en otros linajes. En Actinobacteria PriA se ha subdividido en subfamilias. Entre ellas subHisA, que se ha subfuncionalizado a la función HisA, subTrpF que se ha subfuncionalizado a la función TrpF y finalmente PriB, que aún retiene la función TrpF pero con baja actividad catalítica.

operon trp completo [41] y en la familia subtrpF se conserva solo a la función TrpF debido a la perdida del operon his [Juarez vazquez et al 2015 in prep]. Existen al menos 43 familias de Actinobacteria sin explorar respecto a la funcionalidad de PriA. En la Figure 0.0.9 se muestra la constantes catalíticas de PriA estimadas en diferentes organismos para estos dos sustratos.

Table: Capacidad catalítica respecto a PRA y ProfAR de miembros de la familia PriA

Fuente	Familia	HisA	HisA <i>in vivo</i>	TpF <i>in vivo</i>	$K_{cat}^{ProfAR}[M^{-1}s^{-1}]$	$K_m^{ProfAR}[\mu M]$	$K_{cat}^{PRA}[M^{-1}s^{-1}]$	$K_m^{PRA}[\mu M]$	$K_{cat}^{PRA}/K_m^{PRA}$	Referencia
<i>Escherichia coli</i>			-	-	1.6	4.9	3.1	-	-	Henn-Sax 2002
<i>Escherichia coli</i>	TpF	-	-	-	-	0	12.2	34.5	2.82	Stern et al. 1996
<i>Mycobacterium tuberculosis</i>	PriA	-	-	19	0.23	12	21	3.6	0.17	Due 2011
<i>Mycobacterium smegmatis</i>	PriA	*	*	2.6 ± 0.5	0.85 ± 0.04	0.33	7.9 ± 2.4	3.1 ± 0.43	0.39	Verduzco-Castro 2016
<i>Streptomyces globisporus</i>	PriA	*	*	4.2 ± 0.8	0.74 ± 0.03	0.18	11 ± 1.0	3.8 ± 0.2	0.34	Verduzco-Castro 2016
<i>Streptomyces coelicolor</i>	PriA	-	-	3.6 ± 0.7	1.3 ± 0.2	0.4	5.0 ± 0.08	3.4 ± 0.09	0.7	Noda-Garcia 2010
<i>Streptomyces ipomoeae</i>	PriB	*	*	3.8 ± 0.2	0.82 ± 0.02	0.21	60.8 ± 1.1	8.25 ± 0.4	0.14	Verduzco-Castro 2016
<i>Streptomyces</i> sp. Mg1	PriB	*	*	13.2 ± 3.4	0.92 ± 0.19	69	129.6 ± 34	0.29 ± 0.04	0.0022	Verduzco-Castro 2016
<i>Streptomyces</i> sp. C	PriB	*	*	11.4 ± 3.4	2.53 ± 0.74	0.22	149.9 ± 29	1.4 ± 0.12	9	Verduzco-Castro 2016
<i>Streptomyces sphaeus</i>	PriB	*	*	3.9 ± 0.89	0.69 ± 0.04	0.18	24.5 ± 4.0	1.6 ± 0.29	67	Verduzco-Castro 2016
<i>Corynebacterium diphtheriae</i>	subHisA	-	-	4.4 ± 0.5	2.6 ± 0.3	0.59	-	-	0	Noda-Garcia 2013
<i>Corynebacterium jeikeium</i>	PriA	-	-	2.3 ± 0.2	0.9 ± 0.08	0.39	5.1 ± 1.0	1.6 ± 0.16	0.31	Noda-Garcia 2013
<i>Corynebacterium striatum</i>	subHisA	-	-	6.9 ± 0.7	2.1 ± 0.5	0.3	-	-	0	Noda-Garcia 2013
<i>Corynebacterium diphtheriae</i> La18-L-T50L-T80S	subHisA	*	*	4.5 ± 1.5	0.6 ± 0.08	0.13	133 ± 10	0.05 ± 0.01	0.0004	NNoda-Garcia 2013
<i>Actinomyces wogenstalii</i> DSM 15434	PriB	*	*	2.1 ± 0.5	1.8 ± 0.2	0.9	26.3 ± 6.3	0.37 ± 0.09	14	Verduzco-Castro 2016
<i>Actinomyces odontolyticus</i> ATCC 17982	subTpF	*	*	-	-	0	-	-	0.02	Juarez-Vazquez 2017
<i>Actinomyces oris</i> K20 BABY01	PriA					0.02			0.01	Juarez-Vazquez 2017
<i>Actinomyces</i> sp. oral taxon 171 str. F0337	PriA	*	*	-	-	0.01			4	Juarez-Vazquez 2017
<i>Actinomyces</i> sp. oral taxon 848 str. F0332	subTpF	*	*	-	-	0			0.0001	Juarez-Vazquez 2017
<i>Actinomyces wogenstalii</i> DSM 15434	PriA	*	*	-	-	0.01			0.02	Juarez-Vazquez 2017
<i>Bifidobacterium adolescentis</i> L2-32	PriA	*	*	-	-	0.2			0.1	Juarez-Vazquez 2017
<i>Bifidobacterium gallicum</i> DSM 20093	PriA	*	*	-	-	0.1			0.04	Juarez-Vazquez 2017
<i>Bifidobacterium longum</i> ATCC 15697	PriA	*	*	1.7 ± 0.1	0.3 ± 0.03	0.2	40 ± 7	3.5 ± 0.04	0.3	Juarez-Vazquez 2017
Camera CAM1	Metagenoma			1.7 ± 0.2	0.1 ± 0.01	0.06	32.2 ± 1.7	1.9 ± 0.1	0.06	Noda-Garcia 2015
CAM1 AS1G	Metagenoma			4.0 ± 0.9	0.2 ± 0.03	0.04	23.5 ± 6.5	0.5 ± 0.1	0.02	Noda-Garcia 2015
CAM2	Metagenoma			n.d.	0	n.d.	-	-	0	Noda-Garcia 2015
PriA Ancestral	Ancestral			9.4±1.6	0.3±0.009	0.03	4.3±0.4	0.6±0.02	0.13	Verduzco, Noda, sin publicar
PriA SubHisA	Ancestral			3.7±1.01	0.5±0.03	0.1	-	-	0	Verduzco, Noda, sin publicar
SubHisA Ancestral	Ancestral			6.3±0.7	0.15±0.03	0.02	-	-	0	Verduzco, Noda, sin publicar
SubHisA PriA	Ancestral			27.7±3.4	0.05±0.005	2	167.82	0.03±0.002	0.0001	Verduzco*
<i>Streptomyces acidiscabies</i>				163.6	0.1					Juarez*
<i>A. visco</i>				46	1.37	36	3.4			

## 0.1. Modelos computacionales

Para desarrollar herramientas se adoptó el enfoque de los contenedores bioinformáticos, todo el código fue depositado y documentado en GitHub y distribuido a través de un contenedor Docker

# Antecedentes

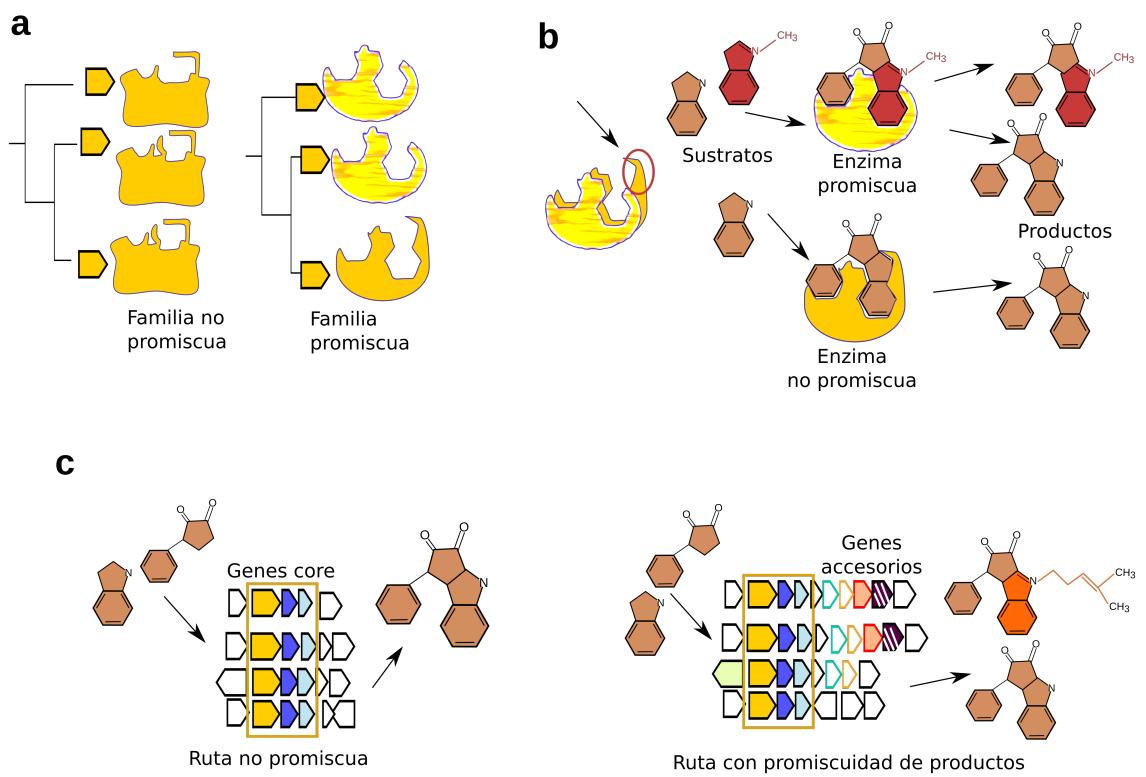


Figura 2: Antecedentes Conceptuales

## 0.2. La promiscuidad puede abordarse a distintos niveles incluyendo enzima, familia y ruta de metabólica.

Si se entiende a la promiscuidad como funciones alternativas de alguna unidad molecular, puede observarse promiscuidad a distintos niveles: desde un mismo gen que presenta splicing alternativo, una misma enzima con funciones alternativas, o bien una familia enzimática donde al menos algunos homólogos codifican enzimas promiscuas [20]. Existen también rutas que generan productos metabólicos alternativos[68], si se considera como la unidad de estudio a una ruta biosintética podemos generalizar la noción de promiscuidad al concepto de rutas promiscuas. La *Figura 1* muestra tres niveles en los que se puede estudiar la promiscuidad: i) Distinguendo familias de enzimas promiscuas de familias especialistas, ii) Distinguendo enzimas específicas de enzimas especialistas en una familia enzimática promiscua y finalmente iii) encontrando promiscuidad en rutas de metabolismo especializado.

PriA en Actinobacteria y HisA en enterobacteria son familias que ambas isomerizan proFAR en la ruta de síntesis de histidina pero solo la familia PriA es promiscua pues puede además isomerizar el sustrato PRA durante la síntesis de triptofano [5]. Sin embargo dentro de Actinobacteria, existen miembros no promiscuos de PriA, en algunas especies de Actinomyces los homólogos de *priA* codifican para enzimas monofuncionales en alguno de los dos sustratos, al menos en análisis *in vitro* [70]. Así pues dentro de una familia promiscua no todos los miembros tienen esta propiedad.

En este trabajo se buscará encontrar marcas de promiscuidad a nivel familia, enzima y ruta biosintética en el metabolismo especializado.

## 0.3. Antecedentes conceptuales

Se ha intentado identificar enzimas promiscuas mediante aprendizaje máquina utilizando únicamente la secuencia de aminoácidos. Estos enfoques no han distinguido entre identificación de familias promiscuas e identificación a nivel de enzima. [36] Hasta ahora al utilizar únicamente la información de la secuencia no ha sido posible identificar una familia promiscua sin conocer previamente al menos un miembro promiscuo de ella. Por otra parte, diferenciar la promiscuidad a nivel de enzima se dificulta cuando la identidad de secuencia es alta, como en el caso de las PriA que han perdido la promiscuidad en Actinomyces [70]

Para mejorar nuestro entendimiento del fenómeno además de la comparación de secuencias es necesario integrar otros elementos al análisis, Figura 2. Es difícil medir la promiscuidad en términos absolutos, por ejemplo, no se puede aseverar que una enzima es no promiscua sin haber previamente descartado todos los posibles sustratos del universo químico. Además incluso enzimas que resultan no promiscuas en análisis

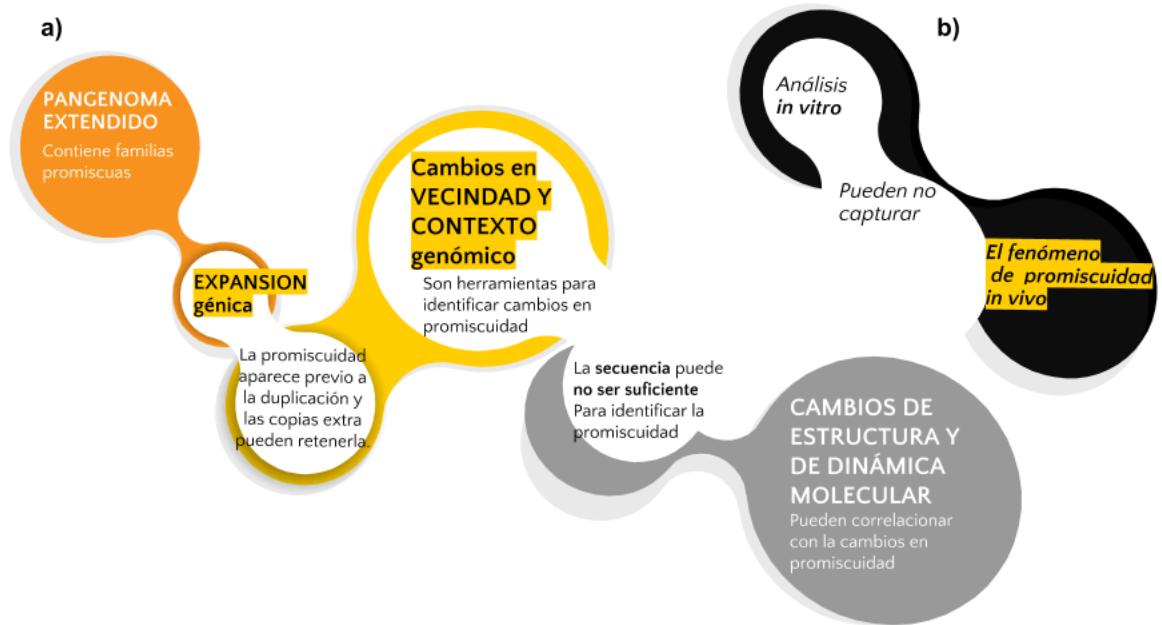


Figura 3: Antecedentes Conceptuales

in vivo si son promiscuas al examinarlas in vivo [71]. Sin embargo debido a que al adquirir una nueva función existe un umbral donde la función ancestral es conservada, es plausible estudiar transformar el problema de encontrar promiscuidad al de encontrar cambios en promiscuidad al relacionar estos últimos con las huellas que dejan los cambios funcionales[33]. Entre los elementos relevantes que correlacionan con la adquisición de una función alternativa se encuentran además de divergencia de secuencia, diversidad en vecindad genómica[72], la perdida o ganancia de genes[70], las expansiones genicas, i.e. el crecimiento del pangenoma endtiro de un grupo taxonómico [26] y finalmente cambios estructurales o de flexibilidad durante la dinámica molecular[13]. Estos elementos tienen en común que reflejan un cambio en alguna propiedad genómica o biofísica observable en el registro evolutivo, de lo que se deriva que el buscar cambios en la promiscuidad de una enzima, familia o ruta, resulta mas factible por ahora que la búsqueda intrínseca de promiscuidad.

Debido a la abundancia de datos genómicos los primeros capítulos de este trabajo se centran en encontrar variaciones en secuencia, distribución del pangenoma, y vecindades genómicas para encontrar candidatos de familias, enzimas y rutas promiscuas.

## 0.4. El establecimiento de un marco de conservación permite distinguir cambios

La función de una enzima es un concepto jerárquico, dependiente de la filogenia de un organismo [73]. Por ello, para poder encontrar marcas de cambio funcional, por ejemplo diferencias en número de copias de una familia, primero es importante trabajar en la construcción de un marco filogenético consistente que permita ordenar inclusive organismos de la misma especie. La dificultad de esta tarea consiste en que si los organismos que se desea ordenar son muy cercanos, marcadores clásicos como el 16s son también muy parecidos en secuencia y no permiten resolver las relaciones entre ellos. Este caso dificultó la construcción de un árbol filogenético de *Actinomyces* y con ello se imposibilitaba encontrar patrones en la matriz de presencia / ausencia de genes [70].

Para solventar la falta de resolución de genes individuales en organismos cercanos puede utilizarse el conjunto de todos los genes comunes en un linaje. Este conjunto es conocido como el core genome. Se han desarrollado herramientas bioinformáticas para este problema, por ejemplo phyloPhlan fija 400 genes comunes en bacteria y trata de localizarlos dado un conjunto de genomas sin importar su linaje [74]. Sin embargo el contenido de genomas procariontes suele ser muy variable debido a mecanismos como transferencia horizontal y duplicación génica [75]. Esta variabilidad puede dificultar encontrar muchos de estos 400 genes o bien puede ser que en cierto linaje no sean tan informativos. Entre organismos de la misma especie pueden suceder fenómenos como que el core siempre se reduzca al aumentar un nuevo genoma y que el conjunto total de familias genéticas (pangenoma) siempre aumente. Aunado a esta observación biológica están también las limitaciones técnicas, hay genes que no aparecen en un genoma porque este fue mal secuenciado o ensamblado y por lo tanto estos genes disminuyen el tamaño del core.

Así pues para distinguir cambios genómicos ayuda establecer primero un orden entre los genomas del linaje a analizar. Para ello, un camino es localizar los genes del core exclusivos de cada grupo de genomas y libres de parálogos. Aunque al momento existe publicada *metaphor*, una herramienta de selección de ortólogos [77], al comenzar este trabajo no existía un método disponible para ello. Finalmente una vez obtenidos los genes del core es deseable contar con un algoritmo que además los concatene y entregue un árbol filogenético.

## 0.5. La genómica comparativa como herramienta en la distinción de familias y enzimas promiscuas que participan en el metabolismo especializado.

Una parte del metabolismo especializado está compuesta por familias enzimáticas que evolucionaron de rutas de metabolismo central [78]. En las familias expandidas, ya sea por duplicación o por transferencia horizontal, las expansiones pueden retener la función química de las rutas centrales [79], así como también la función alternativa suele estar presente aún a bajos niveles antes de la divergencia o duplicación [33]. Por tanto las familias con expansiones en un linaje son candidatas a ser familias promiscuas en él. Se ha notado que en el linaje en que una familia enzimática es promiscua hay una zona de cambio en promiscuidad [39], Las expansiones de rutas centrales que participan en la síntesis de productos naturales son candidatos a presentar cambios en promiscuidad tanto a nivel familia como a nivel enzima.

por ello la observación de la retención de función ancestral al aparecer una función alternativa proporciona una zona favorable para la búsqueda de promiscuidad a nivel de enzima. En la familia existirá un gradiente de promiscuidad, las más cercanas a la zona de duplicación o divergencia tienen más posibilidades de tener un cambio en promiscuidad que las más conservadas y cercanas al metabolismo central.

Después de que en la sección anterior se estableció la posibilidad de mejorar las relaciones filogenéticas de un linaje, se abre la posibilidad de buscar en él expansiones de familias génicas de metabolismo central. Estas copias extra son candidatas a pertenecer a rutas de metabolismo especializado. Como prueba de concepto esta idea de minar genomas incorporando información evolutiva permitió la identificación de la biosíntesis de arsenolipidos [52]. La búsqueda de productos naturales cuenta entre sus premisas que estos se producen en vecindades genómicas llamadas clusters y que además clusters cercanos (ya sea en contenido genético o en la secuencia de sus componentes), exploran variaciones metabólicas, es decir sus enzimas catalizan reacciones sobre sustratos parecidos aunque no idénticos [52].

La primera versión de evomining cuenta con 200 genomas de Actinobacteria, una base de datos de secuencias de enzimas de productos naturales y otra base de datos de secuencias de enzimas de rutas centrales curada a mano.

1 ARTS [80] 2 EvoMining [52] Se ha avanzado en archaea [81]

Desarrollarla en combinación con algoritmos de búsqueda de cambios en la vecindad genómica la harán una plataforma ideal para abordar el problema de las familias, proporcionando una solución a la dificultad de no tener conocimiento previo de un miembro promiscuo en la familia investigada.

Respecto al problema de los miembros, se propone explorar variaciones en vecindad genómica, flujo genético y dinámica molecular, como candidatos a reflejar la variación

en promiscuidad. Finalmente,

tomando como modelo biológico el phylum Actinobacteria, un grupo de bacterias reconocido por su diversidad metabólica donde se ha probado la existencia de promiscuidad enzimática.

Evomining es una plataforma bioinformática pensada para la identificación de productos naturales

Si se combinara evomining con la premisa de que vecindades distintas son marcadoras de funciones químicas distintas, al encontrar una familia expandida con vecindades genómicas diferentes se podría solventar la deficiencia de otros métodos bioinformáticos consistente en que para identificar familias promiscuas se debe conocer previamente un miembro promiscuo de la misma. (Fig 4) Así pues al combinar evomining con herramientas de vecindad genómica tanto de comparación como de visualización estaremos mejorando su funcionalidad en la identificación de familias promiscuas. En la siguiente sección BLA BKA

la variación es la materia prima de la evolución.

## 0.6. La genómica comparativa como herramienta en la priorización de clusters promiscuos

La promiscuidad nos interesa por su producción de variantes. Si bien en rutas centrales rescata la función en metabolismo secundario crea nuevas variantes moleculares que permiten adaptación, de hecho pangénomas grandes correlacionan con aparición de nuevas funciones enzimáticas. Considero que el concepto de promiscuidad puede ser extendido a un nuevo nivel. Promiscuidad enzimática, promiscuidad de familia, promiscuidad de cluster. En rutas centrales robustez, pero en metabolismo secundario platicidad. Pangéoma abierto cerrado, aquí proponemos organizar los clusters y finalmente la medida de su apertura. Variantes de clusters producen nuevos compuestos, ya sea por promiscuidad enzimática en un core conservado o por variación en la presencia / ausencia de genes. Un cluster recibe los mismos sustratos y los transforma en diferentes productos.

### 0.6.1. Expansión y contextos genómicos como herramienta de anotación funcional

Al evaluar la herramienta de análisis de promiscuidad PROMISE [36] en un set de datos de la familia HisA/PriA [39] obtuve que en su mejor desempeño es (huella molecular de tamaño 6) clasifica correctamente casi todas las no promiscuas, (HisA) pero no sucede lo mismo con la familia PriA donde tiene éxito en 16 de 45 casos. Al aplicar el mismo tamaño de huella a 9 miembros promiscuos de la familia IlvC no consigue predecir correctamente ninguno de ellos reflejando tal vez que en su conjunto

de entrenamiento no había miembros promiscuos ilvC. Por lo menos para estas familias el conjunto de entrenamiento o los descriptores no son suficientes para la anotacion de promiscuidad.

La diversidad enzimática existente es el resultado de un proceso de expansion, mutacion y seleccion que se ha desarrollado durante el transcurso de la historia evolutiva [1]. Existe evidencia de que cierto grado de promiscuidad o divergencia funcional precede a la duplicacion genica [62]. Por este motivo detectar expansions ya sea duplicaciones o transferencias horizontales [83], puede ser un buen punto de partida para determinar divergencia funcional y promiscuidad. No todas las expansions denotan cambio de funcion enzimatica, algunas pueden ser meros accidentes, sin embargo dado que la funcion de una enzima suele estar relacionada con sus vecinos [84], una expansion en una vecindad genomica diferente de la tradicional sera un referente de adquisicion de una nueva funcion y entonces un indicador de existencia previa de promiscuidad.

Para sistematizar el estudio de contextos y vecindades genomicas se desarrollo Search Tool for the Retrieval of Interacting Genes/Proteins STRING [85], que cuenta con una anotacion de ortologia jerarquica y consistente, realizada en 2000 organismos en cuyo marco interacciones de proteinas con implicaciones funcionales son predichas tanto de novo por informacion genomica de co-ocurrencia como por mineria de datos en articulos publicados. STRING es una base de datos, y como tal no permite agregar nuevos genomas para su analisis. Sus 2000 organismos incluyen especies tanto bacterianas como eucariotas. Al existir tanta diversidad, los genomas disponibles para un genero o clase especificos son escasos, p. g. de los mas de 300 genomas disponibles de *Streptomyces* solo 24 estan incluidos.

Para resolver la baja cobertura de STRING hacia ciertos grupos taxonomicos se pueden desarrollar scripts de vecindad genomica utilizando RAST (Rapid Annotation using Subsystem Technology); un servicio interactivo de anotacion automatica de genomas de bacterias y arqueas [86] donde la funcion de cada gen se asigna de acuerdo a conocimiento previo de subsecuencias de organismos cercanos filogeneticamente, cuando es posible se incluye en un subsistema metabolico. Estamos en una era de explosion de datos genomicos, proximamente se espera contar con millones de genomas bacterianos incluso provenientes de bacterias no cultivables, por ello los algoritmos deben ser constantemente optimizados a los nuevos volumenes de datos [44]. Ante esta expectativa seria muy util desarrollar algoritmos de analisis genomico que sean de codigo libre o al menos interactivos para que cada laboratorio pueda personalizarlos para sus propios genomas.

Finalmente, no solo la vecindad genomica inmediata puede ser utilizada como distin-tivo en la busqueda de promiscuidad, diferencias en el contexto genomico en genes relacionados con una enzima promiscua, sin importar su ubicacion dentro del genoma tambien pueden ser relevantes para la perdida o ganancia de funcion quimica [41], [70].

### 0.6.2. Contexto y vecindades genomicas

En 2012 fueron analizados 102 genomas de 29 familias de Actinobacteria [71]. sugiriendo que al menos en *Corynebacteria* el contexto y la vecindad genomica incidian en la sub-funcionalizacion de PriA en subHisA [41]. Respecto a IlvC, otra familia involucrada en la sintesis de aminoacidos fue estudiada y caracterizada bioquimicamente en 1 *Corynebacterium* y 8 *Streptomyces* [42]. Para ampliar estos resultados, utilizando la anotacion de RAST y una generalizacion de la definicion de vecindad de STRING, se diseño un algoritmo para identificar vecindades similares asi como uno de visualizacion de contexto, ambos disponibles como software libre en github nselem/perlas .

El algoritmo de clasificacion de vecindades permite agruparlas en clusters y calificar estos clusters segun su conservacion dado un grupo de bacterias. La definicion de vecindad y similitud de vecindad esta descrita posteriormente en los metodos. El algoritmo fue aplicado a la familia IlvC en 290 *Streptomyces* resultando 9 clusters Datos entre los mas poblados el primero cuenta con 279 elementos, otro con 9 elemento y dos mas con 7 miembros (Fig 3), resultados experimentals son congruentes con que existe divergencia funcional entre miembros de clusters distintos [42]

Natural products genomic era[88]

## 0.7. Estudio de una familia PriA

### 0.7.1. Caracterizacion in vivo

No todas las familias promiscuas provienen de expansiones, tal es el caso de PriA en Actinobacteria, donde no tiene expansiones y hasta el momento no se le conoce participacion en rutas de metabolismo especializado.

cuya promiscuidad es debida a la perdida de TrpF Algunas enzimas PriA no han mostrado promiscuidad in vitro pero si in vivo ya que sobreviven en un medio sin triptofano, es decir in vivo complementan la funcion trpF. Para la construccion de cepas de *Streptomyces* con variantes no nativas de priA minimizando la modificacion genomica y el efecto de sobreexpresion, se planea utilizar *E. coli* como intermediario para realizar seleccion por auxotrofia. Se cuenta con un conjunto de plasmidos para transformar a *E. coli* asi como con las mutantes sencillas de *E. coli* para trpF y hisA que permiten realizar seleccion por auxotrofias. Ademas tenemos una colección de cepas nativas de *Streptomyces* asi como un mutante de PriA de *S. coelicolor*. Se optimizo una reaccion de PCR para la amplificacion de un segmento de DNA de *S. coelicolor* que contiene a priA.

### 0.7.2. Caracterizacion bioquimica in vitro.

De la familia PriA y sus subfamilias se han caracterizado bioquimicamente miembros selectos de Actinomycetaceae, Bifidobacteriaceae, Micrococcaceae, Acidimicrobiaceae, *Corynebacterium*, Mycobacteriaceae, Streptomycetaceae, Camera (provenientes de metagenoma), reconstrucciones ancestrales, 80 mutantes de *Corynebacterium*, y 2 mutantes de Camera mediante cineticas enzimaticas para calcular las constantes  $K_{cat}, K_m$ . El genero *Streptomyces*, el que cuenta con mayor cantidad de genomas disponibles representa una oportunidad muy poco explotada de explorar la influencia del contexto y la vecindad genomicas en secuencias de PriA (Tabla 3, Figura 5).

### 0.7.3. Modelado de dinamica molecular

La dinamica es un metodo que permite hacer simulaciones de particulas que sirve para obtener informacion de propiedades macroscopicas de un conjunto de atomos [89]. Es util en el marco de mi proyecto porque permite la exploracion del espacio conformacional, y se ha visto que este esta relacionado con la actividad de la enzima [91], ademas dado un conformero permite verificar su estabilidad. Resuelve la ecuacion de movimiento de Newton con base a una configuracion inicial, las fuerzas interatomicas como los enlaces covalentes, las fuerzas de Van der Waals y la carga de las particulas[45]. Entonces para generar una simulacion de dinamica molecular, debe contarse con una estructura como punto de partida, ya sea esta cristalografica o modelada de novo o por homologia. El laboratorio de bioinformatica y biofisica computacional ha desarrollado un protocolo de generacion de modelos homologos estructurales y dinamicas moleculares (Carrillo-Tripp et al 2015 in prep); con este pipeline se han generado dos estructuras de Camera [39], 30 estructuras y dinamicas de miembros de Actinobacteriaceae y Bifidobacteriaceae (Vazquez-Juarez et al in prep.) y finalmente una estructura de subHisA de *Corynebacterium diphteriae*. En la familia *Streptomyces*, interesante debido a su variacion en contexto genomico y en mediciones in vitro aun no se modelan dinamicas moleculares aunque 40 estructuras por homologia estan en proceso.

En un estudio de subHisA [71] se utilizo el metodo de dinamica molecular y se comparó el numero de confórmeros entre miembros de subHisA y PriA, resultando mayor el de PriA como corresponde a una enzima promiscua. El estudio sobre la relación dinámica-flexibilidad de  $\beta$ -lactamasas utiliza replica exchange, una variacion de dinamica molecular que corre replicas en paralelo a distintas temperaturas [92]. Una desventaja de este metodo es que por el costo computacional de las replicas agregar explicitamente otras moleculas a la simulacion como el solvente no es posible en tiempo razonable. Una vez generadas las dinamicas moleculares se procedera a calcular tanto el numero de conformeros como el indice de flexibilidad dsi [13]. Se esta desarrollando PEDB, promiscuous enzyme database, una base datos genomicos, evolutivos, bioquimicos y estructurales y de metabolismo de PriA en Actinobacteria donde se procedera al analisis de los mismos (<http://148.247.230.43/nselem/PHP/queries.html>).

En conclusion la promiscuidad enzimatica es un fenomeno complejo debido a multiples causas. Existe una gran variedad de estudios con enfoques puntuales sobre aspectos estructurales, dinamicos y evolutivos sin embargo hasta ahora no se han reportado trabajos multidisciplinarios que involucren a todas las partes involucradas

# Pregunta biológica



# **Objetivos**

## **0.8. Objetivo General**

Estudiar el fenomeno de promiscuidad enzimatica tanto desarrollando estrategias para identificar familias promiscuas dentro de un grupo taxonomico, como comparando variaciones de promiscuidad in vitro e in vivo con variaciones en contexto genomico y flexibilidad en miembros de una familia. (Figura 7)

## **0.9. Objetivos particulares**

Mejorar evomining como metodo de identificacion de familias enzimaticas promiscuas aprovechando los cambios en vecindades genomicas como caracteristicas informativas provenientes de datos filogenomicos. Estudiar la relacion entre historias filogenomicas y procesos biofisicos con la promiscuidad in vitro, a traves de mediciones de ciertas caracteristicas de la familia PriA. Caracterizar cambios de promiscuidad enzimatica in vivo mediante perfiles metabolomicos de actividades de PriA y enzimas asociadas.



# Estrategias

**Obtener informacion genomica del phylum Actinobacteria.**

Colectar genomas de Actinobacteria de NCBI y de colecciones privadas.

## **0.9.1. Anotar consistentemente las secuencias codificantes de estos genomas.**

Utilizar un anotador automatizado y desarrollar los scripts necesarios para anotar los genomas.

**Establecer las relaciones filogeneticas de los genomas colectados.**

Mediante el uso del core genome construir un arbol filogenomico que permita establecer un marco sobre el cual hablar de cambio y que facilite reclasificar los genomas mal nombrados.

## **0.9.2. La promiscuidad en familias enzimaticas.**

Mejorar Evomining mediante la identificacion de cambios de vecindad genomica en familias selectas de metabolismo central convirtiendola en una plataforma de codigo libre disponible para otros investigadores.

**Identificar cambios en la vecindad genomica en familias selectas de enzimas de metabolismo central.** Clasificar sistematicamente las secuencias de familias codificantes segun su similitud en familias enzimaticas.

Desarrollar las herramientas bioinformaticas necesarias para separar clusters de vecindades genomicas.

**Promiscuidad in vitro dentro de miembros de una familia promiscua de enzimas.**

Dados los sustratos conocidos de PriA investigar las posibles correlaciones entre mediciones de constantes catalíticas, contexto genómico, vecindad genómica, número de conformeros e índice de flexibilidad.

**Sistematizar Evomining para convertirla una plataforma descargable y utilizable en cualquier set de datos bacterianos relacionados taxonomicamente proporcionados por el usuario.**

Ampliar el contenido de Evomining al integrar los genomas colectados de Actinobacteria. Sistematizar la base de datos de metabolismo central.

Desarrollar la visualización e integrar la clasificación de vecindades genómicas como una herramienta adicional en la búsqueda de promiscuidad.

**Seleccionar miembros homólogos de la familia de enzimas.**

Se escogieron 41 Streptomyces repartidos en un árbol de rpoB de 400 Streptomyces con genoma disponible. Esta selección incluye los seis Streptomyces de los que se cuenta con cinética enzimática de PriA, tres de ellos con estructura cristalográfica.

**Medir cinéticas enzimáticas, contexto genómico, vecindad genómica, flexibilidad y número de conformeros.**

Determinar la pertenencia a uno de cuatro posibles contextos genómicos respecto al gen trpF. Estudiar la existencia de distintas vecindades genómicas. Determinar la cinética enzimática de 9 enzimas más buscando variabilidad en contexto genómico (sugeridas en la tabla 4). Obtener mediante una colaboración 37 modelos estructurales por homología y modelar dinámica molecular.

La siguiente tabla contiene la diversidad de contextos y vecindades genómicas de 41 Streptomyces respecto al gen trpF.

**Determinar posibles correlaciones entre los datos producidos.**

Número de conformeros e índice de promiscuidad.

índice de flexibilidad y número de conformeros.

Número de conformeros y contexto genómico.

índice de flexibilidad y contexto genómico.

Contexto genómico e índice de promiscuidad I.

Analizar las vecindades genómicas e índice de promiscuidad I.

**0.9.3. Desarrollar una metodologia para la deteccion in vivo de promiscuidad enzimatica.**

Debido a cambios en flexibilidad o cambios de contexto genomico, se puede sospechar de diferencias en la funcion quimica de dos miembros de una familia de enzimas, sin conocer las diferencias a nivel de sustratos. Para investigar estos cambio in vivo se propone estudiar diferencias en perfiles metabolomicos de una colección de cepas en condiciones diversas.



# Metodologia

A continuacion describiré la metodologia para cada una de las estrategias expuestas previamente. Todos los scripts desarrollados fueron escritos en perl y estan disponibles en github <https://github.com/nselem/perlas>.

## 0.9.4. La promiscuidad en familias enzimaticas.

### Actinobacteria genomica

Para obtener informacion genomica del phylum Actinobacteria mediante la colección de genomas de NCBI se revisaron todas las familias de Actinobacteria de la base genoma de NCBI y se seleccionaron los genomas con minimo 5 genes por contig. Se crearon scripts para utilizar la interfaz e-utils de NCBI y descargar estos genomas desde la terminal a partir de una lista de identificadores.

### Annotation

Para anotar consistentemente las secuencias codificantes de estos genomas se utilizo el anotador automatizado RAST y se desarrollaron los scripts necesarios para anotar los genomas desde la terminal, conectado asi NCBI y RAST.

### Genomic DB phylogeny

Establecer las relaciones filogeneticas de los genomas colectados. Mediante el uso del core genome para construir un arbol filogenomico, para reclasificar los genomas mal nombrados.

Para obtener el core genome y en base a el reclasificar los genomas se diseño el algoritmo estrellas basado en Best Bidirectional Hits (blast all vs all).

Estrellas. Se realiza un blast all vs all de genomas deseados. Para cada secuencia, centrado en cada genoma se realiza una lista (estrella) de sus mejores hits bidireccionales. Si las listas de todos los genomas coinciden es un BBH multiple y se agrega la lista al core genome. (Fig 9) Una vez con el core genome completo se puede reconstruir

la filogenia. Este metodo fue exitoso en la detección de una familia marcadora de *Clavibacter michiganensis* (2014 Rodriguez-Orduña in prep).

### **0.9.5. Identificar cambios en la vecindad genomica en familias selectas de enzimas de metabolismo central.**

Clasificar sistematicamente las secuencias de familias codificantes segun su similitud en familias enzimáticas.

Como se menciono en los antecedentes, se han separado 888 genomas de Actinobacteria en 3 grupos taxonomicos utilizando para la anotacion la tecnologia de subsistemas de RAST. Para la separacion en familias iso funcionales (ortologos, paralogos y expansiones) se utilizo RAST, especificamente el script What Changed (WC) que asigna un numero a cada familia, esta herramienta esta basada en k-mers, su codigo esta disponible en github: ([https://github.com/kbase/kbseed/blob/master/service-scripts/srv\\_CS.pl](https://github.com/kbase/kbseed/blob/master/service-scripts/srv_CS.pl)). Ademas de en los tres grupos ya mencionados, tambien se realizara una clasificacion para trescientos genomas de Actinobacteria distribuidos en todas sus familias taxonomicas.

Para desarrollar las herramientas bioinformaticas necesarias para separar clusters de vecindades genomicas a continuacion se describe detalladamente como se definio vecindad genomica y la relacion implementada de similitud.

1. Un conjunto expandido es un conjunto que contiene secuencias homologas asi como sus expansiones: paralogos y transferencias horizontales. Dado un conjunto de genomas, se pueden calcular y enumerar todos sus contextos extendidos utilizando WC.
2. Un PEG es un elemento de un conjunto expandido. Dado un PEG p, se define  $CE(p)$  el numero del conjunto expandido de p, como el numero asignado por WC al conjunto expandido a que p pertenece.
3. La vecindad de un PEG es el conjunto de PEGs cercanos a el. Dado un umbral en terminos de distancia de pares de bases entre puntos medios para precisar la definicion de cercano, se pueden calcular todos los contextos de un genoma.
4. Una vecindad A es n-similar a otra vecindad B si  $C = \{aA \mid bB, CE(b) = CE(a)\}$  tiene al menos cardinalidad n. Es decir si existen al menos n elementos de A que pertenecen al mismo conjunto expandido que algun elemento de B.
5. Un conjunto de vecindades es un conjunto de PEGs clusterizado segun la relacion n-similaridad. Si A es n-similar a B y B es n-similar a C entonces, aun si A no fuese n-similar a C, los PEGs generadores de A,B,C son agrupados dentro del mismoconjunto de vecindades.
6. Un cluster es un conjunto de conjunto de vecindades.

7. Los clusters son evaluados segun el numero y la cardinalidad de sus conjuntos de vecindades.

Sea  $Cl$  un cluster, donde  $CC_i$  es un conjunto de contextos y  $n_i$  es la cardinalidad de  $CC_i$   $Cl=\{CC_1, CC_2, \dots, CC_k\}$

Sean  $M$  la cardinalidad maxima de un conjunto de contextos y  $m$  la cardinalidad maxima sin considerar  $M$ .

$$\#\$ M \neq \max_{i=1,2,\dots,k} n_i \quad m \neq \max_{i=1,2,\dots,k} n_i \quad n_i \neq M$$

$$\sum_{j=1}^n (\delta\theta_j)^2 \leq \frac{\beta_i^2}{\delta_i^2 + \rho_i^2} \left[ 2\rho_i^2 + \frac{\delta_i^2 \beta_i^2}{\delta_i^2 + \rho_i^2} \right] \equiv \omega_i^2$$

$M$  representa el contexto mas difundido de la enzima, dentro del grupo taxonomico considerado; mas relevante porque si  $m$  es grande significa que hay un segundo contexto genomico conservado en dicho grupo taxonomico, y entonces posiblemente una ganancia de funcion.

La evaluacion de  $Cl$  esta dada por una combinacion lineal de  $k, m$  y  $M$   
 $S(Cl) = f(k, m, M) = c_1k + c_2m + c_3M$

Este algoritmo se puede mejorar considerando la orientacion de los genes del cluster asi como clusters de los vecinos.

### Organizar y presentar los datos en una plataforma.

Para contribuir al desarrollo de la plataforma Evomining se desarrollaran scripts de visualizacion de arboles filogeneticos y contextos genomicos.

Para facilitar el analisis visual de una vecindad genomica ya la vez generar imagenes de alta calidad facilmente exportables para su uso en publicaciones, se desarrollaran scripts de visualizacion que utilizaran el formato Scalable Vector Graphics (SVG), dicho formato es basicamente un archivo de texto XML que contiene instrucciones para que el navegador realice un dibujo (W3school/SVG 2015). Al ser vectores, las imagenes generadas en SVG no pierden resolucion al ser escaladas y justamente por ser escalables permiten explorar con detalle grandes cantidades de datos organizados por ejemplo en arboles filogeneticos. Los scripts a desarrollar extraeran para cada gen informacion necesaria como coordenadas, direccion, funcion quimica, etc, proveniente de la anotacion de RAST y de los scripts de comparacion de vecindades genomicas. La primera version de evomining fue desarrollada en el lenguaje perl; este lenguaje cuenta con un modulo para facilitar la elaboracion de SVG (perlmaven/SVG 2015) por lo que al utilizar SVG no se agregan nuevos requerimientos a su desarrollo y se facilita su portabilidad.

Se amplificara Evomining de los 200 genomas con que contaba su version inicial a los 880 colectados mudando la curacion manual de su base de datos de rutas centrales

a la anotacion por subsistemas de RAST. Finalmente se presentara la variacion en vecindades genomicos como una herramienta adicional que ayude en la busqueda de promiscuidad en familias de enzimas pertenecientes al metabolismo central.

### 0.9.6. Promiscuidad in vitro

#### Datos cineticos:

En todos los ensayos enzimaticos se busca medir una señal que permita una distincion clara entre sustrato y producto [93]. La cinetica enzimatica de PriA proveniente del genero Streptomyces sera determinada como ya se ha reportado previamente, mediante el monitoreo de cambios en fluorescencia (isomerizacion del sustrato PRA) o en absorbancia (isomerizacion del sustrato PROFAR). En el caso de la isomerizacion de PRA, debido a que contiene un anillo de antranilato, la fluorescencia del sustrato PRA es 50 veces mayor que la del producto 1-(2-carboxyphenylamino)-1-deoxy-D-ribulose 5-phosphate (CdRP) por lo que se utiliza la disminucion en fluorescencia como medida de la conversion del sustrato en producto [94]. Se mandaran sintetizar estas variantes para posteriormente sobre expresarlas en *E. coli*. Se creceran cepas modificadas de *E. coli* (W-, H-) en medio minimo M9 enriquecido con una mezcla de aminoacidos excepto L-histidina y L-triptofano y se seleccionaran por rescate de auxotrofia. Para obtener la enzima necesaria para los ensayos enzimaticos se utilizaran plasmidos disponibles para construcciones de sobreexpresion de proteina, despues de la produccion la enzima se purificara utilizando cromatografia por afinidad a niquel [67].

Finalmente se recopilaran datos cineticos de PriA tanto privados como los publicos reportados a la fecha en la BRaunschweig ENzyme DAtabase BRENDa [95]. Una vez colectados los datos se anotaran en PEDB, nuestra base de datos ad hoc, y se tomara como medida de promiscuidad el I-index [12] que se define como:  $I = \frac{1}{N} \sum_{i=1}^N \frac{1}{K_i} \text{cat}_i / \text{Kim}$

#### Dinamica molecular

Para generar dinamicas moleculares primer lugar se recolectaran las estructuras tridimensionales de miembros de PriA de Actinobacteria. Despues se procedera a modelar por homologia las estructuras tridimensionales faltantes utilizando el pipeline del laboratorio de bioinformatica y biofisica computacional. Este pipeline utiliza el software Rosetta para el modelado para las estructuras y GROMACS Groningen Machine for Chemical Simulation, [96] para el modelado de la dinamica molecular. Esta parte del trabajo se realizara en colaboracion con el laboratorio de bioinformatica y biofisica computacional.

### 0.9.7. Promiscuidad in vivo

Se realizaran construcciones con variantes no nativas de priA y/o trpF en Streptomyces coelicolor. Para las construcciones se amplificara mediante PCR un fragmento alrededor de PriA que se insertara en un vector. Este vector recombinara en E. coli con un casete provisto de un gen marcador de resistencia a antibiotico y este gen recombinado se pasara por conjugacion a S. coelicolor donde se espera que realice una doble recombinacion. El paso por E. coli es llevado a cabo porque Streptomyces no se puede transformar por electroporacion. Se seleccionaran las cepas de Streptomyces resistentes al antibiotico como prueba de que ya no poseen su priA nativa. Posteriormente, mediante un procedimiento analogo se sustituira el gen marcador, por variantes no nativas de priA/trpF.

La cromatografia se refiere a un conjunto de metodos que separan y analizan mezclas de moleculas. Basicamente estos metodos se basan en diferencias en el tamaño, intercambio de iones y afinidad. [45] Posteriormente se combinan con espectrometria de masas que es una tecnica que mide el radio masa-carga de las particulas fragmentadas en iones. [45]. Los datos obtenidos de espectrometria de masas se procesaran utilizando redes moleculares, que consiste en agrupar los productos segun la similitud de sus partes. Plan: 3 replicas tecnicas, 2 replicas biologicas de 5 cepas.

### 0.9.8. Consideraciones

Falsos negativos respecto a promiscuidad estan muy extendidos en la literatura y en las bases de datos, en parte porque la mayoria de las funciones son asignadas por similitud de secuencia y dado un falso negativo el error se propaga en secuencias similares. Por otro lado es muy dificil demostrar un verdadero negativo a menos que se prueben todas las posibilidades de sustrato para la enzima. Sin embargo el espacio de sustratos puede acotarse gracias a tecnicas como el docking que esta intimamente relacionado con la dinamica molecular [45]. Limitar el espacio de sustratos puede retroalimentarse con el estudio de la promiscuidad in vivo y viceversa.

Con los metodos propuestos en este trabajo solo se podra detectar perdida o ganancia de promiscuidad entre enzimas de organismos respecto a otros miembros dentro un grupo taxonomico, no asi el estado de promiscuidad intrinseco a la enzima. Si dada una enzima no se detectan variaciones en contexto, vecindad genomica o flexibilidad dentro de un grupo taxonomico cercano, entonces no podemos decir en principio nada acerca de la promiscuidad de la variante, posiblemente es promiscua pero al mantenerse constante en todos los parametros descritos, con estos metodos no se puede sugerir promiscuidad. Es posible que al mirar en un grupo taxonomico mas amplio se detecte una neofuncionalizacion de la familia aunque tambien es posible que exista una variable z como la flexibilidad de sustrato [20] que no se este considerando y que explique o sea el mejor indicador para esta familia de promiscuidad enzimatica.

Se debe considerar que si existe una correlacion vecindad genomica-promiscuidad, esta

no indica causa efecto, mas bien, es plausible que la vecindad sea un amplificacion de diferencias en secuencia, a un numero igual de variaciones en secuencia la existencia de un cambio de vecindad indica un proceso mas largo y mas cambios, es una amplificacion de las marcas dejadas por transformaciones funcionales.

Si bien no se resuelve el problema de anotar promiscuidad automaticamente, este trabajo pretende aprovechar que los contextos genomicos ayudan a la identificacion de familias promiscuas para mejorar una plataforma de productos naturales, pretende tambien una confirmacion de que los cambios en la dinamica molecular ayudan a identificar los miembros mas promiscuos hacia actividades recien adquiridas, asi como tambien ser pionero en la investigacion de promiscuidad in vivo.

Gene cluster plants[98]  
Archaeal core [99]  
Methanosarcina reconstruction [100]  
Archaea phylum[101]  
Prediction for possible products of promiscuous enzymes[102]  
Saxitoxin [103]  
Plants clusters [104]  
MiBIG [105]  
Metagenomics on Streptomyces [106]  
Sulfolobus reconstrucion [107]  
Archaeal Natural products[108]  
Computational Pangenomics [109]  
Cuántos genes “obtenidos por EvoMining” son core/ cloud/stand alone  
Qué porcentaje de genes únicos recupera EvoMining  
Eucarya paralogs reshape gene clusters [110]  
Microbial dark mater [111]  
Archaea anaerobica carbon [112] Archaea Eucarya gap loki[113]  
Archaea and eucarya[114]  
BPGA [115] genes esenciales bacteria minima[116]  
Radical [117]  
RaxML large phylogenies [118]  
R phylogenies [119]  
Streptomyces exploradores [120]  
LUCA [121] Luchando por el reconocimiento de Archaea[122],[123] The primary kingdoms [124]  
Prediccion aRchaeas [125]  
RASt archaea [126] Book Archaea [127]  
Computational methods for bacterial and archaeal genomes [128]  
Archaeas boook [129]  
GC content plasmido genoma [130] Genoma minimo[131]  
Phylogeny R [132]  
Cyanobacteria fluctuacion genomica y adaptacion [133]  
Ecology of cyanobacterua [134]  
Histidine biosynthesis[135]

- PriA reconstruction [136]
- Escala temporal bacterias [137]
- Pangenome size [138]
- variabilidad del 16s [139]



# Capítulo 1

## Desarrollo de Orthocore e implementación de otras herramientas computacionales para entender el pangenoma de un linaje genómico.

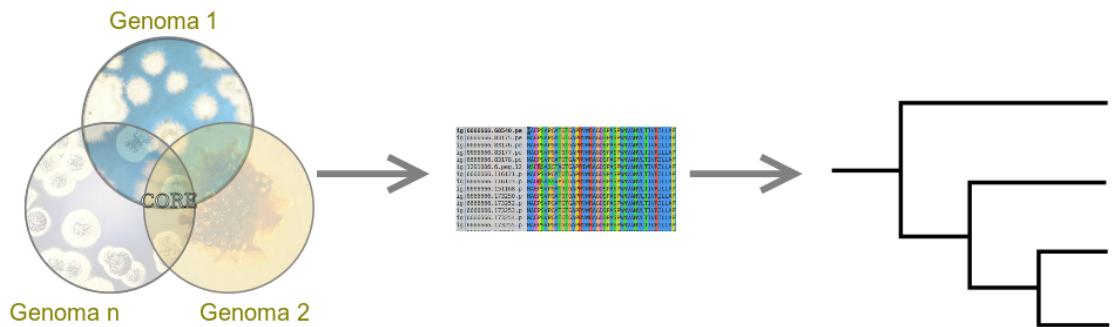


Figura 1.1: Orthocore calcula las familias génicas comunes de un linaje taxonómico. Después de un proceso de filtrado, alineamiento y curación, Orthocore concatena las secuencias de estas familias y entrega una reconstrucción filogenética.

La organización filogenética de un linaje genómico permite la observación de dinámicas de pérdida y ganancia de familias génicas en organismos cercanos. Si los organismos están desordenados taxonómicamente es difícil apreciar la dinámica de aparición-desaparición por genoma de ortólogos que son miembros de una familia génica. Así pues, ordenar los genomas de un linaje de acuerdo a su historia evolutiva facilita apreciar cambios en el número de copias de una familia. Esta consideración es relevante en el marco de esta tesis ya que cambios en la ocurrencia de promiscuidad

pueden estar relacionados a copias extra de organismos cercanos[42]. Como se explicó previamente en la introducción el pangenoma de un linaje taxonómico está dividido en el *core*, *shell* y *dispensable genome*. Orthocore es un algoritmo que se desarrolló para automatizar la obtención del *core conservado* de un linaje genómico. Con este *core* realiza una reconstrucción filogenética de los organismos del linaje Figure 1.1. En esta sección se explica el funcionamiento de Orthocore así como casos en donde fue utilizado, específicamente en casos de estudio de los órdenes *Actinomycetales*[70], *Micrococcales*[140] y *Nostocales*[141] y en el género *Salmonella* [142]. Finalmente se describen otros métodos implementados para entender el pangenoma de un linaje genómico que han sido utilizados en aplicaciones relacionadas al patógeno del tomate *Clavibacter michiganensis*[140].

## 1.1. La distribución de la función metabólica de las familias del pangenoma depende de la variabilidad del linaje seleccionado.

El número de familias génicas presentes en el pangenoma, así como su distribución en el *core*, *shell* y *dispensable genome* depende de la elección de los genomas y del linaje genómico. Para entender esto se puede pensar en un ejemplo extremo, consideremos una bacteria de 1000 familias de genes de la cual se obtienen secuencias de diez genomas de la misma cepa. Estas secuenciaciones deberían ser prácticamente idénticas y en ese caso el *core genome* sería 1000 familias, el *shell* y el *dispensable genome* serían cero. En este caso, todo el metabolismo, tanto el central como el especializado estarían conservados dentro del *core genome*, ya que el *shell* y el *dispensable genome* se encuentran vacíos. Sin embargo, si variamos el linaje taxonómico, y ahora estudiamos el pangenoma de 10 especies distintas del género *Streptomyces* ahora el core genome estará compuesto por aproximadamente un tercio de su tamaño promedio, y dentro del *core genome* es donde se encontrarán muchos de las familias dedicadas al metabolismo central o conservado (por ejemplo, familias de la glicólisis o síntesis de aminoácidos). En cambio muchas de las familias dedicadas al metabolismo especializado y pertenecientes a clústers biosintéticos de productos naturales (BGCs) estarán en el *dispensable genome* pues *Streptomyces* es productor de una gran variedad de metabolitos especializados y cada especie suele tener su producto característico. Figure 1.2.

El *core genome* de un linaje, además de tener familias conservadas y prácticamente presentes en todo el dominio Bacteria también puede contener familias marcadoras Figure 1.3 . Estos genes marcadores permiten realizar pruebas de diagnóstico para colonizaciones bacterianas. A las familias que están presentes en el *core genome* de un linaje A, pero que están completamente ausentes de un linaje B se les llama marcadoras. Por ejemplo genes conservados en la especie *Streptomyces coelicolor* pero no conservados en *Streptomyces rimosus* son genes marcadores de *Streptomyces coelicolor* respecto de *Streptomyces rimosus*. Estos mismos marcadores tal vez no sean

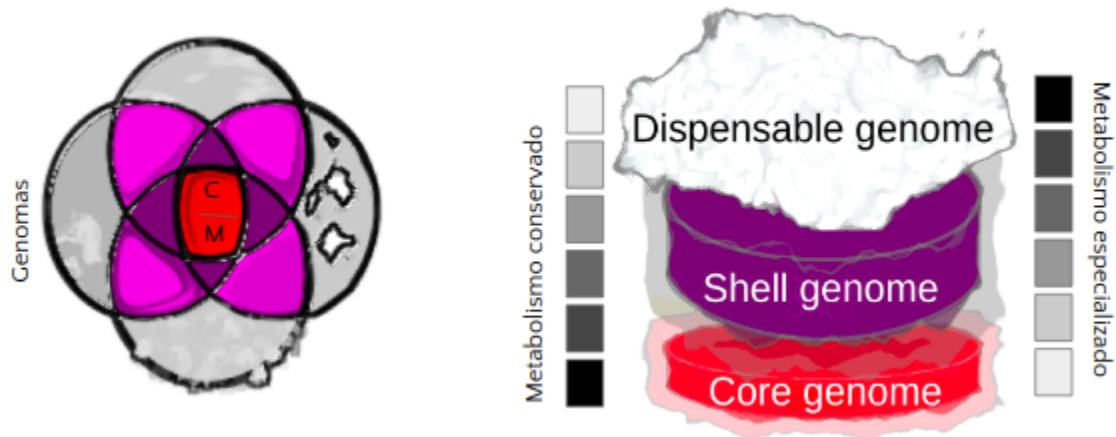


Figura 1.2: El pangenoma de un conjunto de genomas de un linaje puede ser clasificado en varios grupos. En este ejemplo, en el lado izquierdo de la figura se observa en gris el *genoma dispensable* compuesto por familias génicas presentes sólo en un genoma. En dos tonos de morado observamos el *shell genome*, familias que están presentes en la mayoría de los genomas del linaje, en este caso dos o tres genomas. Finalmente en rojo se muestra el *core*, aquellas familias presentes en todos los genomas del linaje. El *core* contiene tanto familias muy conservadas con una sola copia por genoma ( C ), como familias expandidas. Las familias marcadoras ( M ) pueden ser parte del *core* conservado o de las familias expandidas. Del lado derecho se muestra una representación del pangenoma para cualquier número de genomas. Familias de metabolismo conservado tenderán a estar concentradas entre el *core* y el *shell genome*, mientras que el metabolismo especializado tendrá más representantes en el dispensable que en el *core genome*. Sin embargo tanto el tipo de metabolismo como el tamaño del *core*, *shell* y *genoma dispensable* pueden variar según la diversidad de los organismos seleccionados.

marcadores respecto de *Streptomyces lividans*, a pesar de la cercanía taxonómica entre estos organismos. La presencia de genes marcadores en el core depende de ambos linajes, por lo que es importante contar con algoritmos que permitan automatizar su cálculo.

El número de familias en el pangenoma, ya sea en el *core*, *shell* o *dispensable* genome no sólo depende de la divergencia o proximidad taxonómica de los organismos del linaje seleccionado, también depende de lo variable que sea el contenido génico en los genomas del linaje. A esta característica se le conoce como apertura. Hay especies, por ejemplo algunos patógenos, cuyo pangenoma se encuentra sumamente cerrado en el sentido de que no importa cuántos genomas se agreguen, el número de familias parece converger y ser asintótico rápidamente a una cota superior. En cambio especies o géneros que viven en una gran diversidad de hábitats suelen tener un pangenoma abierto. Esto significa que cada vez que se agrega un nuevo genoma aparecen otras familias que no estaban en los genomas anteriores. En los linajes con pangenoma abierto el número de familias nuevas al agregar un genomas seguirá una tendencia creciente y no asintótica.

Además de la apertura, existen otros intentos de cuantificar la diversidad génica de un linaje. Están por ejemplo la fluidez, definida como el promedio de familias únicas entre familias totales por pares de genomas. El pangenoma bacteriano total, es decir el total de familias génicas en el dominio Bacteria es considerado abierto.

Finalmente la distribución de las funciones metabólicas encontrada en los subconjuntos del pangenoma (*core*, *shell* y *dispensable* genome) está relacionada a la proximidad filogenética de los organismos seleccionados en el estudio. Entre más diversos sean los organismos menos familias dedicadas exclusivamente a metabolismo especializado abundarán en el *core/shell genome*. La diversidad provocará que lo único que tengan los genomas de estos organismos en común sean funciones conservadas por una amplia variedad de especies bacterianas. Ahora bien, muchas familias de metabolismo especializado provienen de reclutamientos de copias extra de familias de metabolismo conservado. Así pues aunque decrezca el número de familias con exclusividad en metabolismo especializado en el *core y shell genome*, estos subconjuntos del pangenoma aún pueden contener familias conservadas que tengan copias extra en proceso de reclutamiento para algún Cluster biosintético de genes (BGCs) de metabolismo especializado. Considerando las reflexiones anteriores, entre más diverso sea un linaje, más tenderá su *core genome* a contener exclusivamente familias de metabolismo conservado mientras que su *dispensable genome* estará formado mayormente por familias de enzimas del metabolismo especializado.

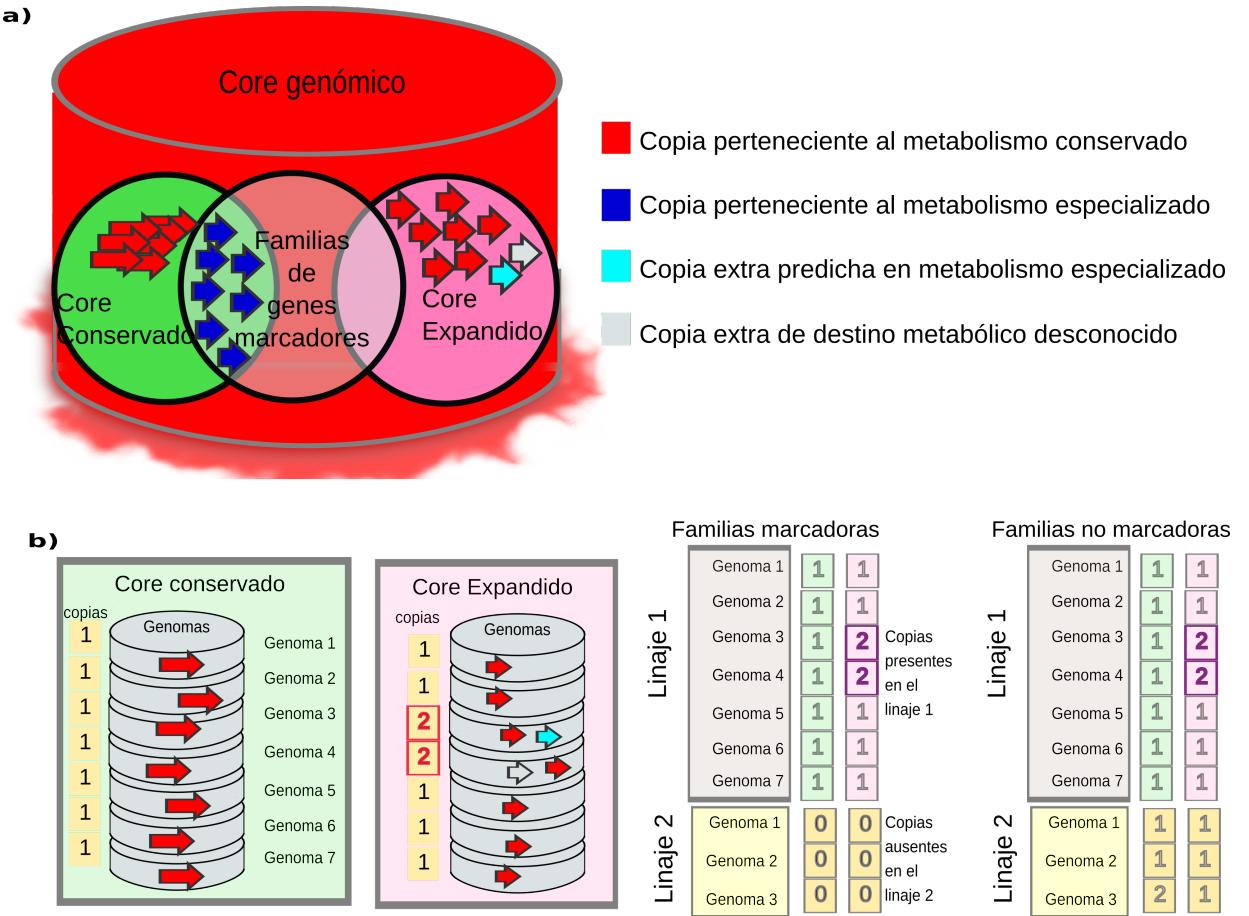


Figura 1.3: El core genome puede contener familias con funciones en distintos grupos metabólicos así como diversidad en el número de copias. Arriba se muestra que en el core pueden coexistir familias tanto de copia única como con expansiones. Las familias con funciones en el metabolismo conservado suelen concentrarse en el *core genome* (rojas), pero también dependiendo de los organismos seleccionados pueden encontrarse ya sea familias enteras o algunas copias dedicadas al metabolismo especializado (azul). No de todas las copias se conocerá su función, algunas pueden tener un destino metabólico desconocido (gris) o bien ser predichas por algún algoritmo como parte del metabolismo especializado (cyan). Abajo a la izquierda se comparan familias del *core conservado* con exactamente una copia por genoma contra familias del *core expandido*. Ambas pertenecen al *core*, pero en el *core expandido* hay dos genomas que tienen una copia extra en esta familia, una cyan y una gris, que podría dificultar la elección de los verdaderos ortólogos. A la derecha se ejemplifican familias de genes marcadores, útiles para identificar un linaje genómico. Tanto familias del *core conservado* como del *core expandido* pueden ser familias marcadores, siempre que exista al menos una copia de cada familia en el linaje 1 y ninguna copia en el linaje 2. Las familias dejan de ser marcadoras cuando el linaje 2 contiene al menos una copia en algún genoma.

## 1.2. El core conservado permite la reconstrucción de filogenias complicadas

Orthocore es el desarrollo bioinformático que realicé para calcular las familias génicas más conservadas del *core genome*. Dos genes son homólogos si poseen un ancestro común, entre los principales grupos de homólogos están ortólogos y parálogos. Los ortólogos provienen de eventos de especiación de un ancestro común mientras que los parálogos evolucionan por eventos de duplicación. Orthocore obtiene un subconjunto del *core genome*: el *core conservado*, es decir, familias de ortólogos presentes en todos los genomas del grupo y que además son libres de parálogos de difícil identificación. El *core conservado* facilita la organización en árboles filogenéticos de organismos de un linaje genómico.

La comparación de la variación molecular entre ortólogos ha sido utilizada para establecer relaciones filogenéticas entre organismos. Esta técnica ha dado lugar a grandes descubrimientos. Por ejemplo comparar la secuencia de la subunidad 16s del gene RNA ribosomal condujo a Woese al descubrimiento del dominio Archaea en 1977 [124]. Un árbol de especies suele hacerse con secuencias de familias que pertenecen al core genome de un Dominio, por ejemplo las familias 16s o RpoB en los Dominios Bacteria y Archaea. Algunos autores realizan árboles multilocus para mejorar la resolución de árboles de especies realizados mediante la comparación de secuencias de 16s. Los genes seleccionados para los árboles multilocus deben estar en todos los organismos y no tener copias extra tan parecidas que puedan confundirse y entorpecer la reconstrucción filogenética, es decir, las familias seleccionadas deben ser parte del *core conservado*. Orthocore automatiza la identificación de estas familias.

Entre los factores importantes para establecer las relaciones filogenéticas diferenciando entre Archaea y Bacteria están los siguientes: 1) la presencia conservada de la subunidad de 16s en los tres dominios mencionados, y 2) la suficiente divergencia entre estas secuencias en los organismos de dichos dominios. Ahora bien, establecer relaciones filogenéticas entre Archaea y Bacteria es en cierto sentido más sencillo que establecerlas entre organismos pertenecientes al mismo género o inclusive a la misma especie. En ocasiones, como en el caso del género *Streptomyces*, la secuencia de 16s por sí sola no posee la suficiente variación para resolver la filogenia [143]. En *Streptomyces* la variación entre estas secuencias suele ser menor al 1 %. Para resolver el problema de escasa variación en secuencias de 16s se pueden concatenar las secuencias de otros ortólogos, siempre que estos aparezcan en todos los organismos que se estén estudiando, es decir, siempre que sean parte del core genómico.

## 1.3. El algoritmo de Orthocore

Los ortólogos suelen identificarse por similitud de secuencia, pero si se realiza la identificación manualmente también se suelen capturar parálogos que pueden confundir

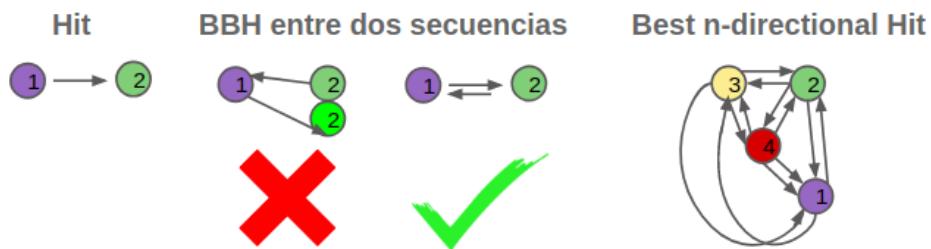


Figura 1.4: Orthocore utiliza los mejores hits n-direccionales para obtener grupos de ortólogos. Un hit es el mejor resultado de una secuencia en otro genoma. Un Bidireccional Best Hit es el mejor hit bidireccional. La secuencia 2 es el mejor hit de la secuencia 1 en el genoma 2 y reciprocamente, la secuencia 1 es el mejor hit de la secuencia 2 en el genoma 1. La existencia de una copia extra muy parecida a la secuencia 2 puede romper el BBH. Un mejor hit n-direccional debe ser BBH todos contra todos garantizando que estas secuencias están muy conservadas entre sí.

la elucidación de eventos de especiación. Orthocore automatizó la búsqueda de ortólogos y el filtrado de parálogos en genomas procariontes mediante la generalización de la definición del mejor hit bidireccional (BBH por sus siglas en inglés) Figure 1.4 . Dos secuencias son BBH si cada una es el mejor hit de un algoritmo de distancia (BLAST usualmente) en el genoma de origen de la otra. Una primera generalización para obtener el set de ortólogos de una familia del core es definir un genoma de referencia y tomar los BBH respecto a ese genoma. En la práctica, esta definición da como resultado distintos resultados según el genoma de referencia, haciendo que algunos parálogos no sea filtrados.

Para solventar esta dificultad se definió en Orthocore el concepto de mejores hits multidireccionales. Un conjunto de genes son mejores hits multidireccionales si todos entre sí son BBH por pares. Es decir si cada gen fuera un punto y ser mejor hit se expresara como una conexión con dirección todos los puntos estarían conectados por una flecha de ida y otra de regreso. Con este método se eliminó la dependencia de un genoma de referencia. Esta restricción también ocasiona que en grupos muy grandes por ejemplo más de 100 genomas de distintas especies, o muy diversos de distintos dominios, o muy fragmentados como con contigs de en promedio 3 Mbp, el core conservado puede quedar vacío.

### 1.3.1. Componentes técnicos de Orthocore

Orthocore es un tubería escrita en perl que incorpora los hits multidireccionales permitiendo obtener y usar el core conservado para realizar una reconstrucción filogenética mediante los siguientes pasos:

- Obtiene el core conservado: los mejores n-direccional hits (blastp).
- Alinea cada familia del core conservado (MUSCLE).
- Cura automáticamente cada familia del core conservado (Gblocks).

- 
- Concatena las familias del core conservado formando una matriz de aminoácidos.
  - Realiza una reconstrucción filogenética de la matriz de aminoácidos (FastTree)
  - Provee las funciones del core conservado según la anotación funcional de RAST.

Existen otros algoritmos como OrthoMCL[53], y fastOrtho[144] que dividen pangenos en clusters de familias de genes, get\_homologues [145] y metaphor que obtienen el core y filtran buscando verdaderas relaciones de homología, y finalmente BPGA[115] que hace reconstrucciones filogenéticas tanto según el core como según el pangenooma. Sin embargo Orthocore resolvió en su momento la necesidad específica de proporcionar una matriz concatenada de genes del core conservado lista para utilizarse en un árbol multilocus. Adicionalmente, como Orthocore fue diseñado para trabajar con la anotación de la plataforma RAST, también se obtiene la anotación funcional tanto de familias del core como del complemento.

Orthocore incorpora todas las dependencias en un contenedor de docker disponible en <https://github.com/nselem/orthocore>. Además en este contenedor está un script que permite bajar genomas de NCBI masivamente y posteriormente anotarlos en RAST desde la terminal. Los protocolos de uso se encuentran al final de este capítulo.

### 1.3.2. Aplicaciones de Orthocore, identificación del core conservado y de familias de genes marcadores.

Cuatro aplicaciones de Orthocore serán presentadas en las siguientes secciones de este capítulo. En la primera aplicación el core conservado en *Actinomycetales* permitió organizar filogenéticamente a este orden. Esta organización facilitó el entendimiento en cambios de promiscuidad de la familia enzimática PriA mediante la distinción de patrones de pérdida y ganancia de genes en las rutas de síntesis de histidina y triptófano. En la segunda aplicación cepas de *Salmonella* fueron ordenadas filogenéticamente. La tercera aplicación permitió realizar una reconstrucción filogenética del orden *Nostocales* del phylum *Cyanobacteria* y comparar así patrones de presencia y ausencia de clusters de genes biosintéticos. Finalmente, en organismos del microbioma del tomate orthocore de utilizó para identificar genes marcadores que permitieran distinguir cepas de *Clavibacter Michiganensis* de otras especies de *Micrococcaceae*.

Perfiles de promiscuidad de PriA en el orden *Actinomycetales* se relacionan con la especiación. Para mejorar el entendimiento de los cambios en promiscuidad enzimática de PriA, se necesitaba entender filogenéticamente al orden *Actinomycetales*, un camino era obtener su core conservado. Orthocore fue diseñado para resolver este problema. Con el resultado de Orthocore se realizó un árbol de especies donde se observaron patrones de pérdida y ganancia de genes en la vecindad genómica del gen que codifica para PriA. Se encontró que hay clados de *Actinomyces* donde los genes correspondientes a la síntesis de histidina no estaban en la vecindad genómica de PriA, y mediante la realización de cinéticas enzimáticas se comprobó que la actividad de catalizar la reacción correspondiente a HisA estaba perdida en estos organismos. A estas enzimas se les llamó subTrpF ya que sólo poseían la capacidad de catalizar la reacción correspondiente

a la familia TrpF. Del mismo modo existían clados que perdieron los genes de síntesis de triptófano en la vecindad de PriA y estas enzimas se subfuncionalizaron a la familia subHisA. De estos datos se observa que en estos organismos la especiación coincidió con el cambio de promiscuidad en la familia PriA, acorde a la pérdida y ganancia de genes vecinos. La promiscuidad puede co-ocurrir con variaciones en el contexto genómico, pudiendo estos cambios ser una marca para sugerir cambio funcional en una familia.

### **1.3.3. Genes de islas de patogenicidad de *Salmonella* en México están conservados en la mayoría de los genomas.**

Para estudiar la diversidad de *Salmonella* presente en alimentos en México, primero se ensamblaron y anotaron genomas secuenciados para el trabajo “distribución de los genes de la toxina VirB/D4 en plásmidos de bovino asociados a *Salmonella* no tifoidal”. Los genomas fueron ensamblados en Patrick y se desarrolló myRAST, una tubería previa a Orthocore para la anotación automática de genomas ensamblados en RAST. El protocolo de myRAST puede encontrarse al final de este capítulo.

Orthocore fue usado aquí para reconstruir filogenias de *Salmonella* y además fue integrado como parte de CORASON, el cual se reporta en el Capítulo 3, es el algoritmo que sirve para visualizar variantes de clusters organizados filogenéticamente. Los clusters de genes pueden ser biosintéticos, islas de patogenicidad, operones o cualquier región parcialmente sintética de un genoma bacteriano centrada en un gen. En este caso se observó una alta conservación de toxinas tifoidales en islas de patogenicidad de *Salmonella*. Éstas fueron identificadas en 76 % de las cepas analizadas y posteriormente visualizadas mediante CORASON.

### **1.3.4. Nostoc provenientes del metagenoma de cíadas se agrupan Cyanobacteria**

Las Cyanobacterias son un phylum de bacterias que se han adaptado a diversos ambientes. Aunque muchas de ellas son marinas algunas Cyanobacterias viven como simbiontes de plantas. En particular las cíadas han desarrollado un tipo especial de raíz donde se sabe que vive como simbionte el género *Nostoc*. La presencia de *Nostoc* en la raíz coraloides de las cíadas es fácilmente distinguible por la formación de un anillo verde conocido como anillo cyanobacterial. En la Figure 1.5 se muestra la filogenia de 76 Cyanobacterias de 7 órdenes distintos construida con 198 proteínas del core conservado obtenidas por Orthocore. En esta reconstrucción se puede observar que los *Nostoc* asociados a plantas tienden a agruparse en la filogenia.

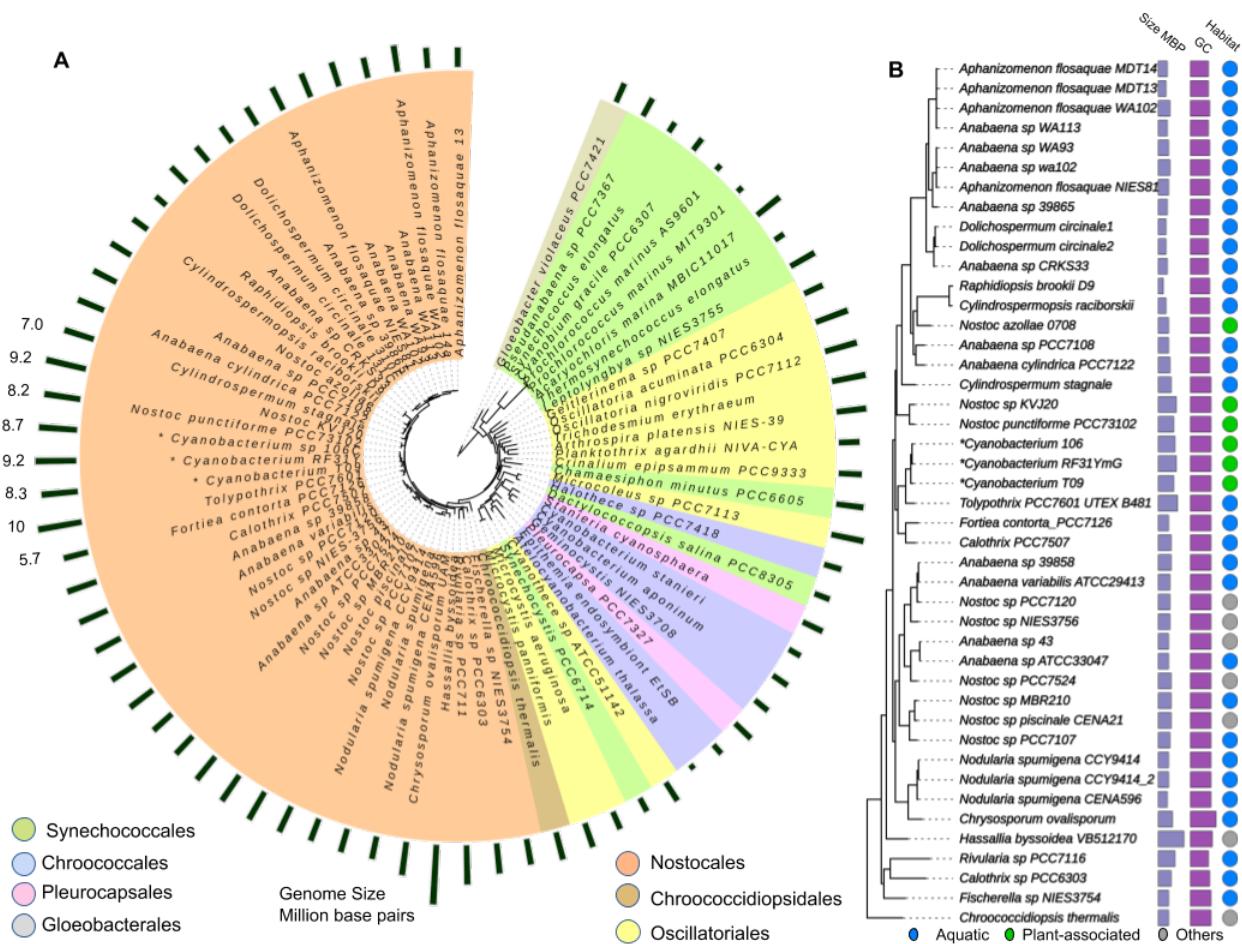


Figura 1.5: Reconstrucción de 76 taxa provenientes de 7 órdenes de Cyanobacteria.

La matriz final incluyó 45,475 aminoácidos curados de 198 familias de proteínas pertenecientes al *core conservado*. A la derecha se muestra un zoom sobre el orden *Nostocales*. En este orden se incluyen algunas bacterias simbiontes de cíadas. Metadatos como el tamaño de genoma, contenido de GC y hábitat de origen muestran una posible tendencia de incremento de tamaño en los genomas provenientes del microbioma de plantas.

### 1.3.5. Identificación de genes marcadores de *Clavibacter michiganensis*

*Micrococcales* es un orden de Actinobacteria que contiene a *Clavibacter*, *Micrococcus* y *Microbacterium*, entre otros microorganismos. El género *Clavibacter* comprende especies que pueden causar enfermedades en diversas plantas. En particular la especie *Clavibacter michiganensis* es una bacteria causante de la enfermedad del cáncer del tomate. *Clavibacter michiganensis* ha sido frecuentemente aislada en compañía de otros *Micrococcales* morfológicamente parecidos. La distinción entre microorganismos debida a la comparación de la secuencia de 16S no era suficiente para distinguir entre *Micrococcales* del microbioma del tomate, por lo que una prueba de diagnóstico se hacía necesaria. Se habían utilizado como marcadores genes como *tomA*, *ppaC* y *celA* entre otros, sin embargo estas elecciones en ocasiones resultaban en falsos positivos según árboles de especies de 16S, por lo que nuevos marcadores eran necesarios.

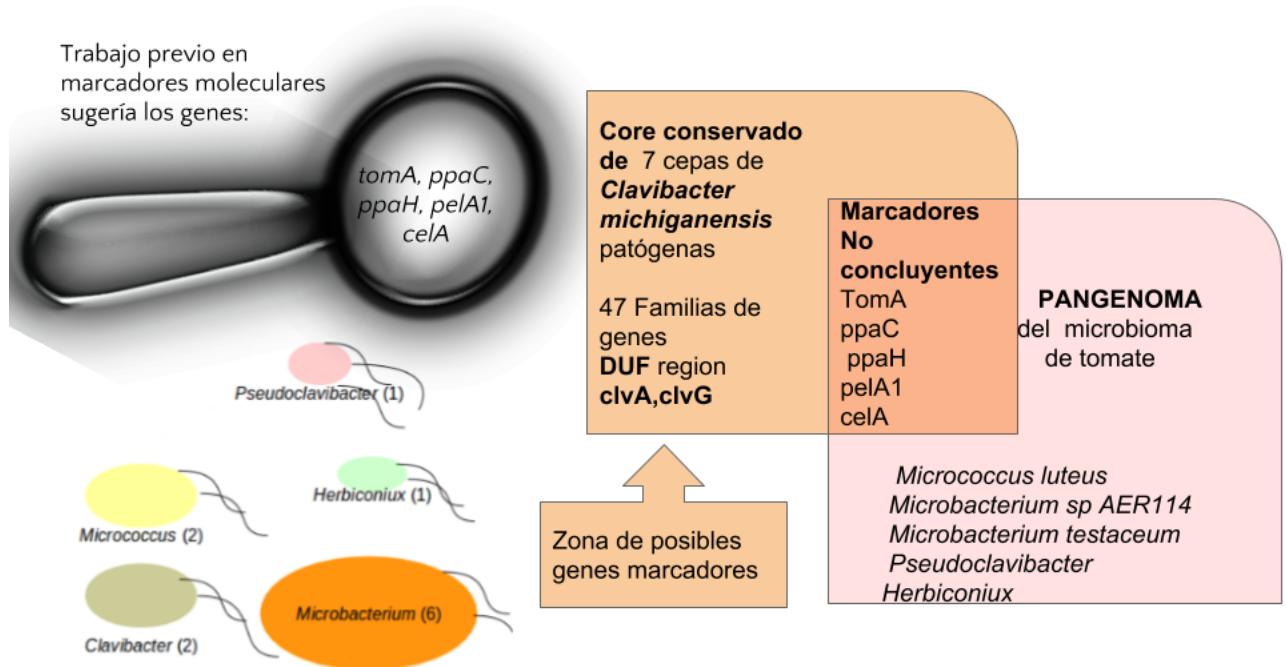


Figura 1.6: Los marcadores moleculares previos a este trabajo no permitían diferenciar correctamente a *Clavibacter michiganensis* respecto al microbioma del tomate en invernaderos mexicanos. Arriba a la izquierda se muestran los antiguos marcadores *tomA, ppaC, ppaH, pelA1, celA*. Abajo organismos pertenecientes al microbioma del tomate. Orthocore obtuvo el core conservado de siete cepas de *Cmm* patógenas, y este core se utilizó para definir nuevos marcadores. Aunque *tomA* pertenecía al core conservado de *Cmm*, estaba también incluido en el pangenoma del microbioma del tomate. Después de calcular la intersección core conservado de *Cmm* y pangenoma del microbioma se obtuvieron entre las familias marcadoras genes *clv* parte del cluster biosintético de clavidicina (michiganina).

Al analizar en RAST genomas de *Microbacterium* y *Micrococcus* aislados de tomate

se encontró que en efecto, *tomA* y los otros marcadores propuestos previamente no eran exclusivos de *Clavibacter michiganensis* (*Cmm*). Al utilizar Orthocore en siete genomas de *Cmm* encontramos que varios genes del cluster biosintético de michiganina (BGC0000528 en MIBiG) codificado por los genes *clvAFGLKM* pertenecían al core conservado, pero que al agregar los genomas no *Cmm* del resto del microbioma del tomate los genes *clv* se pierden. El descubrimiento de que *clv* pertenecía al core de *Cmm* se realizó con secuencias de genomas muy fragmentados, en la Figure 1.6 se muestran las cepas originales que fueron analizadas.

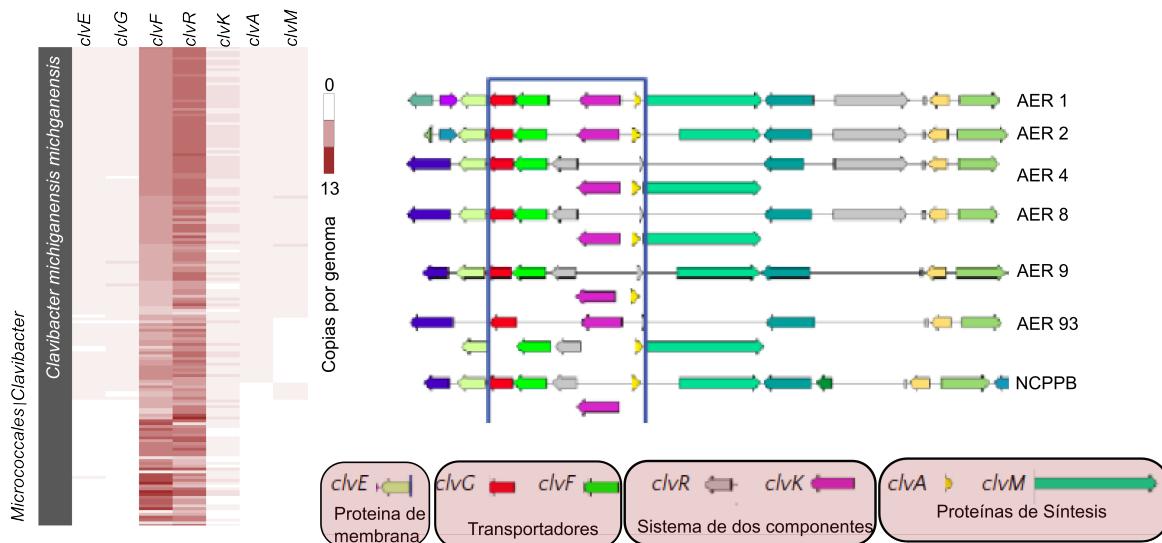


Figura 1.7: El Cluster *clv* es un marcador de la especie *Clavibacter Michiganensis* en el orden *Micrococcales*. Los genes *clvEGAM* que son una proteína de membrana, un transportador y dos proteína de síntesis del BGC *clv* están muy conservados en la especie *Clavibacter Michiganensis*. *clvFRK*, es decir un transportador y el sistema de dos componentes están presentes en otros *Micrococcales* pero con baja identidad de secuencia y nunca en el contexto del BGC de clavidicina

Esta observación se corroboró con más genomas, Figure 1.7 en este caso se muestran como ejemplo 10 genomas de bacterias del microbioma del tomate, entre ellas siete *Clavibacter*, seis de tomates de invernadero y uno *Clavibacter* proveniente de tomate silvestre, *Clavibacter RA1B* que fue reportado en la tesis de maestría de Yanez. Con Orthocore vemos que el tamaño del core decrece al ir agregando genomas de *Clavibacter* y decrece aún más rápido al agregar los genomas de *Micrococcus* y *Microbacterium*. La reconstrucción filogenética de este microbioma, ubica a *Clavibacter RA1B* cerca de los otros *Cmm* pero no en un clado junto con ellos. Una búsqueda por blast revela que los genes marcadores de *Cmm*: *clvF*, *clvR* son también marcadores de *Clavibacter RA1B* pero no así *clvA* y *clvG* que solamente están presentes en el core de *Cmm*. Sin embargo *clvF* y *clvR* no están en el contexto del cluster de michiganina en RA1B y su similitud de secuencia es menor que la que se observa entre los otros *Cmm*.

De hecho al considerar más genomas dentro del microbioma del tomate, la familia *clvF* no solo no está en el core conservado de *Microbacterium* y *Micrococcus*, sino que

no está presente en ningún otro genoma distinto a *Clavibacter*. Con distintos niveles de conservación de secuencia los genes *clv* son un buen marcador para distinguir *Cmm* de otras especies, por esta razón estos genes aún se encuentran en uso como genes marcadores de *Cmm*. Esta misma conclusión se reporta en el trabajo de yasuhara, su observación de que las familias *clvA*, *clvF* y *clvG* son exclusivas de *Cmm* es independiente de la nuestra, fue hecha sin análisis genómicos y basada en evidencia experimental [146]. Este descubrimiento ha permitido bajar los costos de identificación de *Clavibacter*, ya que ahora en lugar de enviar a secuenciar el genoma es suficiente identificar por PCR *clvF* en conjunto con otro marcadores.

Con Orthocore además de obtener los genes marcadores podemos obtener también la matriz del core conservado para realizar la reconstrucción filogenética de especies cercanas de *Clavibacter* Figure 1.8 . Debido a que es importante para los agricultores conocer de dónde provienen las bacterias que infectan al tomate, una de las líneas de investigación del laboratorio de evolución de la diversidad metabólica, es la organización taxonómica de cepas de *Cmm* y de otras bacterias del microbioma de tomate. Por este motivo se han secuenciado y mantenido como datos privados unos doscientos genomas provenientes del microbioma del tomate. Orthocore colabora con esta investigación proporcionando matrices multilocus que pueden diferenciar entre cepas de *Clavibacter* de la misma o de diferente especie.

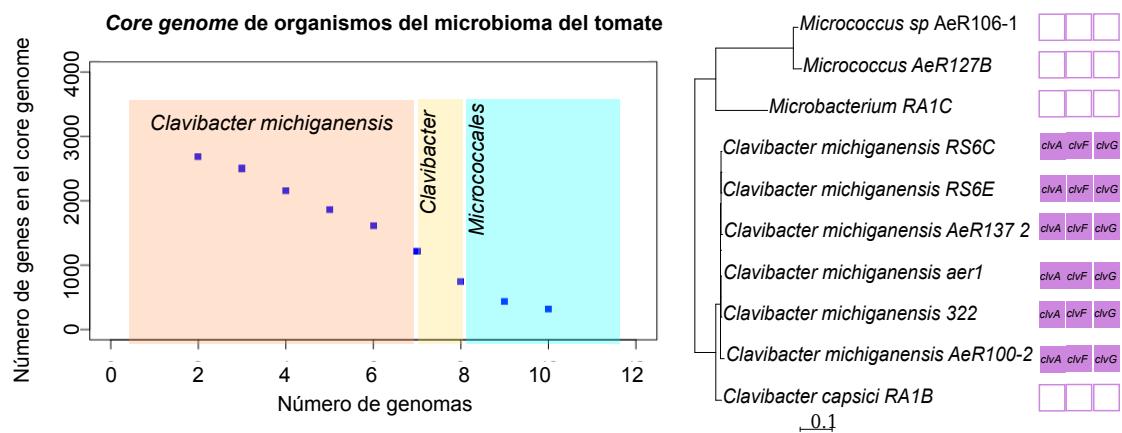


Figura 1.8: El *core genome* de *Micrococcaceales* ayuda a identificar genes marcadores. El *core genome* de siete organismos del microbioma del tomate muestra de

Debido al intercambio génico por transferencia horizontal en las bacterias, es posible que los genes marcadores actuales alguna vez aparezcan en otros organismos. También las bacterias pierden continuamente genes, por lo que es posible que algún gen marcador de *Cmm* se pierda en una cierta cepa . Esto hace que la definición de estar presentes en el core del grupo de interés y ausentes totalmente de cada uno de los genomas de otro linaje ya no funcione completamente. Sin embargo, en estas dos situaciones presentadas, ganancia de genes marcadores de organismos externos al linaje original o pérdida de genes marcadores en algunas cepas, se sigue cumpliendo que los ex genes marcadores, estarán presentes en la mayoría de los organismos del linaje de interés y ausentes de la mayoría de los organismos del linaje externo. Por ello, se pensó que

está definición de genes marcadores se podía generalizar clasificando a grupos de genes ortólogos acorde a sus porcentajes de ocurrencia. Esta idea se desarrolló en la herramienta clavvisual, explicada en la siguiente sección.

## 1.4. Clavvisual: Identificación de genes marcadores a un cierto porcentaje de grupos seleccionados

La idea de que Orthocore puede ser usado para obtener los genes marcadores de un grupo taxonómico frente a otro fue generalizada en el backend del software Clavvisual. Ya se ha explicado previamente que el core puede salir vacío por diversas razones, entre ellas baja calidad de los genomas, o que éstos provengan de organismos muy divergentes, verdaderas razones biológicas como dinámica génica o un core no convergente. Así pues, es posible que si sólo se utiliza el core no se obtengan marcadores. Pero el core puede relajarse de varias maneras una de ellas es el Pseudocore, donde en lugar de multidireccional hits se toman BBH a un genoma de referencia. Otra forma es establecer un porcentaje de presencia /ausencia de interés.

El pseudocore consiste en utilizar información previa específica del core de un linaje para obtener los genes de y la metodología está depositada en github integrada en el repositorio clavigenomics. El blast fue optimizado cambiando hacer un blast todos contra todos por archivos genómicos individuales `genomai_vs_genomaj.blast` que luego son concatenandos según se necesiten.

Los porcentajes de genomas son diferentes porque al no bastar los best bidireccional hits conservados, todo el pangenoma es decir todos los genes contenidos en los genomas del grupo de interés necesitan ser clasificados por familias, para de ahí obtener las familias que tienen presencia en un porcentaje %p y ausencia en un porcentaje a % del grupo externo. Estos perfiles fueron desarrollados para Clavvisual Figure 1.9 utilizando FasthOrtho para clasificar las familias y de ahí obtener los grupos. Con ellos se consiguieron marcadores para *Kurtobacterium*.

Finalmente Clavvisual despliega un árbol realizado con el Pseudocore respecto a un conjunto de genes de Cmm NCPP previamente seleccionados. En este árbol clavvisual permite la visualización de metadatos, como año, género de la bacteria, estado de salud de la planta e invernadero donde fue aislada.

### 1.4.1. El pangenoma de *Clavibacter Michiganensis* es abierto

Después de desarrollar métodos de identificación de genes marcadores y generalizarlo a obtener grupos con patrones de presencia/ausencia definidos por el usuario, quedaba

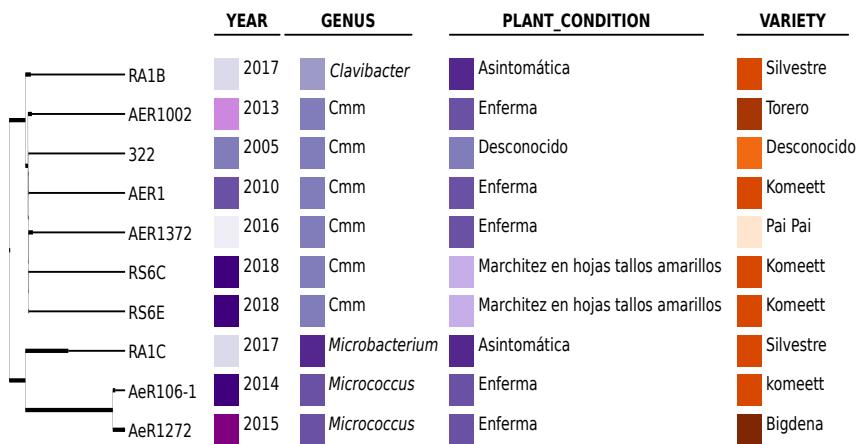


Figura 1.9: Clavisual organiza filogenéticamente cepas de *Clavibacter* y generaliza orthocore. Clavisual no se restringe a la búsqueda del core conservado, permite la identificación de familias conservadas sólo en un cierto porcentaje. Además puede filtrar estas familias permitiendo sólo las que estén ausentes en una proporción mínima de un grupo de genomas de interés. Además orthocore optimiza el funcionamiento de Clavisual ya que fue utilizado para proveer un conjunto de familias génicas conservadas en el género *Clavibacter*. Estas familias son buscadas cada vez que ingresa un genoma nuevo y con ellas se construye un árbol filogenético. Los metadatos disponibles son desplegados al lado de la cepa correspondiente del árbol.

por responder la pregunta cómo es el pangenoma de *Cmm*. Algunos autores consideran que el pangenoma de patógenos es reducido porque sus genomas suelen sufrir proceso de reducción de tamaño debido a la pérdida de genes. Como *Cmm* es un patógeno de planta quedaba por investigar cómo es su pangenoma. ¿Es posible saturar el contenido génico de *Cmm* con sólo secuenciar más genomas?. Aunque actualmente existen ya herramientas web para el análisis de pangenoma, en su momento se utilizó el software bpga que se corre desde la terminal. Para facilitar su instalación se desarrolló un docker que se describe más tarde en este capítulo en la sección de las descripciones técnicas. Como ejemplo de su funcionamiento, se analizó el pangenoma de los mismos diez genomas del pangenoma del tomate utilizados en la visualización de clavisual Figure 1.10 .

Tomando otros siete genomas del género *Clavibacter*, utilizando OrthoVenn tenemos la misma observación, el número de familias de genes agregadas al adicionar genomas, es después de siete genomas casi tan grande como su core Figure 1.11.

## 1.5. Relación entre genes marcadores, Orthocore y la promiscuidad enzimática.

Finalmente, al aplicar Orthocore para detectar genes marcadores se vuelve indirectamente reclutamientos al metabolismo especializado, cómo, pues porque dentro de los marcadores hay productos naturales como, PENDIENTE VER RESULTADOS

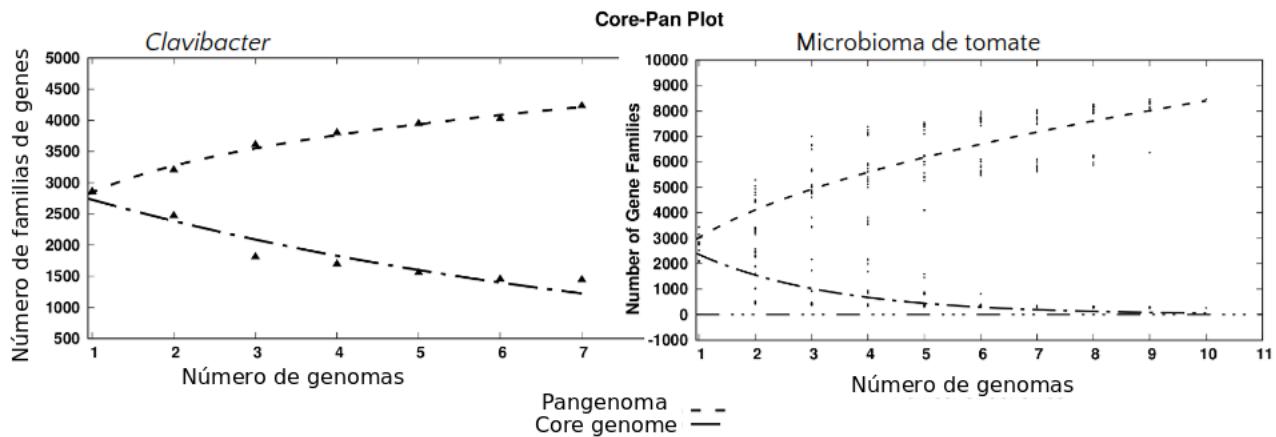


Figura 1.10: El pangenoma de *Clavibacter* es abierto según el análisis de BPGA. En este ejemplo el pangenoma de *Clavibacter* se mantiene creciente mientras que el core no alcanza el cero. En contraste, al ampliar el rango taxonómico el *core genome* sí se acerca a cero según BPGA

RETRO-EVOMINING la clavidicina (michiganina). Ahí vemos que clvABCDEF participan en metabolismo secundario, las enzimas de metabolismo central de este cluster pueden presentar cierta promiscuidad.

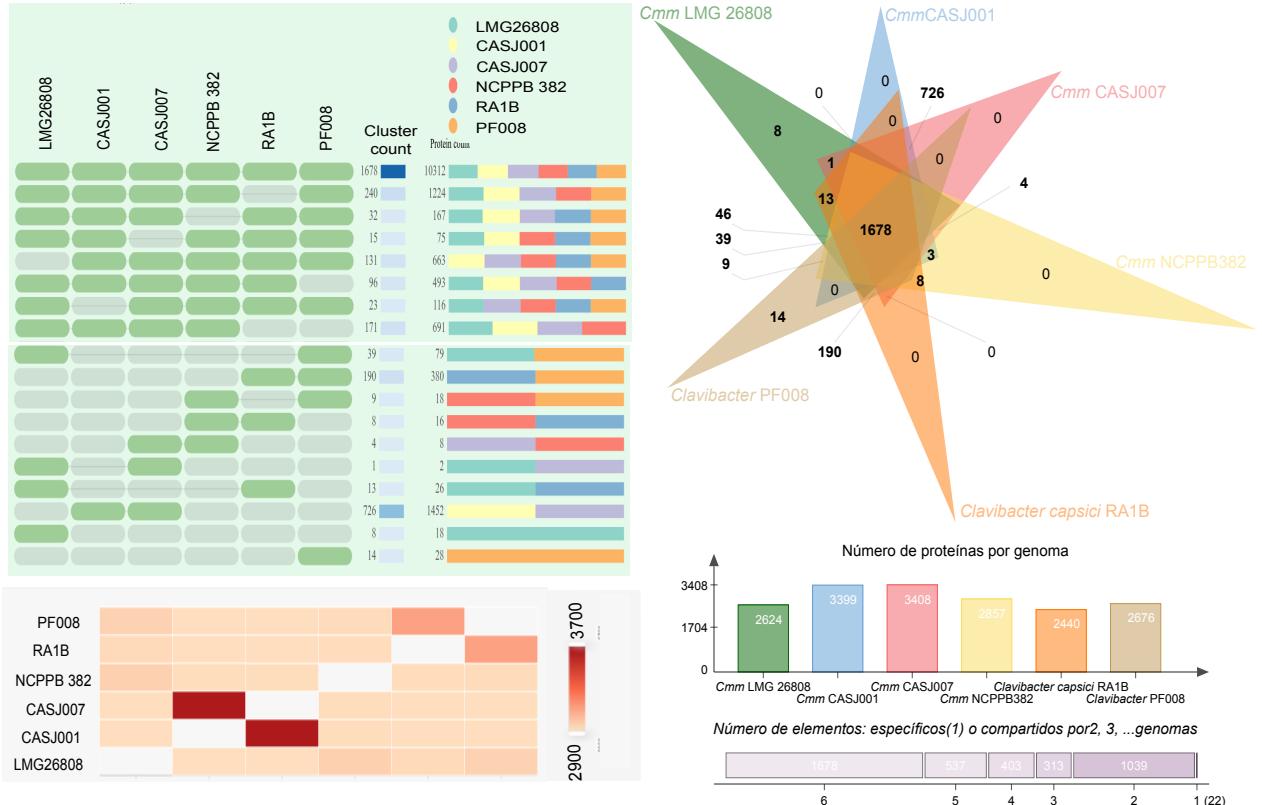


Figura 1.11: Diagrama de Venn del pangenoma de genomas selectos del género *Clavibacter*. Venn-Chart permite visualizar la distribución de las familias génicas mediante un diagrama de Venn siempre que sean pocos genomas. En la parte superior izquierda se muestra el número de familias compartidas entre distintos grupos de genomas. Por ejemplo existen 1678 familias génicas compartidas entre los seis genomas. Entre los grupos de cinco genomas, el más abundante es el que no contiene a RA1B con 243 familias, así pues RA1B es de los genomas más divergentes de este grupo.



# Capítulo 2

## EvoMining como herramienta para identificar el origen y el destino metabólico de familias enzimáticas

### 2.1. Introducción

#### 2.1.1. Las copias extra de familias enzimáticas que son reclutadas para una nueva función están relacionadas con la promiscuidad.

La promiscuidad enzimática puede buscarse en familias envueltas en procesos de divergencia funcional REFERENCIA. Uno de dichos procesos es la expansión de familias pertenecientes a rutas metabólicas conservadas y su posterior reclutamiento hacia el metabolismo especializado REFERENCIA. Cuando se identifica que en una misma familia de enzimas existe un subgrupo de homólogos con una función del metabolismo central bien caracterizada y al menos otro homólogo con evidencia experimental de poseer otra función bioquímica, se puede inferir que la neofuncionalización probablemente ocurrió a través de promiscuidad enzimática. Los homólogos con la función ‘secundaria’ suelen ser producto de expansiones previas de genes con la función ‘primaria’ REFERENCIA. En un trabajo previo se demostró para un caso en el que efectivamente la enzima XXX realiza la función primaria mientras que XXX cataliza la función secundaria mientras que los homólogos intermedios son promiscuos REFERENCIA. EvoMining es un algoritmo que sigue esta estrategia para identificar cambios en la promiscuidad de una familia de enzimas dentro de un linaje definido por el usuario. Al cambiar la familia de enzimas en la que se buscan neofuncionalizaciones se puede encontrar cuáles familias han sido más frecuentemente promiscuas dentro de un linaje, mientras que si se busca cómo han evolucionado las funciones conocidas de

una misma familia en distintos linajes se pueden descubrir los patrones característicos de cada grupo de organismos. Además de que EvoMining permite analizar los orígenes y destinos metabólicos de las familias enzimáticas para entender la evolución del metabolismo, también permite la identificación de rutas de metabolismo especializado que frecuentemente conduce al descubrimiento de nuevos productos naturales.

Los productos naturales o metabolitos especializados son sintetizados generalmente por *clusters* de genes distribuidos en un pequeño porcentaje de los organismos de un linaje taxonómico. Estos *clusters*, conocidos como BGC (*Biosynthetic Gene Clusters*), contienen copias extras de genes de familias que pertenecen al metabolismo conservado. En este trabajo aprovechamos que en la actualidad es posible predecir nuevos BGC mediante estrategias bioinformáticas gracias a la gran cantidad de secuencias disponibles públicamente así como la facilidad para secuenciar nuevos genomas. La similitud de secuencia de los genes que pertenecen a los BGC así como su sintenia en diversos organismos de un linaje hacen que genómica comparativa sea de utilidad para intentar localizarlos.

En este capítulo se explica el desarrollo de EvoMining como plataforma bioinformática dedicada a presentar una visualización del origen y destino de todas las copias de familia escitonemina enzimáticas provenientes del metabolismo conservado. Se discutirá también la evolución de las expansiones de familias génicas en cuatro linajes genómicos Actinobacteria cianobacteria, *Pseudomonas* y Archaea. Finalmente se analizarán BGC que fueron detectados a partir del uso de EvoMining con énfasis en el de la escitonemina.

## 2.2. **EvoMining es un método para encontrar BGC no tradicionales**

### 2.2.1. **EvoMining es un paradigma que permite ubicar copias extra de familias enzimáticas y organizarlas visualmente acorde a eventos evolutivos.**

Existen varias clases de BGC que son arquetipos de los productos naturales. Entre ellas se encuentran las clases *non ribosomal peptide synthetase* (NRPS), *poliketide synthase* (PKS), terpenos, péptidos ribosomales modificados post-traduccionalmente (RIPPs), alcaloides, etc. En estas clases, hay enzimas cuya presencia lleva a la detección de los BGC. Por ejemplo, las sintetasas no ribosomales son las que al encontrarlas dan nombre a los BGC tipo NRPS y las policétido sintetasas son las que dan nombre a los BGC de la clase PKS. No todos los productos naturales están dentro de los BGC clásicos. Dentro de MIBiG (BGC caracterizados experimentalmente) hay 231 BGC [REF 2012,2016,...] (12,7 %) clasificados como “otros”, que de hecho carecen de PKS, NRPS o cualquiera de las otras clases de enzimas características del metabolismo especializado. Como ejemplo se muestra la Figure 2.1 donde se aprecia que un porcentaje de los BGC

reportados tanto en Actinobacteria como en cianobacteria pertenece al grupo “otros”. La ausencia de enzimas biosintéticas conocidas hace que estos BGC sean “atípicos”, difíciles de identificar. Los BGC no tradicionales suelen pasar desapercibidos porque no hay conocimiento previo de ellos que permita reconocerlos.

EvoMining implementa una estrategia de búsqueda de divergencia del metabolismo conservado en lugar de la estrategia de búsqueda de similitud con metabolismo especializado, lo que permite identificar BGC que no pertenecen a ninguna categoría del metabolismo secundario previamente descritos. Para ello facilita la identificación de las ramas divergentes en el registro de secuencias de una familia de enzimas analizada a través de la evolución y las utiliza como una marca que sugiere funciones divergentes del metabolismo conservado. De esta forma se puede localizar alguna enzima de un BGC no clasificado. Como ya se mostró en EvoMining 1.0 encontramos una enzima que FALTA FALTA YA NO HACE LA REACCION SOBRE EL METABOLITO CON AZUFRE SINO CON UN ANÁLOGO QUE TIENE ARSÉNICO EN SU LUGAR, luego de identificar esta enzima divergente de su homóloga de metabolismo central pudimos identificar que era parte de una región con X genes que además mantienen la sintetía en un clado de los ACTINOBACTERIALDETES. Lo que constituyó el primer caso de un BGC con química nueva predicho a partir de secuencias de enzimas que eran divergentes de sus homólogas de metabolismo central[REFYAÑEZ ET AL], XXX FALTA FALTA FALTA HASTA AQUI Otro ejemplo de este escenario es el BGC de la escitonemina [147], un pigmento cianobacteriano que absorbe luz UV. Su biosíntesis requiere de ScyB y ScyA, dos enzimas que sostienen la síntesis de este metabolito especializado [148]. Curiosamente, ScyB y ScyA son homólogos distantes de la glutamato deshidrogenasa (GDH) y la acetolactato sintasa (ALS), respectivamente, que participan en la desaminación oxidativa reversible del glutamato a  $\alpha$ -ketoglutarato y amoníaco [150] y en la síntesis de aminoácidos de cadena ramificada [151], respectivamente.

## 2.3. Algoritmos y bases de datos de EvoMining 2.0

EvoMining está compuesto de dos algoritmos: el primero que encuentra miembros de la familia semilla entre todos los genomas blanco para así detectar genomas donde haya habido expansiones y luego busca cuáles homólogos seguramente tienen la misma función que las secuencias semilla y cuales otros miembros de la familia son más similares a genes que han sido reclutados por BGCs de acuerdo a reportes previos. El segundo algoritmo permite la visualización de todas las copias de una familia expandida en un árbol clasificadas según sus posibles destinos metabólicos. Para ello, los algoritmos EvoMining necesitan tres bases de datos: i) los genomas blanco, ii) las secuencias de enzimas semilla, y iii) la de productos naturales verificados.

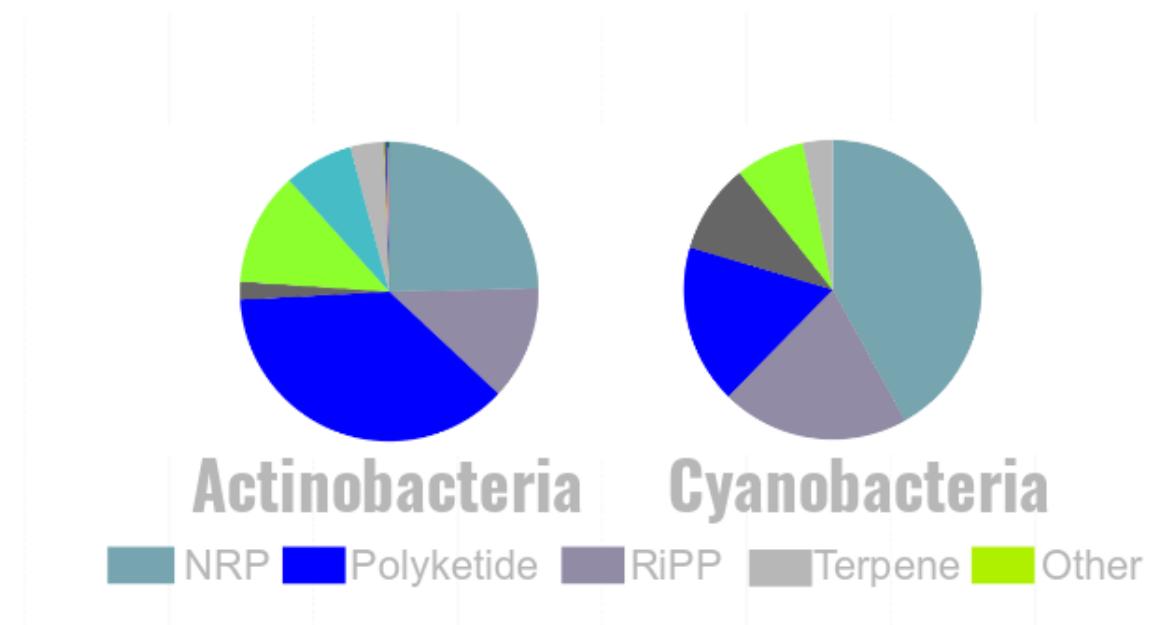


Figura 2.1: Existen cluster biosintéticos no clasificados (other) entre los reportados en MIBiG

### 2.3.1. Algoritmo de expansión y clasificación de grupos de homólogos.

La primera parte de la minería genómica evolutiva de EvoMining consiste en detectar las expansiones de la familia semilla y clasificar los homólogos que son muy similares a la secuencia de las enzimas con la función primaria separándolos de los que son más similares a genes que han sufrido reclutamientos a BGC. Este algoritmo requiere las siguientes entradas:

- Secuencia semilla (Enzyme DB). Es una secuencia o un conjunto de secuencias de la misma familia que serán interpretadas como las enzimas con una función primaria. Por esta razón es importante que el usuario haga una curación minuciosa en la que se sugiere que utilice genes con evidencia experimental que sí sean de una familia enzimática única y conservada. La implementación actual de EvoMining 2.0 permite que se introduzcan en el archivo de entrada varias familias bien identificadas simultáneamente, sin embargo el algoritmo procesa una familia a la vez.
- Base de datos de genomas blanco (Genome DB). Son las secuencias genómicas de todos los organismos en los que se quiera realizar el análisis. Se recomienda que los genomas incluidos pertenezcan a un mismo linaje taxonómico.
- Base de datos de genes reclutados a BGC (NP DB). La versión actual utiliza MIBiG, que contiene genes de BGC que se han demostrado experimentalmente.

La primera parte de este algoritmo consiste en identificar las Familias Expandidas (FE).

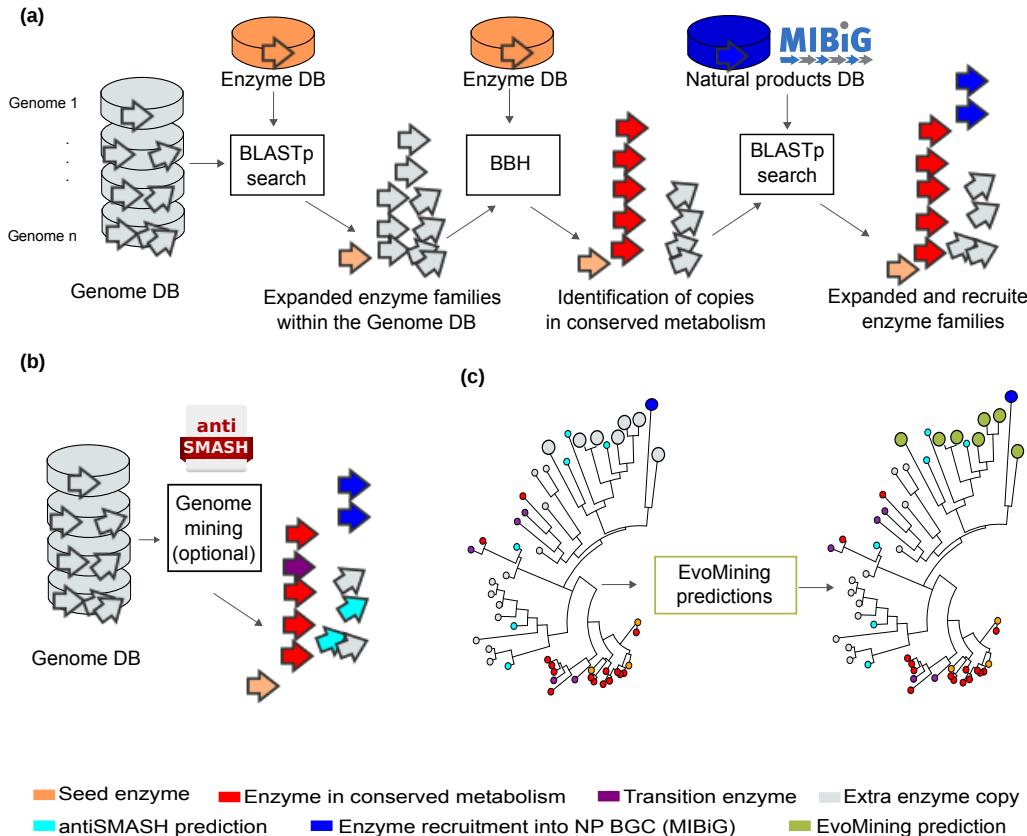


Figura 2.2: Representación de la tubería bioinformática de los dos algoritmos que componen EvoMining. a) Algoritmo de expansión - clasificación/reclutamiento. Se ingresan la Enzima DB (cilindro anaranjado) y la Genome DB (cilindros grises) para identificar por medio de BlastP a todos los que sean miembros de la Familia de las enzimas en Enzima DB, luego se identifican los ortólogos de la base de datos de enzimas (Flechas rojas), finalmente se buscan homólogos de la familia similares a genes reclutados a BGC (Flechas azules), b) EvoMining toma en cuenta los cuando se hace una búsqueda en AntiSMASH. Cuando un gen es encontrado como miembro de un BGC por antiSMASH es cian si no se le había encontrado en las categorías anteriores o morado si antes había sido detectado como rojo. c) El Algoritmo de visualización muestra un árbol donde cada homólogo de la familia expandida tiene una código de color que lo clasifica como se describe en la leyenda

Una FE consiste en todas las copias detectadas mediante una búsqueda con BLASTp, e-value de 0.001 y bitscore de 100, usando como *query* las secuencias de aminoácidos de las enzimas semilla (Enzyme DB) y como base de datos de búsqueda las secuencias de genomas blanco de un linaje taxonómico (Genome DB). En la primera versión de nuestro trabajo definimos que un organismo poseía una expansión si el número de copias de una familia estaba por encima del promedio mas dos desviaciones estándar. Siguiendo esta definición EvoMining colorea en un heat plot las expansiones de las familias enzimáticas respecto a un linaje taxonómico, señalando explícitamente el número de copias de cada familia que se haya agregado como semilla FIGURAX.

Una vez detectados todos los homólogos de la FE sigue encontrar a los ortólogos más parecidos a las enzimas semilla de la enzyme DB mismos que son identificados por *Best Bidirectional Hit* (BBH). Estos ortólogos serán considerados parte del metabolismo conservado y serán posteriormente identificados con color rojo en la visualización. Se infiere que son enzimas que tienen la misma función que las semillas. Por otra parte, las copias extra de la familia expandida hasta este punto son enzimas de las que no se conoce el destino metabólico. Es posible que en los pasos subsecuentes de EvoMining sean reconocidas como posibles reclutamientos a metabolismo especializado con el uso de antiSMASH y MIBiG. En otro caso permanecerán como copias extra con destino metabólico desconocido.

El último paso de este algoritmo consiste en encontrar miembros de la familia que se encuentren reclutados a BGC reportados anteriormente. En este paso definimos un reclutamiento como una copia extra de una familia de metabolismo conservado que ahora participa en un *cluster* de metabolismo especializado. Ejemplos de reclutamientos conocidos han sido observados en BGC reportados en MIBiG. En esta parte, los homólogos de la FE que tengan alta similitud con algún gen de MIBiG se clasifican como casos de reclutamiento. Una vez obtenidas las FEs expandidas e identificado los ortólogos del metabolismo central, se conduce una nueva búsqueda con BLASTp, E-value 0.001, utilizando las EFs como *queries* contra la base de datos NP DB, la de enzimas biosintéticas presentes en BGC. Podrían utilizarse otras bases de datos que contengan enzimas con un destino metabólico conocido en lugar de NP DB. Hasta este punto se tienen clasificados los genes con la función primaria y en este paso se designan los que tendrían una función divergente. Es posible agregar un análisis opcional en el que se agregan predicciones de los genes que están en BGC de acuerdo a las predicciones de antiSMASH. AntiSMASH busca secuencias con dominios conocidos de BGC tradicionales en la base de datos de genomas blanco (Genome DB) y además busca entre los genes circundantes la presencia de otras enzimas que ya hayan sido reclutadas a BGC conocidos. Si alguna enzima de la FE es detectada por antiSMASH nos será también útil para identificarla como el producto de una expansión que está en proceso o se ha neofuncionalizado. La minería genómica tradicional realizada por antiSMASH no es parte de la tubería de EvoMining, pero puede las predicciones de antiSMASH pueden ser calculadas previamente por el usuario y utilizadas por EvoMining. En este trabajo antiSMASH 3.0[152] fue utilizado en los secuencias de los genomas de la Genome DB Figure 2.2 panel (b).

Los usuarios de EvoMining, por lo tanto, deben definir de antemano las familias enzimáticas semilla más apropiadas para un determinado grupo taxonómico. El Enzyme DB seleccionado debe contener un conjunto de familias donde se puedan detectar los patrones de expansión. A su vez, las familias con una distribución restringida a un pequeño porcentaje de genomas no son adecuados para el análisis de EvoMining, situación que se puede detectar en el heat plot en el primer paso del primer algoritmo de EvoMining. También es importante determinar qué familias enzimáticas están compartidas por la mayoría de los genomas dentro de los linajes genómicos de interés, y si esto es importante para el tipo de análisis de EvoMining que se realizarán. El EvoMining DB original incluía las familias curadas manualmente que solo incluían enzimas metabólicas centrales, pero éstas no representaban el repertorio enzimático central de Actinobacteria. Esto se relaciona con la dificultad de definir qué es el metabolismo central; por lo tanto, preferimos utilizar el término enzimas centrales en diferentes umbrales de conservación. En nuestro caso, usamos 50 % para definir las enzimas del *shell*. Esta noción implica la posibilidad de automatizar la integración de Enzyme DB mediante la selección de familias enzimáticas en cualquier linaje genómico dado, evitando la necesidad de definir arbitrariamente qué es el metabolismo central.

En conclusión el algoritmo de expansión - clasificación/reclutamiento trabaja para identificar tres clases de copias en las familias enzimáticas expandidas y tiene como resultado tres salidas. La primera es una matriz que contiene el número de genes de cada familia semilla por cada genoma que se haya provisto y que puede ser visualizado e interpretado independientemente de los otros resultados y procesos de EvoMining. La segunda es una tabla por cada FE que enlista todos los homólogos pertenecientes a ella así como su identificador de a qué categoría pertenece, (i) copias altamente conservadas con algún miembro en el metabolismo conservado; (ii) reclutamientos conocidos en BGC de productos naturales; y (iii) copias extra que no son reclutamientos conocidos ni parte obvia del metabolismo conservado, quedando por definir su destino metabólico. La última salida son las secuencias de los genes de la FE, que servirán para hacer un análisis de su evolución mediante una filogenia en el siguiente algoritmo de visualización.

### **2.3.2. Algoritmo de reconstrucción filogenética y visualización**

Una vez que se tienen definidos los genes de una familia expandida (FE) de enzimas y que ya se conoce cuáles de los homólogos realizan la función primaria así como los que han sido reclutados a BGC para realizar una función divergente, en este algoritmo se visualiza esa información junto con la inferencia de los genes que no han sido clasificados todavía. Para eso se alinean las secuencias de la FE y se construye un árbol filogenético para luego visualizar cada uno de los homólogos como una hoja del árbol con un código de color que clasifica de acuerdo a la inferencia de origen-destino.

Las secuencias de la FE detectadas por el algoritmo de búsqueda - clasificación/reclutamiento se alinean con muscle v3.2 y son curadas con Gblocks. Las posiciones ausentes en más del 50% de las secuencias se filtran y remueven del alineamiento final. Para reconstruir filogenéticamente la historia de las enzimas se utiliza FastTree 2.1[153] que es un método muy rápido pensado para miles de secuencias que utiliza un algoritmo definido como “aproximadamente máxima verosimilitud”. Así se obtiene un árbol en formato newick, mismo que puede ser utilizado con el software de visualización de Microreact[154] y que también es usado por EvoMining para ser visualizado en su propia plataforma.

Los árboles visualizados en EvoMining diferencian entre la función metabólica de cada miembro de una familia génica mediante un código de color Figure 2.2 panel (b, c) . Las secuencias más conservadas se identifican mediante BBH contra la Enzyme DB, los hits de este proceso son considerados copias de metabolismo central, y son marcadas en rojo. En el otro extremo están los reclutamientos conocidos con alguna evidencia experimental que fueron reportados en MIBiG[105]. Estos reclutamientos son marcados en azul. Una vez definidos estos dos grupos, una predicción de EvoMining es definida como aquellas hojas sin una categoría definida en el primer algoritmo y que están más cerca de una hoja azul que de una hoja roja. Es decir, son los homólogos que están en ramas del árbol que contienen genes descritos como reclutados por BGC y que no están en ramas de los homólogos con la función primaria. Estas predicciones consideradas con más posibilidades de pertenecer al metabolismo especializado que al conservado, son coloreadas en verde Figure 2.2 panel (c).

Además de los tres destinos metabólicos descritos marcados respectivamente en rojo, azul y verde, se puede opcionalmente agregar información predicha por antiSMASH. Cuando el usuario provee resultados de antiSMASH sobre qué genes pertenecen a un BGC que contiene una enzima típica de metabolismo especializado, estos se colorean en cian y son llamados predicciones antiSMASH. Si una secuencia es al mismo tiempo predicción EvoMining y predicción antiSMASH se colorea cian y se ignora el verde para enfatizar en los verdes las posibles novedades químicas. Cuando una secuencia está en la intersección de las reconocidas como metabolismo conservado marcada como roja y predicción de antiSMASH es decir color cian entonces es coloreada de púrpura. Estas enzimas de intersección entre metabolismo conservado y metabolismo especializado son definidas como enzimas de transición ya que se podrían pertenecer al metabolismo conservado, al especializado o a ambos. Además las enzimas de transición suelen estar en ramas intermedias entre ramas de metabolismo conservado y ramas de metabolismo especializado. Finalmente para todas las copias extra que no fueron marcadas como rojas, azules, verdes, cianes o púrpuras se les asigna el color gris. Así pues gris son aquellas hojas del árbol de las que no se tiene un clave sobre su destino metabólico, estas secuencias son llamadas de destino metabólico desconocido.

En este trabajo se diseñó e implementó el código para hacer el visualizador de EvoMining que permite hacer zoom en las ramas de interés además de que tiene links en las hojas azules que llevan al BGC en la página de MIBiG. También tiene la función de que cuando se seleccionan hojas de otro color se expande un recuadro que

muestra un mapa del contexto genómico de ese homólogo. Finalmente, este algoritmo también produce archivos de salida con otros metadatos como el número de copias por organismo y para que los árboles producidos por EvoMining sean compatible con la visualización de Microreact[154].

FALTA UNA FIGURA DONDE SE VEAN ALGUNAS CARACTERÍSTICAS DEL VISUALIZADOR O DE LA VISUALIZACIÓN EN MICRORREACT

### 2.3.3. Actualizaciones de las bases de datos de EvoMining

Tres bases de datos son requeridas como variables de inicio de EvoMining, la primera es el conjunto de secuencias de genomas de un linaje, esta base fue llamada Genome DB. La segunda es un grupo de secuencias de enzimas de metabolismo conservado llamada Enzyme DB. La base de datos se secuencias de genes que pertenecen a un cluster biosintético de metabolismo especializado es abreviada como base de datos de productos naturales o por sus siglas en inglés NP DB. Las transformaciones que sufrieron estas bases de datos desde la primera versión de EvoMining hasta este trabajo están resumidas en la tabla uno y serán descritas a continuación( Table 1 ).

Genome DB La primera versión tenía 230 genomas de Actinobacteria, incluyendo 50 géneros diferentes. Gracias a la explosión de datos genómicos disponibles, en EvoMining 2.0 ahora tiene 1245 genomas, incluyendo 193 géneros diferentes. Así pues, adicional a la actualización de Actinobacteria donde fue la primera vez que una prueba de concepto de EvoMining fue probada, tres nuevas bases de datos Genome DBs fueron integradas, incluyendo Cianobacteria (416 genomas), *Pseudomonas* (219 genomas) y Archaea (876 genomas). estas bases están disponibles en el repositorio de datos

público Zenodo con identificador [DOI 10.5281/zenodo.1219709](https://doi.org/10.5281/zenodo.1219709). Estos taxa fueron elegidos por su diversidad de exploración respecto a BGC, por ejemplo Actinobacteria posee 602 MIBiG BGC, Cianobacteria cuenta con 60 MIBiG BGC y *Pseudomonas* 53 MIBiG BGC. Estos tres taxa han sido ampliamente explorados experimentalmente y su riqueza metabólica está fuera de duda; en contraste Archaea sólo posee 1 BGC en la nueva versión MIBiG (v.1.4), y no había ninguno al tiempo de la realización de este trabajo (v.1.3). Por esta razón, incluir el dominio Archaea en los análisis permitía explorar espacios metabólicos previamente ignorados en la minería genómica.

Las predicciones de EvoMining se basan en identificar expansiones de familias de enzimas en lugar de buscar BGC completos, por esta razón los borradores de genomas con un promedio de al menos 5 genes por contig también pudieron ser incluidos en la base de datos Genome DB. Los genomas elegidos fueron recopilados de la base de datos pública NCBI tal y como estaba disponible en Enero de 2017. Las secuencias de DNA de estos genomas fueron anotadas como aminoácidos por la plataforma RAST[87] que a su vez realiza anotaciones funcionales basadas en la homología con otras secuencias con funciones descritas. Estos genomas, previo al análisis de EvoMining fueron minados por antiSMASH[152] con un parámetro `cf_threshold` de 0.7. Estos resultados fueron

suministrados como una base de datos interna, la antiSMASH DB para finalmente esta información ser incorporada a los árboles de EvoMining.

**Enzyme DB** La versión previa de la base de datos de EvoMining Enzyme DB comprendía 106 FEs, de metabolismo central de acuerdo a reconstrucciones metabólicas de los organismos *Streptomyces coelicolor*, *Mycobacterium tuberculosis* y *Corynebacterium glutamicum* [52]. Estos 106 EFs comprenden 339 secuencias de aminoácidos de Actinobacteria, que fueron usadas como secuencias semilla. En la versión actual, las 106 familias fueron filtradas hasta quedar sólo 42 que están presentes en Cianobacteria, *Pseudomonas* y Archaea. Durante el proceso de selección se escogieron genomas semilla que estuvieran contenidos en un sólo contig para evitar excluir familias debido a huecos debidos a problemas técnicos relacionados a la secuenciación o al ensamble de los genomas. Los genomas semillas son los proveedores de las secuencias semilla que conforman la base de datos Enzyme DB. Para cianobacteria, los genomas seleccionados son *Cianothece* sp. ATCC 51142, *Synechococcus* sp. PCC 7002 y *Synechocystis* sp. PCC 6803; para el género *Pseudomonas*, se escogieron *Pseudomonas fluorescens* pf0-1, *Pseudomonas protegens* Pf5, *Pseudomonas syringae* y *Pseudomonas fulva* 12-X; y para el dominio Archaea, los elegidos son *Natronomonas pharaonis*, *Methanosarcina acetivorans*, *Sulfolobus solfataricus* y *Nanoarchaeum equitans* Kin4-M. Las enzimas semilla que conforman la base Enzyme DB, fueron determinadas en los genomas semilla de cada linaje mediante BBH contra la base de datos de secuencias de metabolismo conservado original de EvoMining, la Actinobacteria Enzyme DB [52]. La herramienta Metaphor [77] fue implementada para obtener los BBH, se filtraron aquellas secuencias con menos del 30 % de identidad en un alineamiento del 80 % de la secuencia de las dos proteínas. Como resultado, las 106 familias de Actinobacteria quedaron reducidas a 42 FEs, compartidas por los genomas semilla de Actinobacteria, Cianobacteria, *Pseudomonas* y Archaea. Las bases de datos Enzyme DBs de todos los linajes están disponibles en Zenodo con número de identificación [DOI 10.5281/zenodo.1219709](https://doi.org/10.5281/zenodo.1219709)

**NP DB** Los primeros análisis realizados con EvoMining incluían un base de datos de productos naturales NP DB de 226 BGC reunidos de la literatura y curados manualmente [52]. En este trabajo la base NP DB que se utilizó para los análisis es MIBiG v1.3[105]. La base que viene incluida con el contenedor de EvoMining fue actualizada a la siguiente versión MIBiG v.1.4 liberada en Agosto de 2018. Esta nueva versión comprende 1813 NP BGC y un total de 31,023 secuencias de proteínas.

2.4. Detectamos patrones de neofuncionalización a BGC en distintos linajes y en familias enzimáticas con distintos orígenes metabólicos OR EvoMining detecta distintas dinámicas evolutivas de las enzimas metabólicas que dependen de la familia enzimática y del linaje taxonómico<sup>65</sup>

**2.4. Detectamos patrones de neofuncionalización a BGC en distintos linajes y en familias enzimáticas con distintos orígenes metabólicos OR**  
**EvoMining detecta distintas dinámicas evolutivas de las enzimas metabólicas que dependen de la familia enzimática y del linaje taxonómico.**

#### **2.4.1. Los perfiles de expansión de las proteínas dependen del linaje.**

Para entender la evolución de las enzimas y rutas metabólicas se seleccionaron cuatro linajes de diversas características los phyla Actinobacteria y Cianobacteria , el género *Pseudomonas* y el dominio Archaea. Todos los resultados en las figuras son presentados en este orden. Estos taxa fueron seleccionados para tener un espectro de análisis que abarcara tanto microorganismos ampliamente reconocidos como productores de NP, es decir Actinobacteria (602 MIBiG BGC), cianobacteria (60 MIBiG BGC) y *Pseudomonas* (53 MIBiG BGC); como también Archaea (0 BGC en MIBiG versión 1.3), que representa un dominio poco explorado en lo que respecta a los genes que forman parte de metabolismo especializado [108].

Basándonos en estas bases de datos de genomas blanco (Genome DB), como es explicado en la Figure 2.3 panel (a), un conjunto de familias enzimáticas comunes fue identificado. Notablemente, de las 106 familias actinobacteriales menos del 50 % están conservadas en los nuevos taxa. Cada base de datos, una para cada taxon, contiene sólo 42 FEs (Table S1). La observación de que 64 FEs no están conservadas en los cuatro taxa refleja lo específico del metabolismo en cada linaje con respecto a los otros [155].

Ya que la cantidad de expansiones de un gen en un organismo sería proporcional al tamaño de su genoma, verificamos cómo cambia el número de copias de los genes que pertenecen a las 42 FEs conservadas con respecto al tamaño de los genomas en los cuatro linajes. Todos los linajes tienen patrones de expansión similares en las 42 FEs analizadas hasta un tamaño de genoma 5 Mbp. En genomas más grandes, el número total de secuencias crece más en *Pseudomonas* que en el phylum Actinobacteria, que a su vez es más grande que el phylum Cianobacteria y que el dominio Archaea Figure 2.3 panel(b). Se observa cómo *Pseudomonas* es el linaje en el que su tamaño de genoma crece tanto como el número de copias de genes de familias muy conservadas (Glucólisis, síntesis de aminoácidos, Ciclo de Krebs, etc.). El resultado en Archea se debe a que no se han descubierto organismos tamaños de genoma Archaea de tamaño comparable a los de Streptomyces o *Pseudomonas* (> 5Mbp). Actinobacteria y Cianobacteria aunque tienen genomas > 5Mbp, el incremento en tamaño podría ser porque tienen

expansiones de genes que no son del metabolismo más conservado, porque obtienen más genes por transferencia horizontal u otros mecanismos por los que se incremente el tamaño del genoma. Cianobacteria a pesar de tener genomas grandes no tiene tantas expansiones en estas FEs. Esta observación, es posible que sea generalizable a todas las FEs de metabolismo conservado o bien que se deba a un sesgo en la selección de las familias que componen a las FEs.

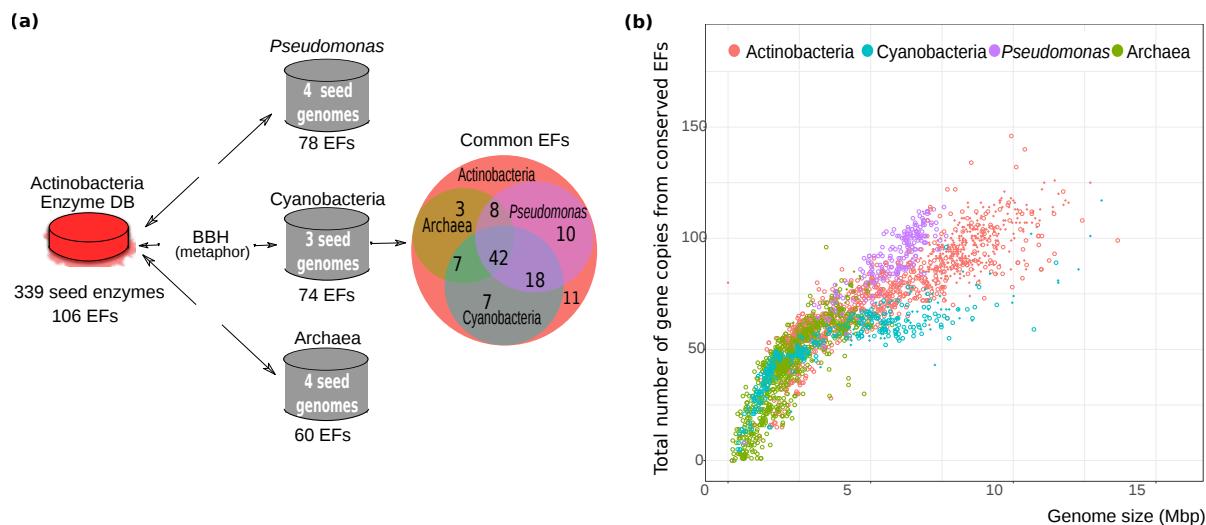


Figura 2.3:

Los órdenes con mayor número de copias fueron en las expansiones de las familias de la Enzyme DB fueron Streptomycetales y Nostocales , en Actinobacteria y Cianobacteria respectivamente. Esta observación es congruente con que estos órdenes tienen un tamaño de genoma grande en sus linajes correspondientes, y además están ampliamente representados en MIBiG como sintetizadores de productos naturales. Interesantemente la clase Halobacteria es la que muestra mayor número de expansiones en Archaea, aunque no es la clase con mayor tamaño de genoma en promedio (Fig. S10). [FIGURA tamaño de genoma]

Esta observación es congruente con que las archaeocinas, dicetopiperazinas, carotenoides y otros productos naturales de Archaea fueron aislados de especies de Halobacteria,

los genes, probablemente en BGC, que sintetizan de estos metabolitos no han sido caracterizados [156]. Por ello EvoMining una herramienta de minería genómica que puede ayudar a explorar linajes poco minados con el potencial de descubrir nuevas rutas metabólicas.

Es claro que en las familias conservadas seleccionadas el número de copias extra correlaciona con el tamaño de genoma los perfiles de expansiones son diferentes en cada grupo taxonómico y que este incremento parece cambiar su patrón en todos los linajes a partir de 5Mbp Figure 2.3 (2b). Por ello se concluye que para ensamblar una base de datos genómica para EvoMining se debe considerar que las expansiones dependen tanto de los distintos linajes taxonómicos como de la diversidad del tamaño de genoma.

#### 2.4.2. El shell genome posee expansiones en sus familias enzimáticas

La figura muestra que *Pseudomonas* posee en promedio más copias por genoma que los otros taxa. De las 42 familias analizadas el 54,8 % tiene su máximo número promedio de copias por genoma en este linaje. (FIGURA FIGURA ig. S11). En contraste, Actinobacteria es el máximo en 26,2 % de las FEs, mientras que Archaea y Cianobacteria empatan en ser el linaje con expansiones sólo en el 9,5 % de los casos (TABLATABLATABL A S1). Aunque existen familias como la acetylornitino aminotransferasa o la acetolactato sintasa (ALS) que están expandidas en todos los linajes (coordenadas A1 y E1, FIGURAFIGURAFIGURA S11), muchas otras familias exhiben expansiones sólo en ciertos linajes. Tal es el caso de la fumarato reductasa subunidad de hierro-azufre (coordinate C3, Fig. S11), muy expandida en Actinobacteria pero con menos de una copia por genoma en promedio en Cianobacteria.

No todas las FEs están expandidas, aunque las 42 familias conservadas están presentes en alguno de los genomas semilla de cada linaje, varias de ellas no se encuentran en la mayoría de los genomas del resto de su base de datos. Este es el caso de AroB en Archaea, donde tiene muy poca representación. Sin embargo, un gran porcentaje de familias muestra perfiles de expansión acordes a la tendencia del número total de copias extra, mostrada en la figura FIGURAFIGURAFIGURA. Por familia, *Pseudomonas* suele ser el linaje con el mayor número de expansiones mientras que Archaea suele ser el linaje menos expandido. Entre las excepciones a esta tendencia está GDH una familia incluida dentro de los ocho casos seleccionados. Para ilustrar esta diversidad que son mostrados en la figura Figure 2.4

panel (a). En esta figura los máximos están marcados con un círculo mientras que los mínimos con un triángulo, los colores de estas formas geométricas representan los mismos linajes que los mostrados en la Figure 2.3. Del total de las 42 familias, GDH es una de las cuatro FEs en las que el mayor número de expansiones se encuentra en Archaea. De hecho, GDH tiene menos de una copia por genoma en promedio en los otros taxa, probando que no se encuentra dentro del core genómico de estos linajes.

Esto contrasta con AroB, que muestra una tendencia opuesta, no es parte del core de Archaea pero muestra copias extra y una presencia mayor que uno en promedio en los otros tres taxa analizados Figure 2.4 panel (b). Los ocho casos mostrados en la figura son todos parte de un cluster biosintético de cianobacteria, descrito en las secciones posteriores de este trabajo.

Con estas observaciones sospechamos que GDH es miembro del shell genome (Ver capítulo 1) [157] en los taxa Actinobacteria , Cianobacteria y *Pseudomonas*, ya que en promedio está cerca de tener una copia promedio por genoma. El promedio no es suficiente para decir que una familia pertenece al shell, podrían suceder casos sobre todo cuando hay mucha variación en un taxon como en el caso de un domino o un phylum en contraposición con taxones conservados como géneros en los que para una cierta familia la mitad de los genomas de un linaje tuviera dos copias y la otra mitad cero. Sin embargo en el caso de la GDH si es consistente en que está presente en más del 50 % de los genomas de cada linaje Figure 2.4 panel (a). Las modas de número de copias también son informativas por ello son mostradas en la figura FIGURAFIGURAFIGURA. Una sola copia extra puede ser la que sea reclutada en metabolismo especializado.

En la figura FIGURAFIGURAFIGURA se muestra un ejemplo donde AroB es una enzima core en oposición a GDH que es una enzima shell. En este esquema conceptual, en algunos de los genomas que contienen a AroB existen copias que se dedican al metabolismo especializado marcadas en color cian, otras copias no tienen un destino metabólico conocido por lo que están marcadas en gris, y otras más marcadas en púrpura son enzimas de transición que están llevando a cabo simultáneamente una función en metabolismo central y otra en metabolismo especializado. En contraste a AroB se muestra GDH, que a pesar de no tener copias en algunos genomas y con un promedio de copias por genoma menor a uno y una moda de uno en esta muestra, GDH existe por duplicado en dos genomas. En uno de esos genomas donde GDH tiene una copia extra, más allá de la moda, esa copia se muestra como un reclutamiento al metabolismo especializado en cian. Figure 2.4 panel (b) .

Las expansiones encontradas por EvoMining en la familia GDH incluían predicciones de antiSMASH para Actinobacteria, Cianobacteria y Archaea, no así para *Pseudomonas*. La secuencia del reclutamiento de GDH por los clusters biosintéticos escitonemina y el policétido pactamycin [158], es suficientemente parecida como para que EvoMining la detecte como expansión en Actinobacteria, Cianobacteria y Archaea, pero es tan divergente respecto a la familia GDH en *Pseudomonas* que EvoMining lo deja fuera de la familia expandida en este linaje. Los árboles donde puede apreciarse esta observación pueden consultarse más adelante en la Figure 2.5 de la siguiente sección.

Los resultados anteriores sugieren que la evolución del metabolismo especializado es linaje dependiente, y más aún que tal y como ya se conocía en las FEs del core genome las enzimas del shell como GDH también poseen el potencial de ser reclutadas en NP BGC. A partir de estos resultados Figure 2.4 paneles (a,b), se realizó un esquema conceptual para explicar cómo en linajes genómicos diversos las familias de enzimas con origen en el metabolismo central que forman parte del core genome o bien las

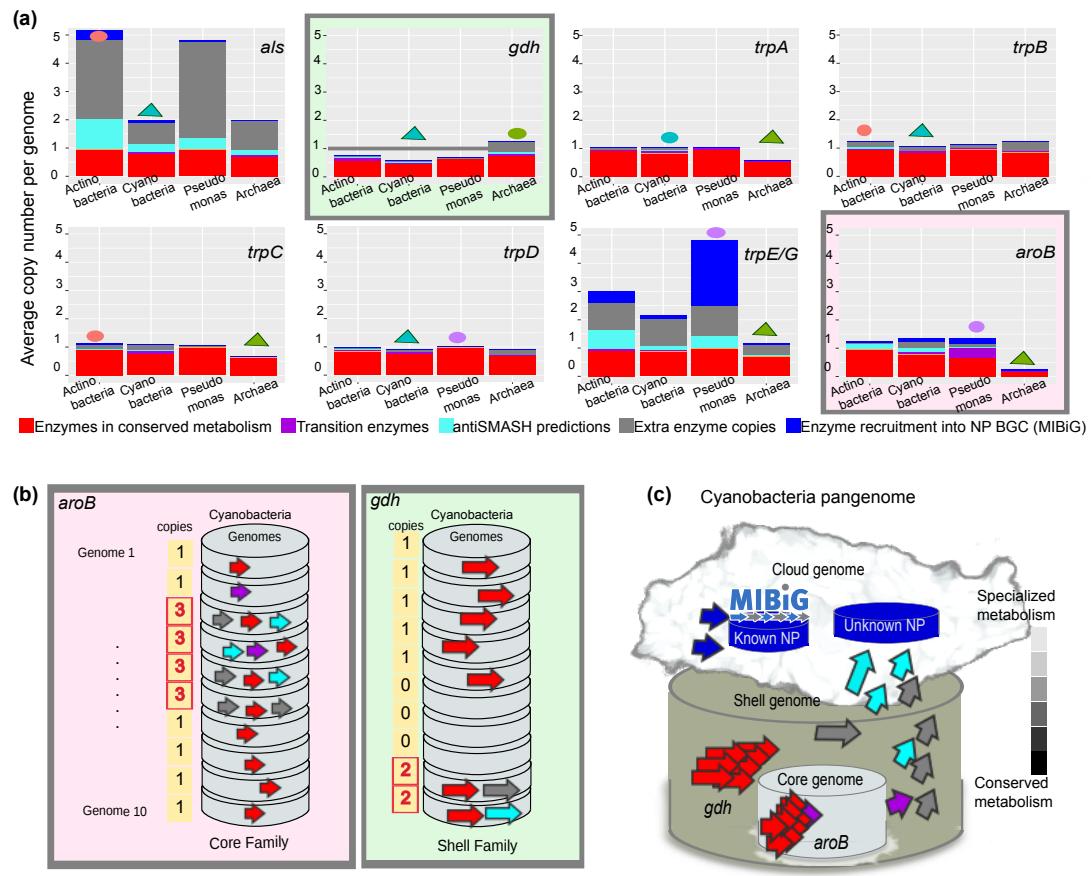


Figura 2.4: Perfiles de expansiones de EvoMining de enzimas conservadas seleccionadas. (a) Patrones de expansión de las ocho familias conservadas cuyas copias adicionales participan en la biosíntesis de escitonenmina. El conjunto completo de 42 EF se muestra en la Fig. S11. La codificación de colores es la siguiente: rojo para el metabolismo conservado, azul para los reclutamientos anotados en MIBiG, cian para las predicciones antiSMASH de metabolismo especializado, púrpura para la intersección entre el metabolismo conservado y las predicciones antiSMASH, y gris para las expansiones sin destino metabólico conocido. El orden en el eje x es Actinobacteria, Cianobacteria, Pseudomonas y Archaea. Los triángulos indican el linaje con el mayor número de copias por genoma en promedio, y los círculos representan el linaje menos expandido. Aunque Archaea tiende a ser los taxones menos expandidos, esta tendencia se revierte en la familia GDH. (b) Se proporciona un ejemplo de un núcleo frente a un shell EF. AroB es un EF básico porque tiene al menos una copia por genoma, mientras que GDH es un EF de shell debido a su ausencia en tres genomas. A pesar de ser un shell EF, GDH tiene copias adicionales que pueden ser reclutadas en un metabolismo especializado. (c) Modelo para la nube o genoma variable compuesto parcialmente por enzimas que pertenecen a BGC NP. En este modelo, el metabolismo conservado se compone de EF tanto de shell como de núcleo. Estos EF pueden sufrir eventos de expansión, y algunas de las copias adicionales son reclutadas para realizar nuevas funciones en el metabolismo especializado.

familias en el metabolismo conservado que incluye tanto al core como al shell genome, evolucionan al metabolismo especializado que tiene una mayor representación en el cloud genome Figure 2.4 panel (c) . Este modelo es relevante porque establece el papel de las familias del shell genome, que no fue considerado en la primera iteración que explotó las capacidades de EvoMining como herramienta de minería genómica para encontrar BGC novedosos[159].

En la siguiente sección se analizarán los patrones de expansión reclutamiento de GDH provistas por EvoMining for GDH y se describirán los árboles filogenéticos de cada linaje mostrados en la Figure 2.5 , así como un árbol que incluye conjuntamente secuencias de todos los linajes (Figs S12 and S13). En Archaea, GDH tiene en promedio 1.23 copias por genoma, mientras en Actinobacteria, cianobacteria y Pseudomonas esta media es de 0.74, 0.56 y 0.65, respectivamente. En estos tres taxa GDH es parte del shell genome (Table S2). Además de GDH se estudiaron las expansiones y los árboles filogenéticos de TrpA, TrpB, TrpC, TrpD, TrpEG, AroB y ALS, todas ellas parte de las 42 FEs conservadas entre los cuatro linajes y a la vez reclutadas en escitonemina [148] un cluster biosintético de cianobacteria.

#### **2.4.3. GDH y ALS en el cluster escitonemina ejemplifican como familias pertenecientes a un mismo BGC pueden tener distintos patrones de expansión.**

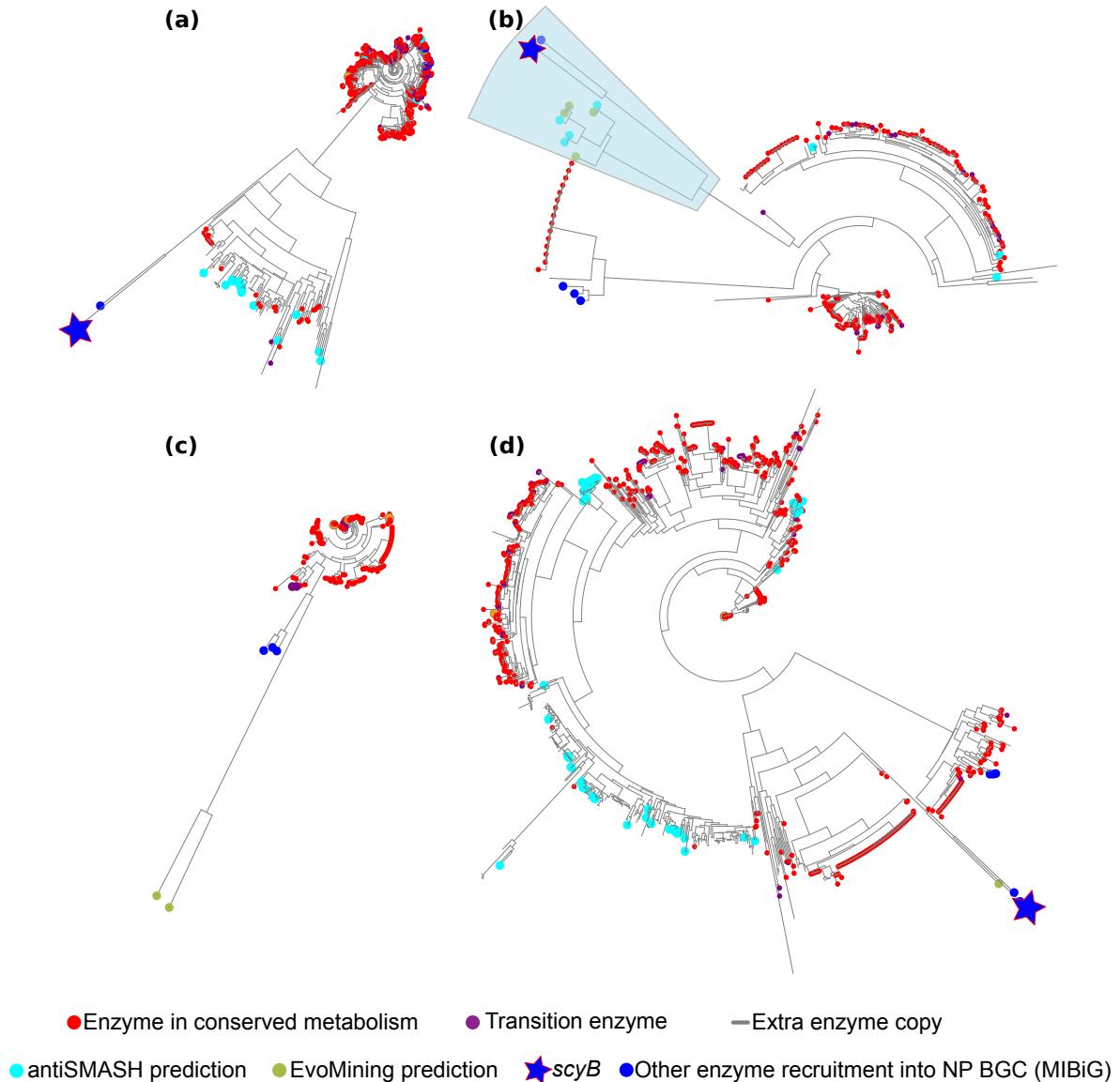
La enzima GDH, se encuentra presente en muchos linajes debido tanto a su origen ancestral como a la transferencia horizontal [ 28, 42 ] (Fig. S12). GDH cataliza la reacción reversible de desaminación oxidativa de glutamato en  $\alpha$ -cetoglutarato y amonio. De acuerdo al uso de cofactores, la GDH puede dividirse en tres clases, la primera usa NAD<sup>+</sup> y es nombrada como GDH(NAD<sup>+</sup>). La segunda clase utiliza NADP<sup>+</sup> y es conocida como GDH(NADP<sup>+</sup>). La tercera clase utiliza ambos cofactores NAD<sup>+</sup> y NADP<sup>+</sup>; por lo que se le conoce como GDH(NAD<sup>+</sup> y NADP<sup>+</sup>)[150].

Aunque existen otras clasificaciones de la diversidad de enzimas GDH esta fue seleccionada porque se relaciona con la historia evolutiva de la enzima. GDH(NAD<sup>+</sup>) es utilizada para la oxidación del glutamato mientras que GDH(NADP<sup>+</sup>) para fijar amonio, algunas enzimas de Archaea funcionan bien con ambos cofactores, es decir tienen promiscuidad de cofactores[150]. La especificidad por NAD<sup>+</sup> o NADP<sup>+</sup> probablemente emergió en repetidas ocasiones, una evidencia a favor de esta hipótesis es que se ha mostrado que algunas mutaciones pueden revertir la especificidad[160]. Esto sugiere que en los cofactores análogamente al caso de promiscuidad por sustrato, la similitud de secuencia no siempre es suficiente para evidenciar la especificidad. En ocasiones la divergencia o cercanía filogenética de los organismos productores de la enzima es una información adicional a la similitud de secuencia, esta consideración es importante al analizar enzimas en linajes muy divergentes.

La familia GDH muestra expansiones aunque no muy abundantes en Actinobacteria Figure 2.5 panel (a) ,y en cianobacteria Figure 2.5 panel(b). Las expansiones están

2.4. Detectamos patrones de neofuncionalización a BGC en distintos linajes y en familias enzimáticas con distintos orígenes metabólicos OR EvoMining detecta distintas dinámicas evolutivas de las enzimas metabólicas que dependen de la familia enzimática y del linaje taxonómico.

71



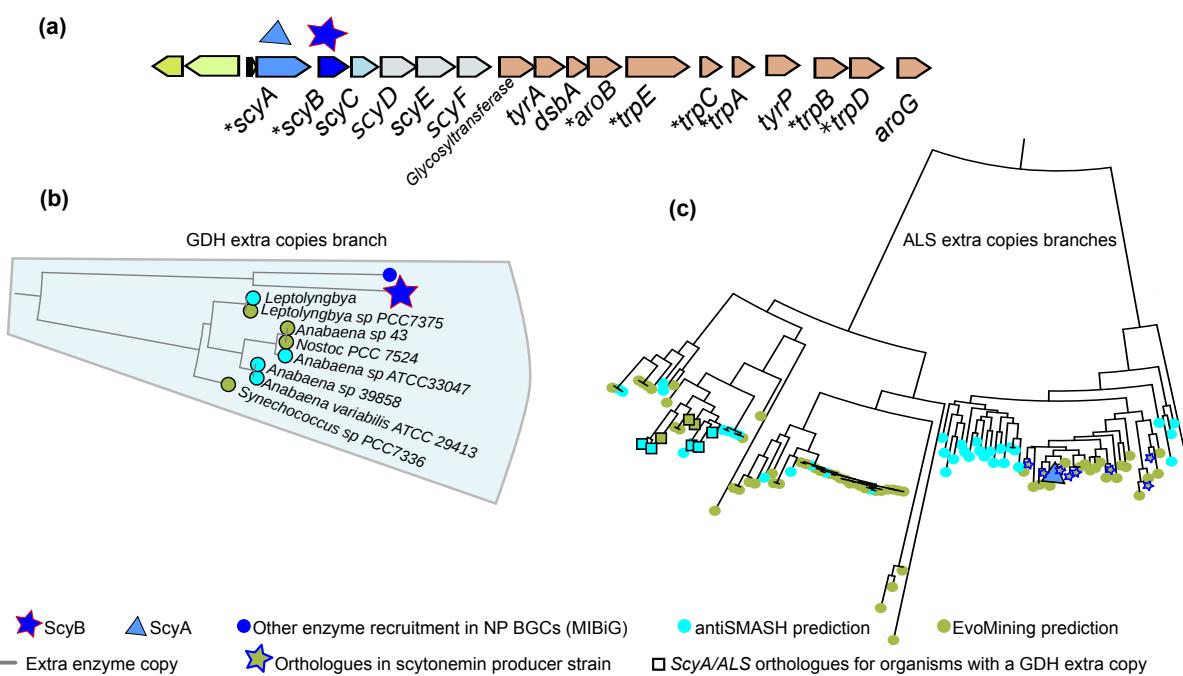


Figura 2.6:

prácticamente ausentes en *Pseudomonas* Figure 2.5 panel (c). En contraste, un número significativo de expansiones es encontrado en Archaea Figure 2.5 panel(d) . El árbol de EvoMining de la familia GDH en Archaea fue enraizado con la secuencia semilla de *Sulfolobus*, que fue predicha por RAST como una enzima dual en el uso de cofactores NAD(P)+[161]. En Archaea las tres clases de GDH alternan en las ramas del árbol (Fig. S13,).

Muchas de las secuencias clasificadas como de metabolismo conservado se concentran volviendo rojas las ramas basales del árbol. Se observa otro clado más grande y diverso compuesto casi exclusivamente por enzimas específicas para NAD(P)[162], incluyendo muchas predicciones de antiSMASH, y sólo dos marcadas como metabolismo conservado. Estas dos marcas pueden deberse a la pérdida real de una enzima d metabolismo central en las ramas centrales o bien a huecos debidos a la calidad del ensamblado y la secuenciación de los genomas. La anotación funcional de estos ortólogos de GDH apunta hacia reclutamientos en el metabolismo especializado. Estos reclutamientos fueron identificados en organismos de los genera *Haladaptatus*, *Haloterrigena* , *Natrialba* , *Natrinema* , *Natriabaceae* y *Natronococcus*. Los genes se encuentran un contexto de posible síntesis de terpenos. Este contexto incluye enzimas relacionadas al geranil pirofosfato, un precursor de todos los terpenos y terpenoides[163]. Este árbol tiene también casos de divergencia reciente. Hay una pequeña rama indistinguible en la figura pero explorable en la plataforma microreact donde los parálogos aparecen junto a las secuencias de metabolismo central. Finalmente a pesar de la divergencia las últimas ramas corresponden a enzimas de metabolismo conservado, es decir son las copias más parecidas en esos organismos a las semillas provistas en la Enzyme DB Figure 2.5 panel(d).

En contraste con la amplia expansión de GDH relacionada a las adaptaciones metabólicas en Archaea, el árbol de Cianobacteria tiene copias extra sólo en el 4,5% de sus genomas ( Figure 2.5 panel(b) , Table S2). En esta rama expandida se encontraron cuatro predicciones de antiSMASH y cuatro predicciones de EvoMining en la rama que contiene a ScyB el homólogo de GDH que fue reclutado por el BGC escitonemina. ScyB es pues parte de la síntesis de escitonemina, un pigmento amarillo producido por muchas cianobacterias como protección contra la radiación UV-A solar[164]. *Nostoc punctiforme* PCC 73102 es el organismo productor de escitonemina cuyo BGC fue caracterizado y anotado en MIBiG. EvoMining sólo unas pocas secuencias GDH copias extra de especies de Nostoc aún cuando se conoce que homólogos de scyB pueden encontrarse en estos genomas. Esta observación puede deberse a la gran divergencia de secuencia entre copias de metabolismo central y de metabolismo especializado en estos organismos.

En la vecindad genómica de algunas expansiones de GDH se observó la secuencia de ALS, un gen identificado en la literatura como homólogo de *scyA*. además, en los BGC conocidos de escitonemina se observó que *scyB* se conserva cerca del gene *scyA* Figure 2.6 panel (a). *scyA* es homólogo de la subunidad larga de ALS. Esta familia tiene un número promedio de copias de 1,87 % en la base de datos Genome DB de cianobacteria. La media es de hecho de 2,1 copias en organismos que contienen al

menos una copia ALS, pero la moda del número de copias es 1 (Table S2). Estos datos indican que muchos organismos tienen más de dos copias de ALS lo que puede correlacionar con que esta familia es más dispersa alrededor de la moda (Fig. S4, blue line).

Al generar el árbol de EvoMining de ALS en cianobacteria , Fig. S11, se observó que *scyA* es un reclutamiento que se localiza en una rama repleta de secuencias de ALS provenientes de *Nostoc spp.*, que fueron etiquetadas como predicciones de EvoMining Figure 2.6 (c). Estas predicciones incluyen más de veinte organismos conocidos como productores de escitonemina [148]. Además, ramas cercanas muestran secuencias de ALS que son predicciones de antiSMASH, reforzando la sugerencia de que esta sección del árbol se dedica al metabolismo especializado. Una última rama contiene a los mismos organismos encontrados en el árbol de EvoMining de la familia GDH. Estos organismos son mostrados en el zoom de la rama de *scyB* Figure 2.6 panel(b).

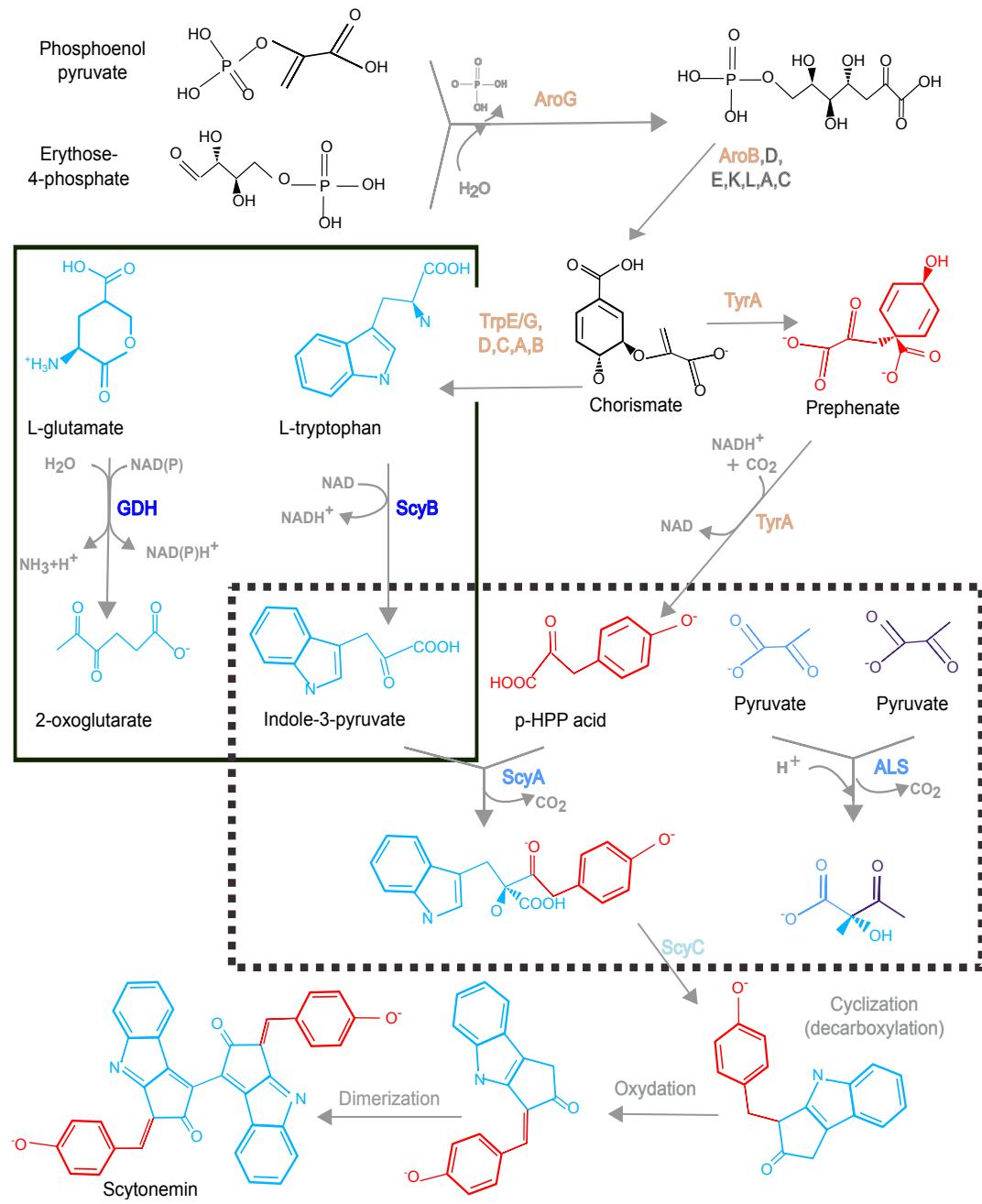
Esta observación sugiere co-diversificación, vía un evento de expansión reclutamiento de ScyA y ScyB a partir de su origen en ALS y GDH, respectivamente. Los perfiles de expansión de estas familias difieren ya que a diferencia de la muy poblada rama de *scyA* en el árbol de ALS, la similitud entre homólogos de la FE de GDH y homólogos de ScyB no fue suficiente para reconstruir una rama de *scyB* con todas las expansiones sugeridas por la ocurrencia del cluster de escitonemina. Estas observaciones son una lección a usar EvoMining como herramienta de minería genómica, que enzimas cercanas pueden co-diversificarse en ocasiones formando parte de un mismo BGC pero a la vez estar sujetas a distintas restricciones evolutivas.

#### 2.4.4. El cluster de escitonemina es un cluster promiscuo.

Definimos *cluster promiscuo* como un BGC que sintetiza distintos productos naturales con la misma estructura base y con algunas modificaciones que son específicas de cada organismo que los contiene o incluso aquellos BGC que en un mismo organismo pueden producir varias moléculas muy similares. El *cluster* de la escitonemina mostrado en la figura Figure 2.6 panel (a) y en Figure 2.7 comprende 18 genes [149]. Además de genes reguladores, este BGC incluye a los genes genes biosintéticos scyABC, los genes conservados con función desconocida scyDEF y los proveedores de precursores: *tyrA*, *dsbA*, *aroB*, *trpE/G*, *trpC*, *trpA*, *tyrP*, *trpB*, *trpD*, *aroG*. Las familias enzimáticas TrpABCDEG y AroB son parte de las rutas de los aminoácidos aromáticos y del ácido shikímico, parecen haber sido reclutadas para proveer de los precursores L-triptofano y prefrenato, que son necesarios para la síntesis de escitonemina. En oposición a los genes del operon *trp* y a AroB que siguen realizando su función de metabolismo conservado aún al ser parte de una ruta de metabolismo especializado están ScyA y ScyB. Estas dos familias también tienen un origen en el metabolismo conservado, ya se ha explicado que tienen su origen en las familias ALS y GDH, pero en este caso sí ha cambiado la especificidad por sustrato al momento de la incorporación

2.4. Detectamos patrones de neofuncionalización a BGC en distintos linajes y en familias enzimáticas con distintos orígenes metabólicos OR EvoMining detecta distintas dinámicas evolutivas de las enzimas metabólicas que dependen de la familia enzimática y del linaje taxonómico.

75



al metabolismo especializado Figure 2.7. ALS une dos piruvatos, transformándolos en S-2-acetolactato [151], mientras que ScyB cataliza la unión de indol-3-piruvato con ácido p-hidroxi-fenil pirúvico. De forma análoga, GDH convierte L-glutamato en 2-oxoglutarato [150], mientras que ScyA cataliza una desaminación oxidativa de triptofano. El producto de estas dos enzimas actuando secuencialmente e un dipéptido, el cual es ciclado por ScyC. La ruta metabólica culmina con una serie de oxidaciones y dimerizaciones hasta llegar al producto escitonemina, aún no es claro como se llevan a cabo estos últimos pasos[148].

Además de GDH y ALS el BGC de escitonemina tiene seis familias que son parte de las 42 FEs analizadas aquí, es decir 8 de los genes que participan en la síntesis de escitonemina tienen un origen en metabolismo conservado y fueron reclutados en el cluster de escitonemina ( Figure 2.4 panel 1, Table S3). ( Figure 2.6 and S3–S9). De estas familias, seis de los siete árboles de EvoMining contienen copias extras identificadas como predicciones de EvoMining, debido a que en su rama de expansión se encuentra el correspondiente gen de escitonemina del cluster reportado en MIBiG. Estas familias incluyen AroB y todos los genes de la ruta del L-triptofano excepto *trpF*. Además los árboles de EvoMining de estas familias incluyen predicciones en ramas que son enzimas provenientes de las rutas de síntesis de otros pigmentos protectores solares como la shinorina y los aminoácidos de tipo mycosporina Figure 2.9 [164] , además de otros productos no relacionados como la welwitindolinona [165], ambiguina [166] y la fischerindolina [167] (Figs S3–S7). Estos resultados ilustran como EvoMining puede complementar a antiSMASH mediante la identificación de secuencias de BGC no tradicionales.

Para investigar la coocurrencia de ScyA y ScyB, que son necesarias juntas para producir escitonemina, reconstruimos su historia evolutiva conjunta. Para ello se obtuvieron las secuencias de lo genomas donde ambas tenían presencia, se concatenaron sus secuencias y se realizó una filogenia con ellas. Las variantes de la vecindad genómica del BGC de escitonemina fueron visualizadas mediante el uso de CORASON [159]. En el siguiente capítulo será explicado con más detalle este software de visualización y organización de vecindades genómicas. Los análisis filogenómicos resultaron en 34 cianobacterias con diversidad química en el BGC de escitonemina. Es decir en conclusión escitonemina es un ejemplo de cluster promiscuo, ya que parece haber un core conservado, pero diversidad en enzimas accesorias y por tanto en sus productos finales.

Se pudieron predecir cinco estructuras putativas que son variantes de escitonemina y correlacionan con episodios de pérdida y ganancia de genes en este locus Figure 2.8. Al core de genes *scyABCD* se incorporan genes que realizan ornamentos como hidrolasas, prenil-transferasas , fosfodiesterasas y monooxigenasas , para formar congéneres de escitonemina como los compuestos 1 y 2. La pérdida de los genes *scyDEF* y la aparición de otras enzimas como la tyrosinasa y o la amidasa pueden derivar en la síntesis de los compuestos 3 and 4. Además encontramos que homólogos de *scyA* y *scyB* son parte de otro BGC que contiene un híbrido NRPS-PKS. Siguiendo las reglas biosintéticas de estas enzimas se propuso el compuesto 5. La diversidad química sugerida en estas predicciones sólo puede ser validado mediante el trabajo experimental.

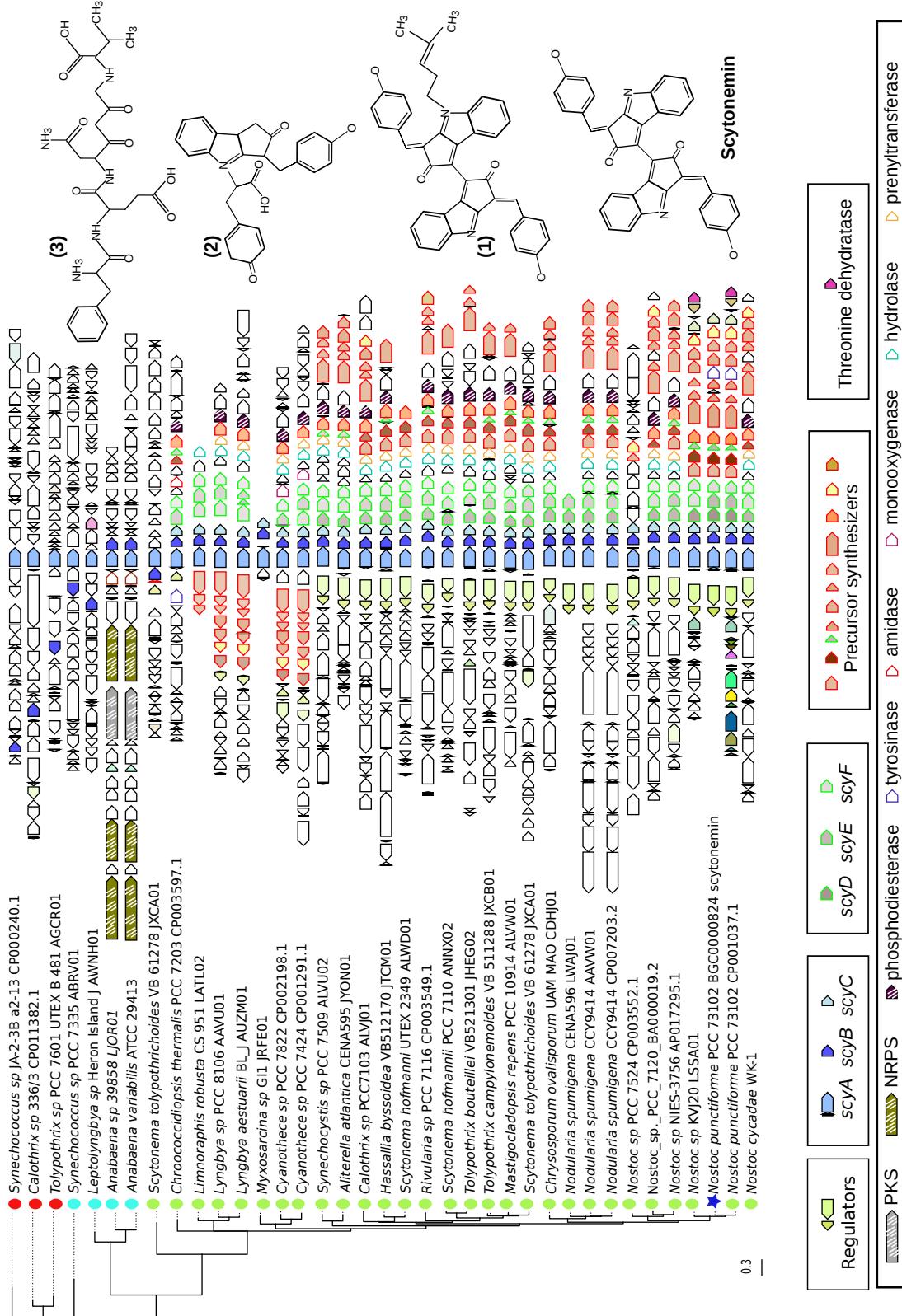


Figura 2.8:

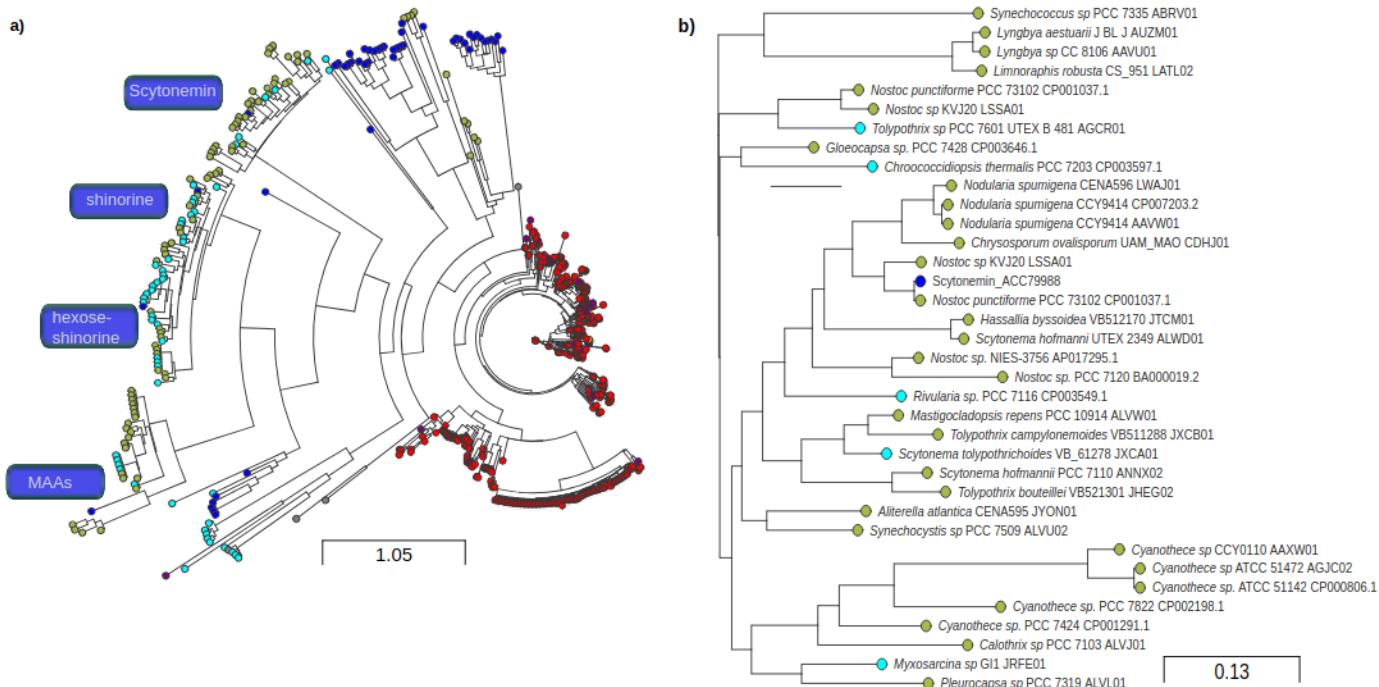


Figura 2.9:

Las variantes producidas por la dinámica evolutiva del metabolismo especializado fueron sugeridas mediante el sólo uso de ScyA y ScyB como semillas de búsqueda. Estos resultados sugieren el poder predictivo de EvoMining para explorar espacios metabólicos típicamente ignorados por métodos de búsqueda tradicionales de BGC que no consideran la evolución dentro de sus algoritmos de minería genómica.

#### 2.4.5. EvoMining en RNA transferasas

Azoxo

#### 2.4.6. La familia *tauD* tiene expansión y reclutamiento tanto en Actinobacteria como en Pseudomonas

TauD es una enzima dedicada al metabolismo de taurina, proviene del operón de *E. coli*. En Pseudomonas también tiene muchas expansiones. Es parte de los *clusters* rimosamide y detoxin, así como de 15 BGC más. Tiene en Actinobacteria una gran rama dedicada al metabolismo especializado, es de una ruta biosintética promiscua. La ruta biosintética será tratado en el siguiente capítulo.

AQUI ARBOL EVOMINING TAUD y GRAFICA CON tauD

#### 2.4.7. Rimosamide y Detoxin tienen en común la enzima TauD,

Un análisis de EvoMining de las expansiones de la familia TauD dioxygenasa mostró que existe una rama dedicada al metabolismo especializado. Dentro de esta rama existen paralogos dentro de géneros como *Streptomyces*, *Rhodococcus*, *Frankia* y *Amycolatopsis* (Fig S13). Encontramos un clado dentro de las expansiones de la familia que contiene quince homólogos de *tauD* que pertenecen a BGC experimentalmente caracterizados y depositados en MIBiG, incluyendo los de la de detoxin y rimosamidas (Table S6). La variedad de BGC mostrada en este clado abre la posibilidad de encontrar variantes moleculares de estas familias.

En Actinobacteria se han reportado dos *clusters* biosintéticos *Streptomyces*, *Rhodococcus*, *Frankia* and *Amycolatopsis*

Cuadro 2.1: Homólogos de *tauD* en BGC reportados en MIBiG.

MIBiG BGC	Compound	Class	Producer Organism
653_ADO85576	pentalenolactone	Terpene	<i>Streptomyces arenae</i>
678_BAC70706	pentalenolactone	Terpene	<i>Streptomyces avermitilis</i>
			NBRC 14893
163_ACR50790	tetronasin	Polyketide	<i>Streptomyces longisporoflavus</i>
961_ABC36162	bactobolin	NRP- Polyketide	<i>Burkholderia thailandensis</i>
287_AAG05698	2-amino-4-methoxy-trans-3- butenoic acid	NRP	<i>Pseudomonas aeruginosa</i> PAO1
846_ctg1_orf9	tabtoxin	Other	<i>Pseudomonas syringae</i>
1183_AGC09526	lobophorin	Polyketide	<i>Streptomyces</i> sp. FXJ7.023
1156_ADD83004	platencin	Terpene	<i>Streptomyces platensis</i>
1140_ACO31277	platensimycin-platencin	Terpene	<i>Streptomyces platensis</i>
1140_ACO31282	platensimycin-platencin	Terpene	<i>Streptomyces platensis</i>
715_ABW87795	spectinomycin	Saccharide	<i>Streptomyces spectabilis</i>
1205_KGO40485	communesin	Polyketide	<i>Penicillium expansum</i>
1205_KGO40482	communesin	Polyketide	<i>Penicillium expansum</i>
1183_AGC09525	lobophorin	Polyketide	<i>Streptomyces</i> sp. FXJ7.023
654_ABB69741	phenalinolactone	Saccharide-Terpene	<i>Streptomyces</i> sp. Tu6071
1070_CAN89617	kirromycin	NRP - Polyketide	<i>Streptomyces collinus</i> Tu 365

La tabla Table 2.1 muestra los reclutamientos de *tauD*.

## 2.5. Consideraciones finales sobre el uso de Evo-Mining

EvoMining fue desarrollado como una herramienta de minería genómica descargable que puede ser aplicada a bases de datos de secuencias de metabolismo conservado (Enzyme DB) provenientes de FEs de distintos phyla. Nuestros análisis llevaron a la conclusión de que los patrones de expansión reclutamiento dependen tanto de la familia enzimática como del linaje genómico en el que se analiza. Una consideración importante al usar EvoMining es que el tamaño de genoma correlaciona con el número de copias extra de familias expandidas. Aunque el tamaño de genoma es importante, también encontramos excepciones donde EvoMining pudo predecir enzimas de BGC no tradicionales en genomas relativamente pequeños, sugiriendo que hacen falta más análisis para estudiar esta relación. En este sentido, optamos por comparar linajes genómicos que no solo son altamente divergentes y, en algunos casos, poco conocidos con respecto a la biosíntesis de NP, sino también desproporcionados en cuanto a su resolución taxonómica y distancias. Por lo tanto, es posible que estos factores hayan impuesto un sesgo al establecer relaciones entre el tamaño del genoma, la tasa de expansión de genes y la diversidad metabólica.

Es interesante observar que las familias más expandidas según los análisis de prueba de concepto anteriores de EvoMining [52] fueron asparagina sintasa, 2-dehidro-3-deoxifosfoheptanoato aldolasa y 3-fosfositomato-1-carboxivinil transferasa, que llevaron al descubrimiento de enzimas biosintéticas de arsenolípidos. Cabe destacar que ninguna de estas enzimas formaba parte de los 42 FE analizados en el presente documento, lo que refuerza la idea de que no solo las enzimas conservadas, sino también las enzimas del *shell* con copias adicionales, pueden servir como semillas para el descubrimiento de nuevos BGC. Después de observar que la familia de la GDH tiene numerosas expansiones en Archaea pero no en otros taxones, proporcionamos un ejemplo de un reclutamiento de una enzima metabólica central por un BGC en Cianobacteria, donde no hay tantas expansiones, lo que sugiere que las predicciones en Arhea deben ser exploradas experimentalmente. Estas observaciones enfatizan la naturaleza predictiva de EvoMining.

En este punto conviene enfatizar cómo fue clave el papel de MIBiG para esta versión de EvoMining [105] ya que permite incrementar consistentemente y sin esfuerzo de curación manual a los BGC reportados por investigadores de todo el mundo. La versión previa de EvoMining no incluía por ejemplo BGC de Cianobacteria o de Archaea. Gracias a esta actualización que la nueva versión de EvoMining pudo identificar correctamente secuencias de genes del BGC de la escitonemina (ver abajo). Sin la presencia de la señal de los BGC de cianobacteria en MIBiG estas hojas habrían sido catalogadas como de destino metabólico desconocido.

En este capítulo se describió el funcionamiento de EvoMining y sus aplicaciones como una herramienta de minería genómica que permite relacionar la historia evolutiva de familias enzimáticas con su función. Específicamente, lo empleamos para señalar familias de genes con expansiones que seguramente tienen elementos con promiscuidad y para ubicar a los homólogos que probablemente forman parte de BGC. Estos últimos seguramente producen nuevos productos naturales, razón por la que mejoramos la predicción BGC a partir de estas enzimas en el siguiente capítulo. Se mostró que el metabolismo se puede expandir a partir de genes de metabolismo conservado o de genes de metabolismo *shell*, que en un BGC se encuentran genes provenientes de la expansión de distintas familias de enzimas, que una misma expansión puede ser utilizada para producir distintos análogos con estructuras similares y que los patrones de expansión son específicos de cada familia y linaje. Finalmente, se describió el descubrimiento de 3 nuevos BGC cuyo producto natural ya ha sido demostrado por colaboradores. Queda por explorar los patrones de neofuncionalización de forma exhaustiva con genes que provengan no sólo de metabolismo conservado sino también del *shell* así como el determinar qué familias del metabolismo conservado dieron origen a los BGC conocidos.



## Capítulo 3

# Desarrollo de CORASON como herramienta para organizar clusters biosintéticos y otras vecindades genómicas conservadas.

CORASON es una herramienta en línea de comandos para organizar filogenéticamente la variación de una familia de clusters, Figura Figure 3.1. Dada la existencia de variantes de BGCs en bases de datos de linajes genómicos se diseñó CORASON. Corason se ejecuta en línea de comandos, su función es identificar el core génico de un BGC y utilizarlo para presentar una visualización de las variantes de la familia organizadas filogenéticamente

Los métodos de minado de genomas han acelerado el descubrimiento de nuevos clusters biosintéticos (BGCs). Casi cada nueva bacteria que es secuenciada aporta alguna novedad al pangenoma bacteriano conocido. Una fracción de estos genes novedosos formará parte de variantes de BGCs previamente conocidos aportando diversidad a las familias de clusters biosintéticos. La diversidad genética que existe en las familias de BGCs está directamente relacionada con cambios moleculares, incluso pequeñas variantes en un metabolito pueden ocasionar diferencias en su función biológica.

CORASON fue diseñado con las siguientes características: (i) una interfaz de línea de comando simple (ii) identificación del core génico del BGC (iii) la reconstrucción filogenética de la familia de BGCs utilizando la información del core (iv) salida visual en formato SVG, que muestra tanto la anotación funcional de los genes como la distancia respecto a sus ortólogos del cluster de referencia.

En este capítulo presento CORASON, CORe Analysis of Syntenic Orthologs to prioritize Natural products biosynthetic gene clusters, una herramienta para explorar la diversidad en el contenido de los BGCs así como su distribución en un linaje genómico proporcionado por el usuario.

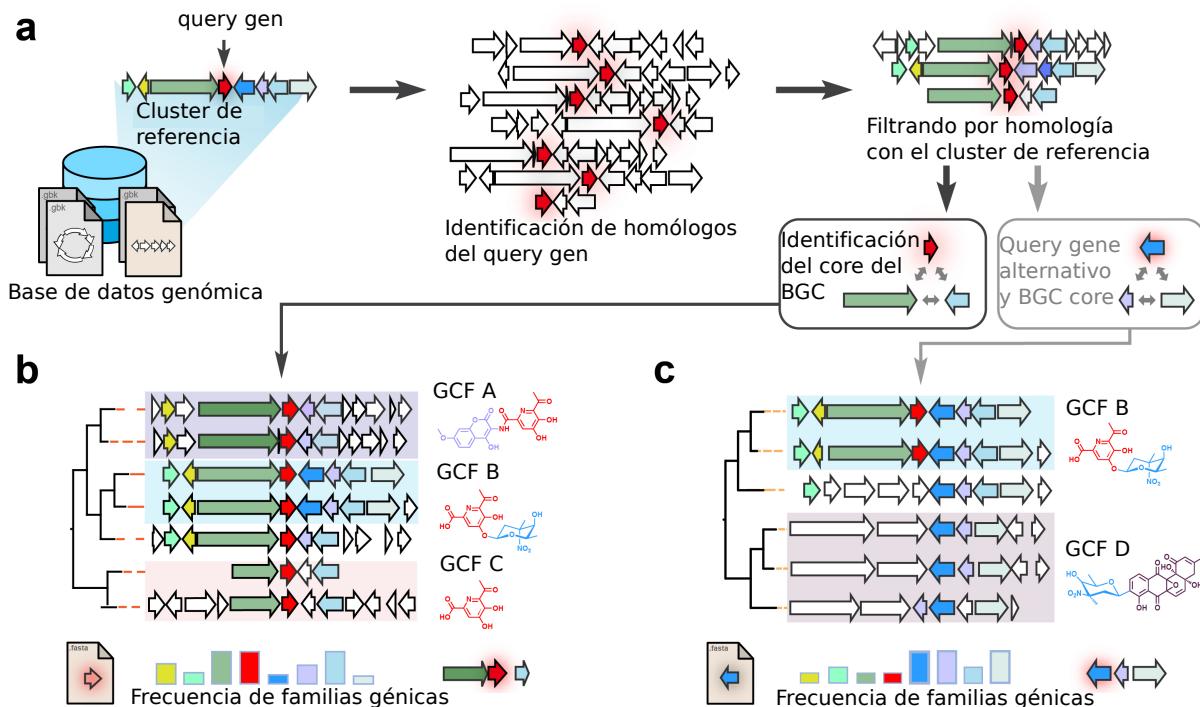


Figura 3.1: La herramienta corason localiza familias de clusters biosintéticos en un linaje genómico partiendo de un cluster y un gen de referencia. Todos los contextos genómicos que contengan ese gen y algún otro gen del cluster de referencia serán encontrados en el linaje seleccionado por el usuario. CORASON identifica el core génico de la familia. La información del core se utiliza para organizar filogenéticamente todos los miembros de la familia del BGC, es decir todas las variantes del BGC serán organizadas. El core génico está relacionado con el core de la molécula, la parte variable del BGC codifica enzimas accesorias que producen ornamentos i.e. variantes moleculares. Al cambiar de gen de referencia CORASON permite explorar otras familias de BGC que contengan las mismas modificaciones. Los resultados se presentan en una visualización que permite al mismo tiempo apreciar variación a nivel de presencia-ausencia de genes entre miembros de una familia de BGCs, como también apreciar variación a nivel de secuencia a través de un gradiente de color entre genes conservados entre una variante y el BGC de referencia.

### 3.1. Algoritmo y características de CORASON

#### CORASON: CORe Analysis of Syntenic Orthologs to Prioritize Natural Products Biosynthetic Gene Clusters

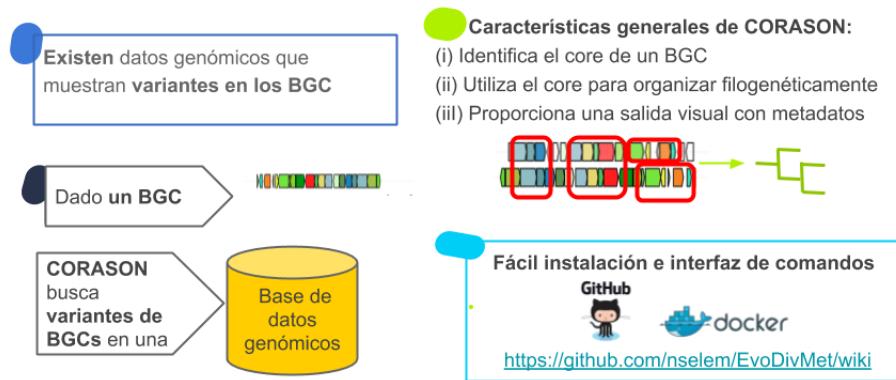


Figura 3.2: CORASON es una herramienta en línea de comandos para organizar filogenéticamente la variación de una familia de clusters. Dada la existencia de variantes de BGCs en bases de datos de linajes genómicos se diseñó CORASON. Corason se ejecuta en línea de comandos, su función es identificar el core génico de un BGC y utilizarlo para presentar una visualización de las variantes de la familia organizadas filogenéticamente

La herramienta corason, como se muestra en la Figure 3.2 localiza familias de clusters biosintéticos en un linaje genómico partiendo de un cluster y un gen de referencia. Todos los contextos genómicos que contengan ese gen y algún otro gen del cluster de referencia serán encontrados en el linaje seleccionado por el usuario. CORASON identifica el core génico de la familia. La información del core se utiliza para organizar filogenéticamente todos los miembros de la familia del BGC, es decir todas las variantes del BGC serán organizadas. El core génico está relacionado con el core de la molécula, la parte variable del BGC codifica enzimas accesorias que producen ornamentos i.e. variantes moleculares. Al cambiar de gen de referencia CORASON permite explorar otras familias de BGC que contengan las mismas modificaciones. Los resultados se presentan en una visualización que permite al mismo tiempo apreciar variación a nivel de presencia-ausencia de genes entre miembros de una familia de BGCs, como también apreciar variación a nivel de secuencia a través de un gradiente de color entre genes conservados entre una variante y el BGC de referencia.

CORASON permite la rápida identificación de variantes de BGCs y organiza los resultados utilizando una aproximación filogenética. En la figura [CorasonPipeline] se muestra esquemáticamente como dado un cluster de referencia y un gen localizado en el cluster, así como las secuencias genómicas de un linaje de referencia corason buscará todas las variantes del cluster de referencia que existen en el linaje genómico seleccionado.

### **3.2. Las familias de BGCs están formadas por variantes del BGC de referencia.**

Así como existen familias génicas, un gen y todos sus homólogos, incluidos parálogos, ortólogos y xenólogos, también existen familias de BGCs. No es tan claro definir un BGC porque no tiene un codón de inicio y un codón de paro como un gen. En algunas ocasiones como en el caso de scytonemin todos los genes del BGC se expresan al recibir un estímulo, como los rayos UV en este caso. Otras veces en la producción del metabolito participan genes que no son necesariamente contiguos, y por ello al cambiar el BGC de organismo y realizar expresión heteróloga no se obtiene el mismo metabolito. Las fronteras es decir los genes de la orilla del BGC no siempre son claras.

Así pues, cuando se habla de un BGC no se debe pensar que este es igual en todos los organismos del linaje, es decir que tiene el mismo contenido génico. Hay variación tanto a nivel de contenido génico como a nivel de secuencia entre los genes en común. Esta variación produce la promiscuidad de producto, una familia de BGCs produce distintos productos a partir de los mismos precursores. En este trabajo vamos a tomar como BGC de referencia a los anotados en MIBiG, que son de los que se tienen datos experimentales respecto al producto reportado. Todas sus variantes que contengan al menos dos genes en común, uno de referencia seleccionado por el usuario y otro cualquiera pero común con el BGC de referencia son considerados parte de la familia del BGC.

Como ejemplo de familias de BGCs podemos pensar a los operones. Un operón es el conjunto de genes dedicados a la síntesis de algún metabolito del metabolismo central, estos genes suelen encontrarse juntos en el cromosoma y transcribirse simultáneamente. Ejemplos de estos operones son los que producen los aminoácidos histidina y triptófano. Estos “BGCs” hace millones de años fueron posiblemente parte del metabolismo especializado y debido a su éxito se fijaron en lo que ahora vemos como BGCs conservados en ciertos linajes genómicos. En estos casos las fronteras son claras y la variación génica es poca. Aún así existen otros ejemplos donde puede constatarse variación en las rutas de síntesis de mecanismos centrales de metabolismo procariota. En el caso de histidina y triptófano, como hablaremos en el siguiente capítulo, es la variación a nivel de secuencia y no a nivel de composición génica la que produce diversidad de producto.

En oposición a los BGCs u operones de metabolismo conservado, están los BGCs del actual metabolismo especializado, así como se encuentran familias con mucha variación génica pueden encontrarse otras muy conservadas. En este capítulo presentaremos varios ejemplos de familias de BGCs.

### 3.3. Aplicaciones de CORASON en Actinobacteria y Pseudomonas

En el capítulo anterior las variantes del BGC scytonemin en Cyanobacteria fueron encontradas al aplicar CORASON a dicho linaje. En este capítulo presento ejemplos del uso de CORASON para investigar los patrones de conservación/variación en familias de BGCs en los linajes Actinobacteria y *Pseudomonas*. Primero, versiones iniciales de CORASON fueron usadas junto con cromatografía y espectrometría de masas (LC-MS) para estudiar la ecología y evolución de los sideróforos del tipo desferroxiaminas en Actinobacteria. Este estudio permitió la identificación de desG, un miembro de la familia penicilin amidasa responsable de la arilación de desferroxiaminas en actinomicetos acuáticos [REF].

En un segundo ejemplo, CORASON fue utilizado para investigar la existencia de variantes en Actinobacteria del cluster productor de arsenolípidos que se encuentra en *Streptomyces lividans*. En la tercera aplicación, se investigó la distribución en Actinobacteria de un BGC de *Streptomyces lividans* que es novedoso porque involucra una enzima que utiliza tRNA de la familia FemXAB. La predicción de producto de este BGC es un metabolito que contiene un grupo azoxy. Esta predicción ha sido confirmada utilizando LC-MS [Aguilar in prep].

Finalmente, un cuarto ejemplo es presentado: el de los contextos genómicos de *tauD* que codifica una dioxigenasa involucrada en el metabolismo de taurina en su papel de enzima del metabolismo conservado. En Actinobacteria *tauD* es parte de 15 BGCs reportados en MIBiG, aunque en *Pseudomonas* no existen BGCs reportados EvoMining predice expansiones en este linaje genómico y CORASON predice conservación en los contextos genómicos de estas copias extra que guardan cierta similitud con variantes de los BGCs conocidos en Actinobacteria.

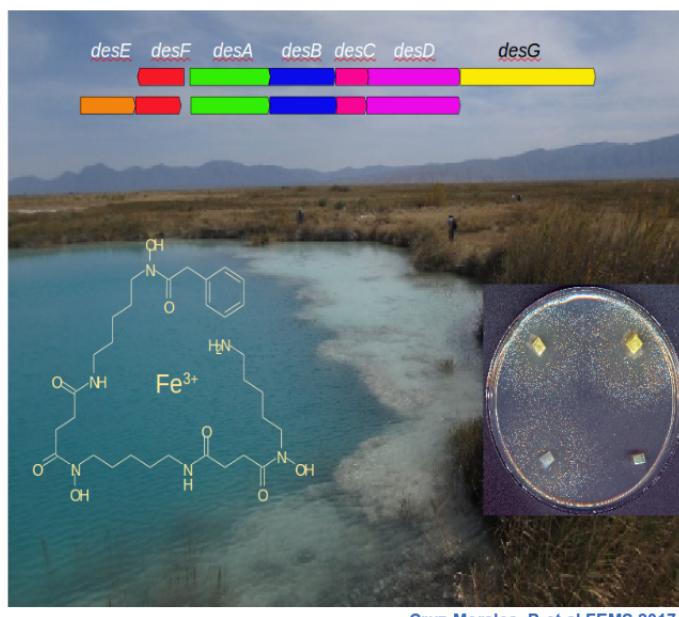
Estos cuatro ejemplo ilustran como CORASON puede ser utilizado tanto para priorizar nuevos BGCs como para ligar la variación génica de familias de BGCs con diversidad estructural. CORASON es una expansión de las habilidades de EvoMining [cruz-morales] de encontrar enzimas en el proceso de diversificación funcional. CORASON encuentra familias de BGCs, y presenta una rápida visualización de sus variantes. Ambas herramientas fueron desarrolladas para expandir nuestro conocimiento sobre nueva química mediante la genómica comparativa. CORASON está disponibles en su contenedor de docker en github <https://github.com/nselem/corason>.

### 3.4. Estas familias pueden clasificarse

Variantes del cluster biosintético de desferroxiamina Figure 3.3 fueron identificadas por CORASON en Actinobacterias de cuatro ciénegas. Una variante del BGC de desferroxiamina fue identificada. Esta variante se diferencia del cluster reportado por

la ganancia de una penicilin amidasa. A este gen extra se le llamó *desG*. Fue verificado que la variación génica está ligada con la variación molecular

El hierro es necesario en el metabolismo de muchos seres vivos. Para poder utilizarlo las bacterias han desarrollados moléculas captadoras de hierro llamadas sideróforos. Un ejemplo de sideróforo es la molécula de desferroxamina sintetizada por los genes *des* en Actinobacteria



Cruz-Morales, P. et al FEMS 2017

Figura 3.3: Variantes del cluster biosintético de desferroxamina fueron identificadas por CORASON en Actinobacterias de cuatro ciénegas. Una variante del BGC de desferroxamina fue identificada. Esta variante se diferencia del cluster reportado por la ganancia de una penicilin amidasa. A este gen extra se le llamó *desG*. Fue verificado que la variación génica está ligada con la variación molecular

### 3.5. Un BGC reportado puede tener muchas variantes que conforman una familia de BGCs

En el linaje de *Streptomyces* EvoMining obtuvo como predicción una arsenoenol piruvato sintasa en una rama del árbol de expansiones de la familia 3-fosfoshikimato-1-carboxivinyl transferasa (AroA). Estudios de mutagénesis y de expresión diferencial génica en presencia de arsénico confirmaron que en *Streptomyces coelicolor* y en *Streptomyces lividans* esta enzima pertenece a un BGC que sintetiza arsenolípidos [Pablo tesis]. El contexto genómico de AroA en estos organismos incluye una enzima PKS localizada a seis genes de distancia. AntiSMASH predice los BGCs tipo PKS de estas enzimas pero no incluye en ellos a la arsenoenol piruvato sintasa. El valor de

EvoMining fue descubrir que esta copia extra de AroA estaba dedicada al metabolismo especializado y por tanto bien podía tener su propio BGC o dada la cercanía con las PKS podía ser parte de estos PKS-BGCs. Esta ruta de síntesis de arseno compuestos, fue descubierta a partir de una rama de copias extra de AroA donde se apreciaba diversidad en las secuencias de aminoácidos. Una pregunta posible es si esta diversidad de secuencia a nivel de enzima ascendía a un siguiente nivel. Es decir si también existe diversidad a nivel del BGC, si se encuentran variantes con distintos patrones de presencia/ausencia en los genes que coponen al BGC de *S. coelicolor*. O más aun, quedaba por investigar si el BGC tenía cierto grado de conservación o era exclusivo de estos dos organismos *S. coelicolor* y *S. lividans*

Una versión preliminar de CORASON, permitió visualizar las variantes de los contextos genómicos de las arsenol piruvato sintasas. Los genomas donde se buscaron estas variantes fueron seleccionados por una búsqueda de blast en NCBI en la base de datos no redundante como aparecía en 2016. Después de organizar manualmente los BGCs como se muestran a la izquierda de la [Fig Arseno-BGC], la visualización permitió distinguir 4 grupos de contextos conservados. El primero independiente de PKS y NRPS mostrado en morado, el segundo con una NRPS-PKS híbrida mostrado en un rectángulo verde y finalmente en rosa y naranja se muestran los grupos tercero y cuarto que contienen una PKS, los primeros del lado izquierdo y los segundos del lado derecho. Presumiblemente estos BGCs aunque contienen un core común producen distintos arseno compuestos.

El contexto genómico conservado en las expansiones de AroA puede apreciarse en la Figure 3.4. La arsenoenol piruvato sintasa fue descubierta en un árbol de EvoMining como parte de una rama de expansiones de AroA en Actinobacteria. El contexto genómico de la arsenoenol piruvato sintasa de *S. coelicolor* tiene un core conservado en Actinobacteria. Este BGC está dedicado a la síntesis de metabolitos secundarios de tipo arsenolípidos. Primero se identificaron en otras secuencias genómicas contextos que contengan el homólogo de AroA y algún otro gen de su vecindad en *S. coelicolor*. A continuación, se ordenaron manualmente los contextos obtenidos y se pudo identificar al menos cuatro diferentes clases de BGCs. La primera clase mostrada en un rectángulo morado no contiene PKS o NRPS, la segunda enmarcada en verde contiene una PKS-NRPS híbrida. La tercera clase incluye una PKS a la izquierda y a más de cinco genes de distancia de la Arsenoenol piruvato sintasa y finalmente la última clase contiene una PKS a la derecha y a sólo un gen de distancia de esta enzima. Esta figura muestra que existen variantes de un BGC

La visualización realizada para los arsenolípidos no incluía ningún tipo de orden y era difícil distinguir grupos de BGCs. Por esta razón se pensó que el orden filogenético ya no de una enzima sino de el core del BGC ordenaría las variantes del BGC. Este orden mostraría en un continuo la dinámica genómica del BGC establecida por los procesos evolutivos. En el caso de los arsenolípidos se utilizó Orthocore para la identificación del core del BGC. En las siguientes versiones de CORASON esta característica fue implementada como parte del algoritmo.

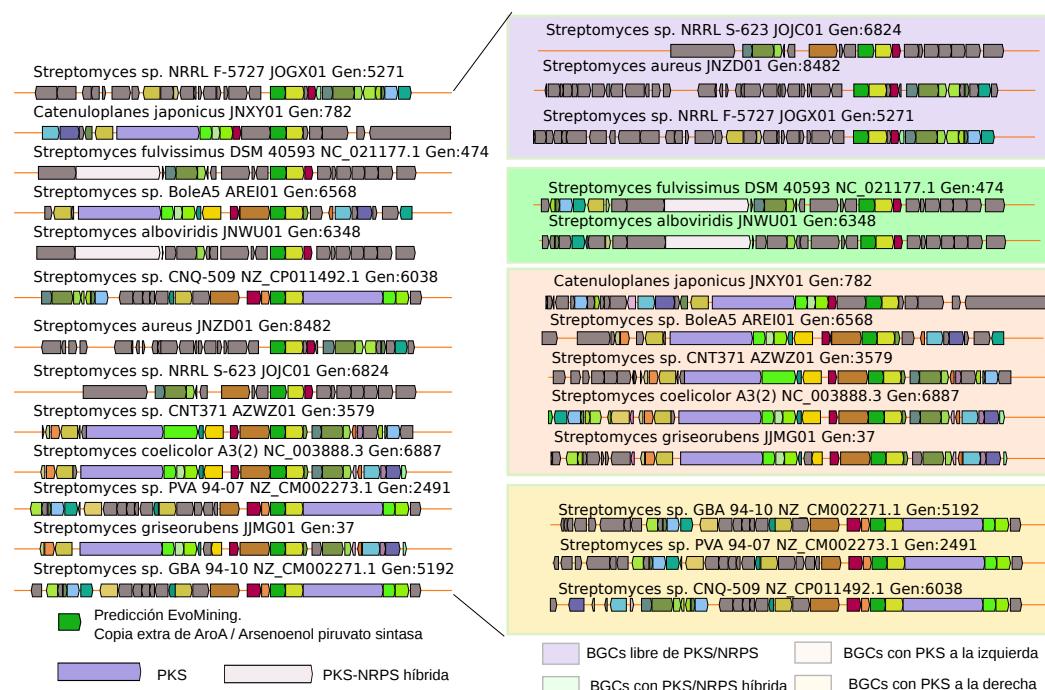


Figura 3.4: Contexto genómico conservado en las expansiones de AroA. La arseenoenol piruvato sintasa fue descubierta en un árbol de EvoMining como parte de una rama de expansiones de AroA en Actinobacteria. El contexto genómico de la arseenoenol piruvato sintasa de *S. coelicolor* tiene un core conservado en Actinobacteria. Este BGC está dedicado a la síntesis de metabolitos secundarios de tipo arsenolípidos. Primero se identificaron en otras secuencias genómicas contextos que contengan el homólogo de AroA y algún otro gen de su vecindad en *S. coelicolor*. A continuación, se ordenaron manualmente los contextos obtenidos y se pudo identificar al menos cuatro diferentes clases de BGCs. La primera clase mostrada en un rectángulo morado no contiene PKS o NRPS, la segunda enmarcada en verde contiene una PKS-NRPS híbrida. La tercera clase incluye una PKS a la izquierda y a más de cinco genes de distancia de la Arseenoenol piruvato sintasa y finalmente la última clase contiene una PKS a la derecha y a sólo un gen de distancia de esta enzima. Esta figura muestra que existen variantes de un BGC

### 3.6. CORASON analiza la conservación del contexto genómico de los EvoMining hits

## 3.6. CORASON analiza la conservación del contexto genómico de los EvoMining hits

La predicción de un BGC para la síntesis de un compuesto tipo azoxy es otro ejemplo de CORASON como complemento de EvoMining para la identificación de los dos niveles de promiscuidad, tanto a nivel familia enzimática como familia de BGCs Figure 3.5. Este trabajo se encuentra en preparación para su publicación por Aguilar-Martínez.

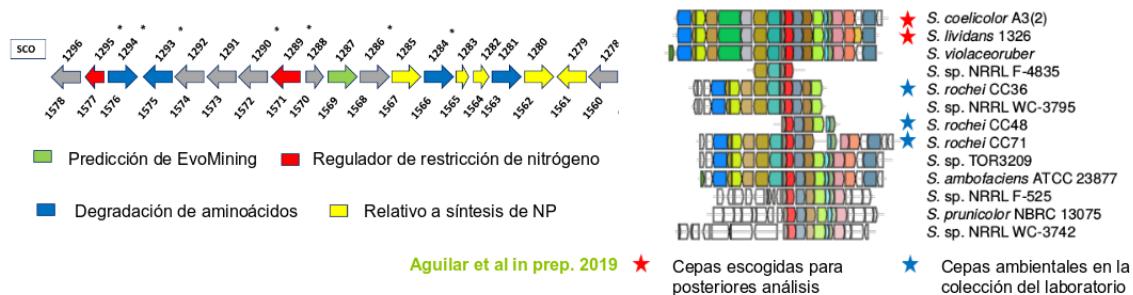


Figura 3.5: Azoxy BGC

### **3.7. El contexto de una rama divergente de *tauD* está conservado en Pseudomonas**

En Pseudomonas el gen *tauD* tiene un contexto que incluye genes de metabolismo especializado. Variantes de este contexto están distribuidas en varios organismos Figure 3.6 . Parecen ser diferentes de los rimosamides y detoxins reportados en Actinobacteria.

### **3.8. CORASON se adaptó a la herramienta BiG-SCAPE de clasificación de familias de BGCs**

BIG-SCAPE es una herramienta bioinformática para separar un conjunto de BGCs en familias de acuerdo al contenido, conservación y distribución de sus dominios. Un dominio es un conjunto de secuencia conservado de un gen, es considerado una unidad de las proteínas. CORASON complemento BIG-SCAPE Figure 3.7 al proporcionar un algoritmo para ordenar la diversidad dentro de la familia. A la vez CORASON permite conectar mediante la evolución a familias aparentemente separadas por BiGSCAPE.

### **3.9. BiG SCAPE y CORASON identificaron nuevos productos variantes de la familia de BGCs Rimosamide - Detoxin en Actinobacteria**

En esta sección utilizamos CORASON y BIG-SCAPE para analizar una base de datos de miles de genomas de ACtinobacteria. Con estas herramientas se organizó la diversidad biosintética de las familias de BGC toxin y rimosamide [168]. El análisis reveló diversidad tanto en los géneros de los organismos que contienen a esta familia, y en la composición génica del BGC. Entre los géneros con alguna variante del BGC están Amycolatopsis, Streptomyces. El core conservado de los BGCs toxin y rimosamide está compuesto por una NRPS, una NRPS/PKS híbrida, y un homólogo de tauD. Este homólogo fue sugerido por un análisis de EvoMining. En *E.coli* tauD se encuentra en el operón tauABCD [MM1] La ruta de síntesis de rimosamide difiere de la de detoxins porque tiene una NRPS adicional, que codifica para una modificación del core de molécula toxin/rimosamide con isobutyrate y glycine39.

El hecho de que el gen *tauD* estuviera presente en todos los miembros de la familia Figure 3.8 captó nuestra atención. El producto del gen *tauD* pertenece a la super familia de enzimas Fe(II)/ $\alpha$ -ketoglutarato-dependientes hidroxilasas. En particular

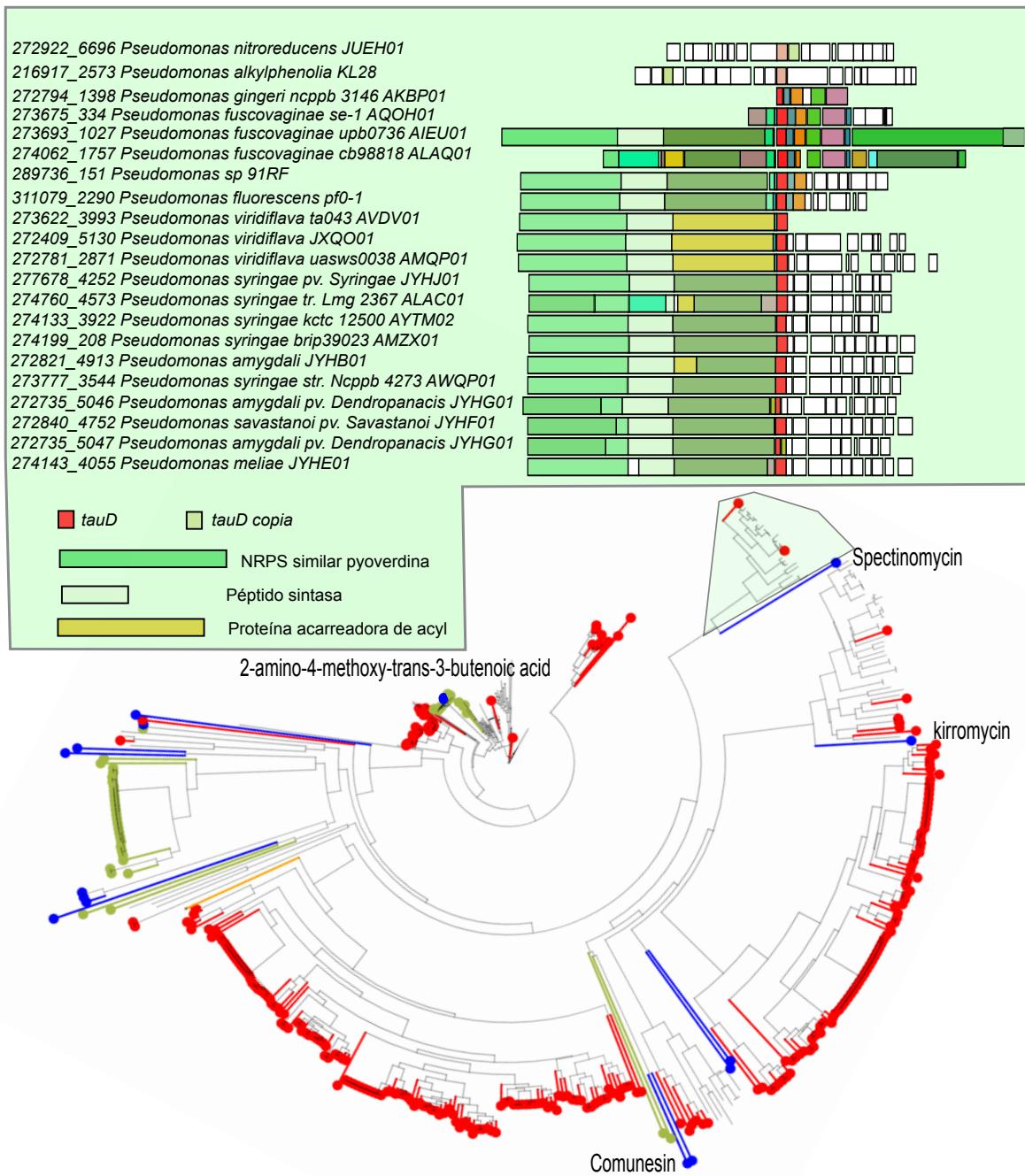


Figura 3.6: En *Pseudomonas* el gen *tauD* tiene un contexto que incluye genes de metabolismo especializado. Variantes de este contexto están distribuidas en varios organismos. En particular

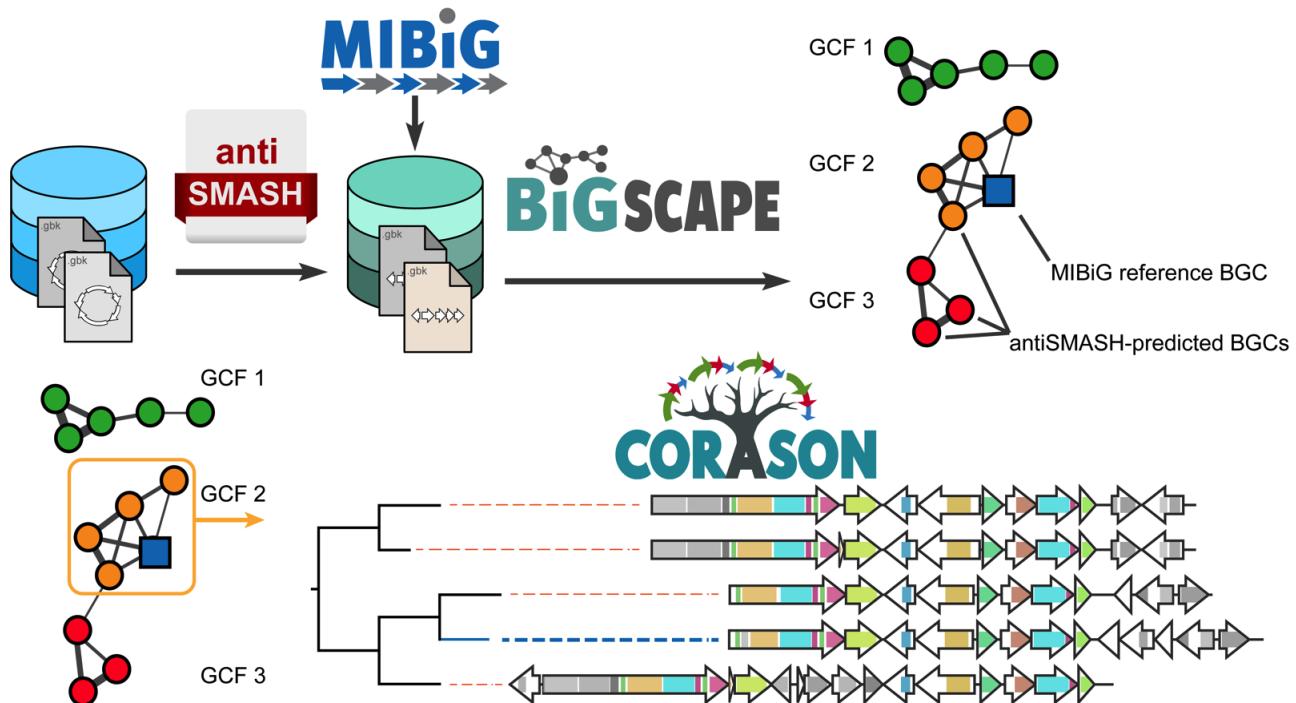


Figura 3.7:

tauD codifica una  $\alpha$ -ketoglutarato-dependiente taurina dioxigenasa involucrada en la asimilación de sulfito por la liberación oxigenolítica del aminoácido taurina41. Interesantemente, esta familia también incluye enzimas en linajes como hongos, bacterias y plantas. Dichas enzimas catalizan hydroxylations, desaturations, ring expansions and ring formations, among other chemical transformations. A la fecha, el rol de TauD en la biosíntesis de los metabolitos detoxin y rimosamide aún es desconocido, se ha sugerido que es responsable de la oxidación de la prolina observada en algunos análogos39.

Para identificar variantes de los BGCs relacionados a detoxin y rimosamide dentro de la rama de metabolismo especializado los 1175 BGCs del dataset que contenían un homólogo de tauD se pasaron por un análisis combinado de BiG-SCAPE/CORASON. Los BGCs fueron procesados con CORASON usando tauD como query gene. Teniendo como resultado que es el único miembro del ‘BGC core’. Es importante notar que el core del BGC puede contener entre 1 y 3 genes, [La NRPS, y la NRPS-PKS híbrida quedan en este ejemplo fuera del core debido a gaps en la secuencia del genoma de organismos como *Streptomyces humi*, *Streptomyces Spectabilis* y *Amycolatopsis Vancoresmycina*]

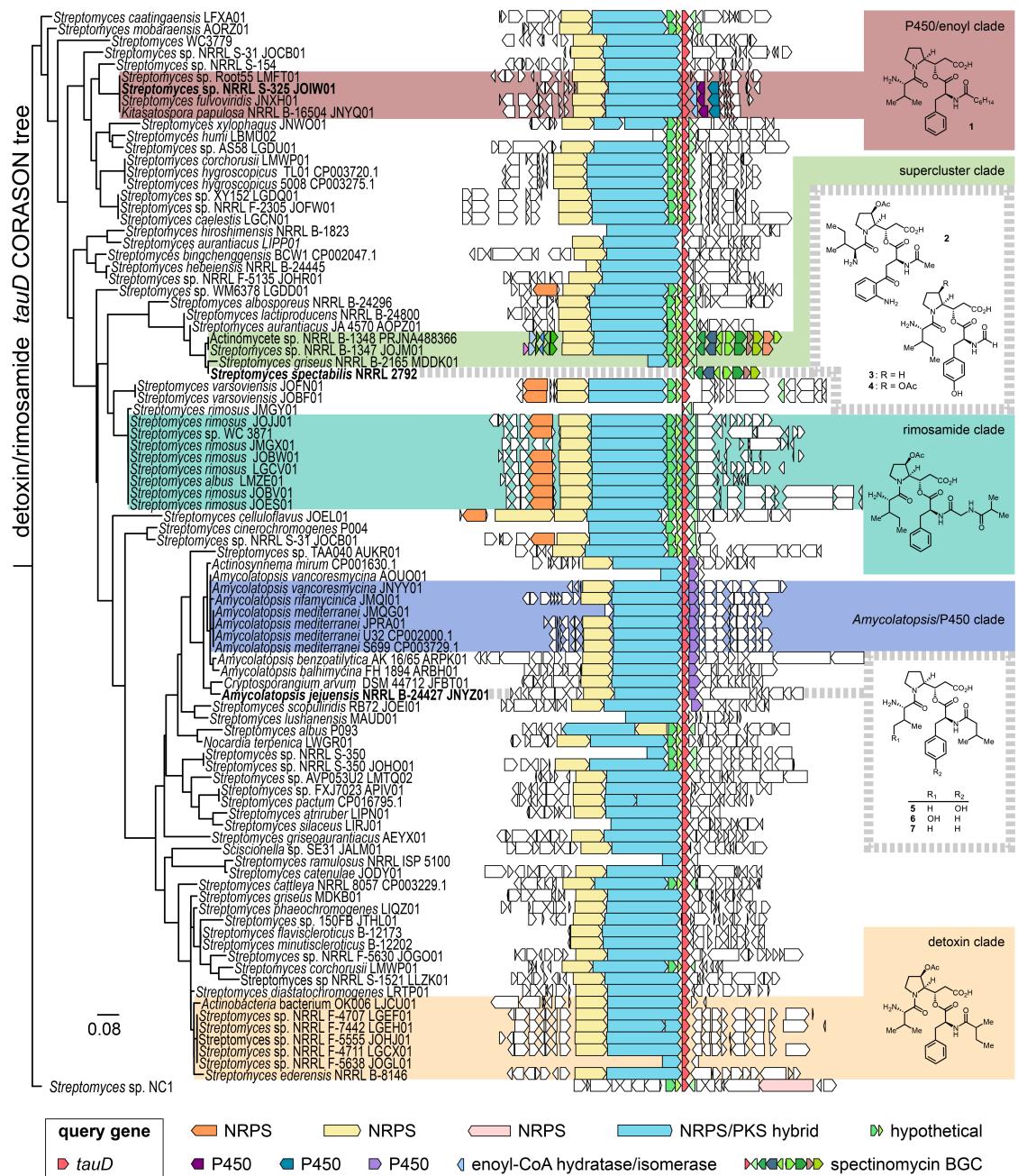


Figura 3.8: Rimosamide

### **3.10. CORASON sugiere que las familias detoxin y rimosamide pertenecen a un amplio clan de familias dedicadas a la síntesis de péptidos.**

El análisis de CORASON reveló que las familias de los BGCs detoxin y rimosamide GCFs identificados en BiG-SCAPE eran parte de todo un clan de biosíntesis de péptidos que incluía clados inexplorados [MM10] del phylum Actinobacteria [Fig TauDActinobacteria]. La organización filogenética de los BGCs provista por CORASON, reveló familias de BGCs que fueron omitidas debido al umbral utilizado por el algoritmo de clustering de BiG-SCAPE. Esto debido a la cercanía de genes biosintéticos detectados por antiSMASH como suficientemente diferentes como para ser clasificados en otras familias (ver Fig. S10).

Hipotetizamos que las detoxins codificadas por los BGCs en los clados inexplorados contendrán cierta novedad química relacionada con las variaciones genéticas. Afortunadamente, 40 de las 152 cepas identificadas como portadoras de un BGCs estaban representadas en nuestros datos metabolómicos de 363–cepas por LC-MS/MS. Análisis de redes moleculares de estos datos indicaron la presencia de tres detoxin conocidas, cuatro rimosamide conocidas y otras 103 variantes de detoxin o rimosamide (Fig. 3c), el basto universo químico sugerido por el análisis BiG-SCAPE/CORASON.

### **3.11. Clados selectos del árbol de CORASON que contiene a las familias rimosamide/detoxin contienen diversidad génica que correlaciona con novedad química.**

Tres de los clados que codifican para detoxin BGC fueron identificados por BiG-SCAPE dentro del árbol de CORASON capturaron nuestro interés ([Fig TauDActinobacteria], in colored boxes). En esta sección se describe el trabajo experimental realizado por Michael Mulloney del grupo de colaboradores de West Chicago, de los tres clados y específicamente de los organismos que seleccioné como candidatos a presentar diversidad química.

#### **3.11.1. El clado P450/enoyl agrega una heptanamida al core molecular detoxin/rimosamide**

El primer clado es el ‘P450/enoyl clade’ contiene genes como el citocromo P450 y un enoyl-CoA hidratasa/isomerasa dentro de cada uno de sus BGCs. Este clado

está marcado en rojo [Fig TauDActinobacteria]. Análisis de datos por tandem MS de extractos de *Streptomyces* sp. NRRL S-325, que se encuentra dentro de este clado, llevó al descubrimiento de la detoxin S1 (1; Fig TauDActinobacteria, S16–17). Este nuevo análogo contiene una cadena lateral de heptanamide, una extructura única entre las toxins y rimosamides cuya instalación posiblemente depende de la enzima enoyl-CoA hydratase/isomerase.

### 3.11.2. El superclado spectinomycin/ detoxin-rimosamide-clan produce cinco variantes de toxin.

El segundo clado de interés,fue nombrado el ‘supercluster clade’ (Fig. 5, en verde claro), comprende los BGCs con genes de detoxin adjacentes al cluster que produce spectinomycin42 [Fig TauDActinobacteria]. El cluster de spectinomycin (MIBiG BGC0000715) contiene en su periferia al gen tauD como se muestra en la línea gris punteada de la [Fig TauDActinobacteria]. La secuencia del cluster de spectinomycin depositada en MIBiG es la única secuencia disponible de *Streptomyces spectabilis* NRRL 2792. Como no se sabe que tauD participe en la síntesis de spectinomycin se hipotetizó que pueden existir los genes del cluster de detoxin al lado de los genes del BGC de spectinomycin en *S. spectabilis* NRRL 2792. Adquirimos esta cepa para determinar si el análisis de CORASON podía ayudar a la predicción de detoxin basado solamente en la presencia del query gen pero en ausencia completa de la secuencia del BGC de detoxin. Análisis en tandem de espectrometría de masas de extracto *S. spectabilis* NRRL 2792 reveló la producción de cinco compuestos tipo detoxin, incluyendo detoxin N1 (2; Fig TauDActinobacteria, S15), detoxin N2 (3; Fig. S15) y su análogo acetoxylated, detoxin N3 (4; Fig. S15).

Los tiempos de retención de ionies y los patrones de fragmentación de los últimos dos compuestos también fueron observados en extractos de *Streptomyces* sp. NRRL B-1347 parte del clado del supercluster, confirmando la habilidad de CORASON para guiar un descubrimiento, mediante la utilización de la filogenia a pesar de lo limitado de los datos en la cepa NRRL-2792. Análisis de LC-MS de cultivos de NRRL-2792 suplementados con isotopos estables etiquetados de amino acidos corroboraron las predicciones estructurales basadas en los análisis de la cepa cercana *Streptomyces* sp. NRRL B-1347 (Fig TauDActinobacteria, S26–31). Los tres nuevos análogos incorporan completamente el 13C6-isoleucine, pero d7-proline solo es incorporado en el compuesto 3. La pérdida de un deuterio de d7-proline en 2 y 4 soporta la asignación de acetoxylation del anillo de pyrrolidine común en toxins y rimosamides<sup>39,43–45</sup>. Características especiales únicas a la serie de N-detoxins incluyen la incorporación de una tirosina N-formylated en 3 y 4 en lugar de fenilalanina, el residuo típico de detoxin/rimosamide, lo que es soportado por la incorporación del anillo d4-tyrosine. El compuesto 2 la incorporación única de un residuo derivado del triptofano en esta posición, haciendo evidente la retención de cuatro deuterios cuando en experimentos de alimentación de indole-d5-tryptophan (Fig. S27). Aunque los datos de MS fueron insuficientes para desenmascarar esta estructura, el compuesto 2 fue producido por *S. spectabilis* NRRL

2792 en suficiente abundancia para el aislamiento y la elucidación estructural por NMR. Varios experimentos 1D y 2D confirmaron las asignaciones por datos de MS y establecieron una N-acetylated kynurenone como la estructura derivada de triptófano en 2 (Figs. S18-S27).

### **3.11.3. El clado *Amycolatopsis P450* produce cinco variantes de detoxin.**

El tercer clado que se estudió del super clan al que pertenecen las familias rimosamide-detoxin, contiene BGCs casi enteramente provenientes del género *Amycolatopsis*. Este clado está marcado en morado en la [Fig TauDActinobacteria]. Este clado de BGCs también contienen un gen P450 único entre los BGCs del árbol, así que fue llamado el clado ‘*Amycolatopsis/P450* clade’. Aunque no se contaba con datos metabolómicos de las cepas del clado de BGCs definidos por BiG-SCAPE como una gen cluster family (GCF), la visualización filogenética de CORASON permitió la selección de una cepa de *Amycolatopsis* de la que se tenían datos metabolómicos con un BGC muy similar, y que también contiene el gen P450 ([Fig TauDActinobacteria], línea gris cerca del clado *Amycolatopsis/P450*). Análisis de datos de tandem MS de extracto fermetado de *Amycolatopsis jejuensis* NRRL B-24427 reveló isómeros de detoxins P1 (5; [Fig TauDActinobacteria], S15) que contienen tirosina, P2 (6; [Fig TauDActinobacteria], S15) mostrando fenilalanina y una valina hidroxilada, así como la detoxin P3, un análogo cercano libre de hidroxilación (7; [Fig TauDActinobacteria], S15). Una validación de la asignación de aminoácidos observados en los patrones de fragmentación de MS/MS se consiguió mediante el uso de experimentos de incorporación de isótopos estables etiquetados de aminoácidos (Figs. S33–S36, S38–S41, and S43–S44).

## **3.12. CORASON permitió expandir las capacidades de EvoMining y explorar la promiscuidad de familias de BGCs de enzimas divergentes de metabolismo central.**

Nuestros resultados ilustran como BiG-SCAPE puede identificar conjuntos de BGCs relacionados, en un gran número de secuencias de genomas. Además al usar CORASON para reconstruir las filogenias de BGC para ordenar visualmente la evolución de un cluster biosintético y su diversidad proveen herramientas poderosas para el descubrimiento de nuevos clados de BGC que codifican en consecuencia para nueva química. Respecto a los BGCs detoxin/rimosamide, CORASON mostró habilidad para ayudar a minar bases de datos genómicas y descubrir siete nuevas detoxins. Específicamente, la organización de las variantes de los BGC facilitó la identificación de los correspondientes variaciones en la estructura química -la presencia de un enoyl-CoA

**3.12. CORASON permitió expandir las capacidades de EvoMining y explorar la promiscuidad de familias de BGCs de enzimas divergentes de metabolismo central<sup>99</sup>**

hidratasa/isomerasa corresponde a la familia de amida ácido graso detoxin S1 y la presencia de un gen P450 corresponde a la presencia de hidroxilaciones en detoxins P1–P3.

## **Discusión**

El cluster des está muy conservado, no así los de arsenolípidos, ni los de rimosamida



# Capítulo 4

## La familia PriA

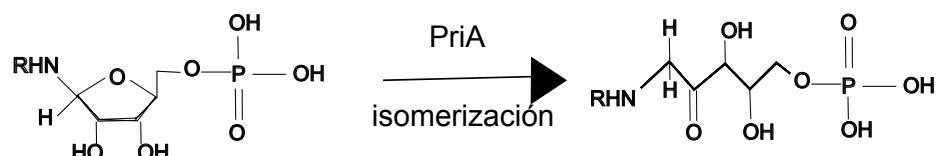


Figura 4.1: La reacción que cataliza PriA es una isomerización donde abre un anillo de cinco carbonos.

PriA es la familia de enzimas de Actinobacteria homóloga a la familia HisA existente en Enterobacteria, Cyanobacteria, *Pseudomonas* y Archaea. Según las definiciones de este trabajo PriA es una familia promiscua ya que varios miembros de PriA han demostrado mediante cinéticas enzimáticas *in vitro*, su capacidad de catalizar tanto la reacción correspondiente a su homólogo HisA como la isomerización que cataliza TrpF. Como ya se explicó en la introducción TrpF es una reacción involucrada en la producción de triptofano. Es decir, se ha mostrado que en varias Actinobacterias PriA participa a la vez en las rutas de síntesis de histidina y triptofano. Los primeros miembros caracterizados como promiscuos en esta familia fueron *Streptomyces coelicolor* y *Mycobacterium tuberculosis* [5] durante el año 2003.

La mayoría de las actinobacterias han perdido el gen *trpF* aunque conservan el resto del operón de triptofano, por lo que se cree que la promiscuidad de PriA está extendida en un gran subconjunto de Actinobacteria. En la ruta de histidina PriA isomeriza el sustrato ProFAR en PRFAR, realizando la función HisA. En la ruta de triptofano PriA lleva a cabo la isomerización de PRA en CdRP actuando análogamente a la función TrpF en la ruta del triptofano. La vecindad genómica de PriA en *Streptomyces coelicolor* y en otras Actinobacterias contiene genes tanto del operón de histidina como del de triptofano.

Además, en Actinobacteria PriA ha mostrado un gradiente funcional. Esta variación divide a la familia en varias subfamilias según su capacidad catalítica en los sustratos ProFAR y PRA. En la introducción mostramos datos que ejemplifican las subfamilias

de PriA, incluyendo ejemplos de miembros de PriB, la subfamilia ubicada en el género *Streptomyces* con baja capacidad de catálisis para la función TrpF. Varios *Streptomyces* con un ortólogo en la familia PriB, se diferencian de otras Actinobacterias en que contienen en su genoma un gen *trpF* localizado fuera del contexto genómico inmediato de los operones de histidina y triptofano. Otra subfamilia de PriA es subHisA, que ha perdido totalmente la actividad TrpF, existen miembros de subHisA en *Corynebacteria* y en *Actinomycetes*. Finalmente en Actinomycetes encontramos la subfamilia *subTrpF* que ha perdido la actividad de HisA.

En este capítulo en las siguientes secciones se exploran cuatro aspectos de la familia PriA. *i)* La distribución y el contexto genómico de PriA en diversos linajes genómicos. *ii)* La información contenida a nivel de aminoácidos en variantes de PriA como medio de estudio de rutas evolutivas y su relación con la reconstrucción de su estructura tridimensional. *iii)* Las posibles afinidades de PriA por otros sustratos con métodos bioinformáticos, y finalmente *iv)* La validación experimental de PriA en sustratos diferentes de PRA y ProFAR o bien en combinaciones de los mismos.

#### **4.1. PriA como modelo de familia enzimática donde las expansiones no son condición necesaria para la promiscuidad.**

##### **PriA en EvoMining**

Para explorar PriA en diversos linajes genómicos se utilizaron las herramientas EvoMining y CORASON descritas en los capítulos previos. Se investigaron las expansiones de la familia PriA en los linajes Actinobacteria, Cyanobacteria, *Pseudomonas* y Archaea. En Actinobacteria, donde se sabe que PriA es promiscua no se detectaron copias extra. En la Figure 4.2 se muestra el número promedio de copias por genoma en los linajes genómicos seleccionados. Según EvoMining en Actinobacteria no hay expansiones, prácticamente todas las copias son reconocidas como de metabolismo central (rojo o morado) aunque algunas PriA además son marcadas por antiSMASH como parte de algún cluster biosintético (morado). En cambio en Cyanobacteria, *Pseudomonas* y Archaea la figura muestra en negro las copias extra de las que no se conoce su destino metabólico. El caso de Archaea es llamativo porque las copias de metabolismo central llegan en promedio hasta .5 copias por genoma, es decir muchos genomas de Archaea no cuentan con una copia de PriA, y en cambio, contrario a Actinobacteria, un 50 % de la copias es marcado en negro, es decir varios de los genomas que tienen al menos una copia de PriA en realidad tienen dos copias. Esta figura muestra que en Actinobacteria PriA constituye un ejemplo de familia promiscua mayoritariamente distribuida con una sola copia por genoma. Esta observación demuestra que para que una familia sea promiscua no es imperativo tener copias extra con marcas de reclutamiento en metabolismo especializado. Aunque las copias extra suelen ser una

indicación de promiscuidad, no son una condición necesaria. Tanto EvoMining como CORASON mostraron que existen excepciones de organismos donde PriA tiene doble copia, tanto en Actinobacteria como en otros linajes genómicos.

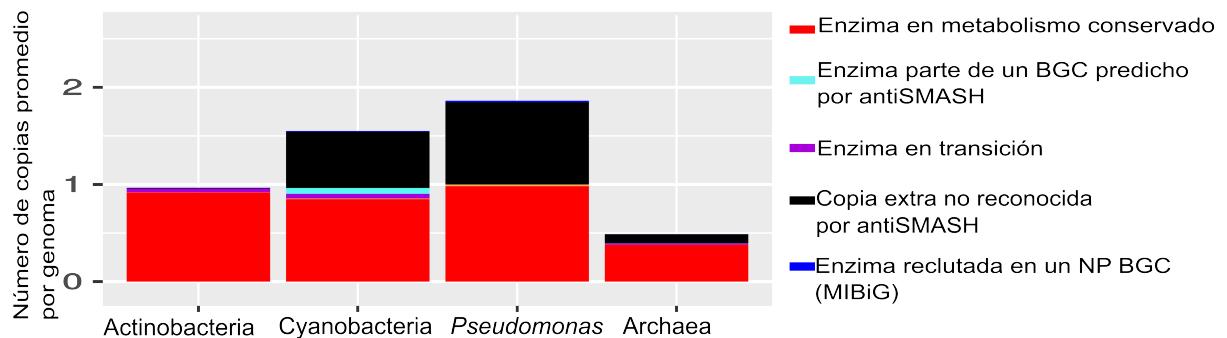


Figura 4.2: Número promedio de copias por genoma de PriA en Actinobacteria, Cyanobacteria, Pseudomonas y Archaea. Los colores muestran el destino metabólico asignado a cada copia según EvoMining. En rojo están los BBH a las enzimas semilla de metabolismo central. En morado las enzimas de metabolismo central también reconocidas por antiSMASH como parte de unBGC y en negro las copias sin un destino metabólico conocido.

Después del conteo de número de copias promedio, se analizaron los árboles de PriA de EvoMining, tanto los coloreados de acuerdo al número de copias, Figure 4.3, como los coloreados según el destino metabólico, Figure 4.4. En Actinobacteria la mayoría de las hojas son verdes reafirmando que existe sólo una copia por genoma en ese organismo. Sin embargo, existen varias hojas de color amarillo, indicando de dos copias en ese genoma. Las Actinobacterias con dos copias son *Ornithinimicrobium pekingense* DSM 21552, *Pseudonocardia* sp. P2, *Serinicoccus marinus* DSM 15273, *Serinicoccus profundi* MCCC 1A05965, *Sphaerobacter thermophilus* DSM 20745 y *Streptomyces* sp. CT34.

En Cyanobacteria, *Pseudomonas* y Archaea en contraste con Actinobacteria, se muestran una mezcla entre organismos que poseen una (verdes) o dos copias (amarillos) de PriA. Sin embargo al analizar detalladamente los árboles producidos por EvoMining en los distintos linajes, observamos que las copias extra tanto de Cyanobacteria como de *Pseudomonas* son en realidad miembros de otras familias enzimáticas, que guardan cierta similitud de secuencia con PriA. En Cyanobacteria y *Pseudomonas* la copia extra está en una rama divergente y muy poblada del árbol. En ambos casos esta segunda copia está en su mayoría anotada por RAST como imidazol glicerol fosfato sintasa ciclase. En Archaea sin embargo, diversas especies de la clase Methanomicrobia sí tienen dos copias de PriA.

Después de explorar cuáles organismos tienen expansiones de PriA, a continuación se muestra en la Figure 4.4 el posible destino metabólico de las copias extra de la familia. El árbol de Actinobacteria está poblado de hojas rojas Figure 4.4, es decir de PriA dedicadas al metabolismo conservado, en este caso relacionado a las rutas de Histidina y Triptofano. Sin embargo hay algunas hojas grises, como es el caso de los

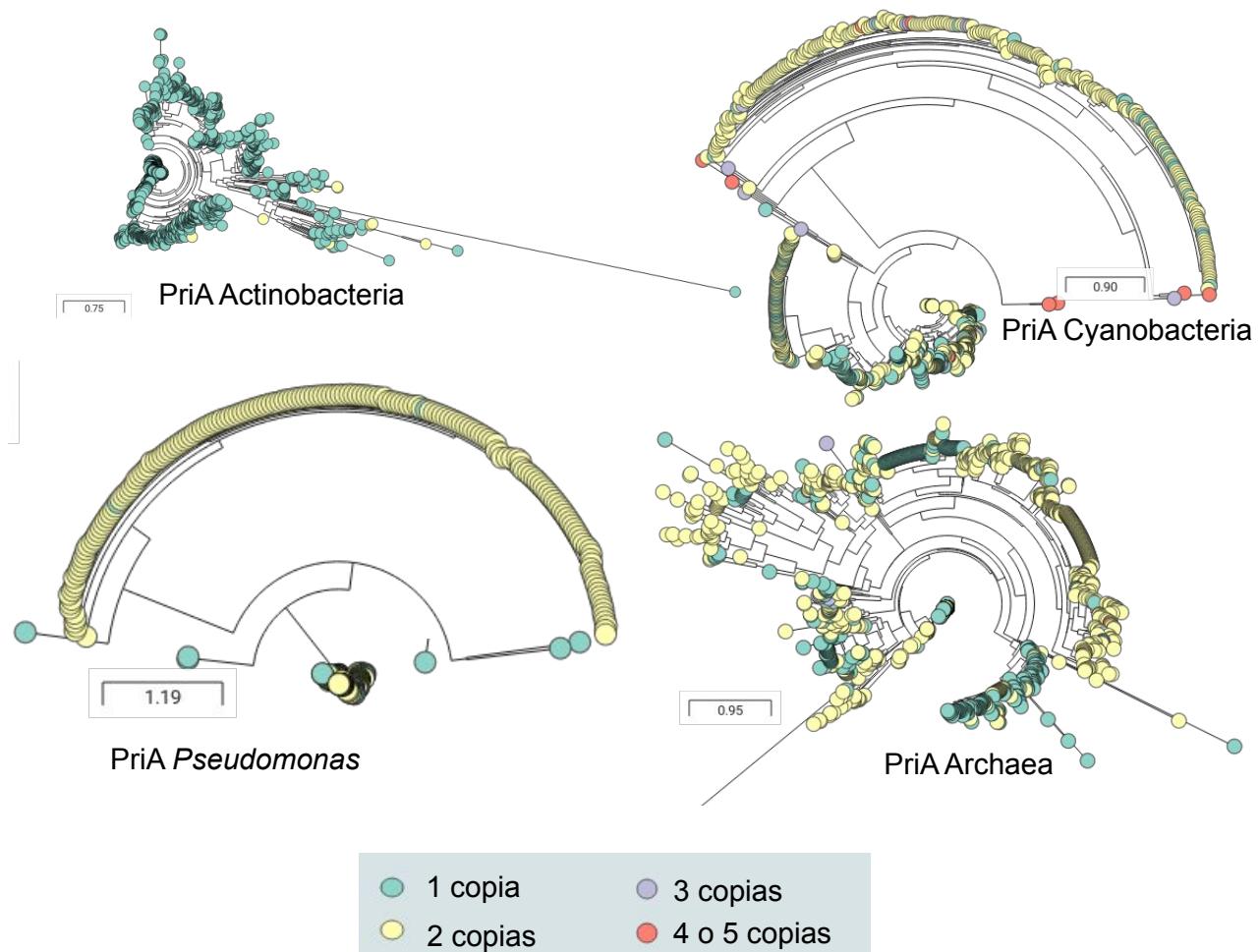


Figura 4.3: Número de copias de PriA en Actinobacteria, Cyanobacteria, Pseudomonas y Archaea. En Actinobacteria, donde se ha comprobado su promiscuidad enzimática la moda en el número de copias es uno (verde). En Cyanobacteria y Archaea se observa una mezcla de hojas verdes y hojas amarillas, donde las amarillas son nodos que pertenecen a organismos con dos copias. También en estos linajes algunos organismos poseen tres copias (morado) y cuatro o cinco (rojo). En Pseudomonas la mayoría de los organismos muestran dos copias

dos *Serenicoccus* mencionados previamente. Es posible que estas PriA puedan tener funciones alternativas. Además, en Actinobacteria la PriA de *Janibacter Hoiley*, la rama más larga del árbol, es muy divergente. Esto se debe a que existe una fusión de PriA con HisH. La fusión de genes bacterianos es un mecanismo común para formar proteínas multidominio [169]. De hecho, dentro del operón de histidina existen fusiones reportadas HisNB y HisIE con proteínas bifuncionales en ciertas proteobacterias [170]. La fusión aquí presentada no parece ser un artefacto de anotación ya que hay otros genomas de *Janibacter* con secuencias de PriA más grandes que el promedio. Los árboles que se produjeron por EvoMining están disponibles para exploración interactiva en Microreact en los links de la Table 4.1:

Cuadro 4.1: Árboles EvoMining de PriA en MicroReact

Linaje	Link al árbol de EvoMining en MicroReact
Actinobacteria	7g2IGfkv9
Cyanobacteria	qF6jWRMox
<i>Pseudomonas</i>	ydff6DWqs
Archaea	Ig-m9Cm6f

En Cyanobacteria hay pocas predicciones de EvoMining, lo que se muestra en escasas hojas verdes Figure 4.4 que además no están localizadas cerca de su reclutamiento la HisA de saxitoxin, el BGC proveniente de Cyanobacteria. Varias de las copias extra de Cyanobacteria corresponden a HisF, la subunidad enzimática imidazol glicerol fosfato sintasa parte también del operón de histidina. HisA proviene de una duplicación de un ancestro común de la mitad de su tamaño y HisF proviene de la duplicación de HisA [171]. Por ello puede considerarse que EvoMining en efecto está encontrando una expansión de HisA que consiste en la familia HisF en el linajes Cyanobacteria, lo mismo ocurre en *Pseudomonas*. En su momento HisF fue parte del *dispensable genome*, no todos los organismos poseían esta expansión, posteriormente, al parecer antes del último ancestro común de Archaea, Bacteria y Eukarya, la síntesis de histidina fue fijada [65]. Este es un caso donde lo que alguna vez fue *dispensable genome* / metabolismo secundario, se ha vuelto *shell genome* / metabolismo conservado y donde se muestra que en estos linajes Cyanobacteria y *Pseudomonas* EvoMining aún es capaz detectar esa marca de expansión. De hecho se ha reconstruido una HisF ancestral con el objetivo de probar si retenía la actividad de HisA, es decir si poseía promiscuidad enzimática [173]. LUCA-HisF mostró no ser promiscua para los sustratos de HisA ni tampoco para su similar el sustrato TrpF.

En *Pseudomonas* hay una gran población de predicciones de EvoMining, pero el árbol tiene similitudes con Cyanobacteria, la rama divergente corresponde en su mayoría a copias de HisF. De hecho, el reclutamiento que hace que el árbol de *Pseudomonas* tenga una rama verde está anotado funcionalmente no como una HisA, sino como una HisF parte de un BGC que produce un lipopolisacárido en la Proteobacteria *Legionella pneumophila*. En Archaea, la rama central con una mezcla de hojas grises y rojas

Figure 4.4 contiene también copias de HisF provenientes de genomas de *Sulfolobus*. Al verificar en el árbol interactivo en Microreact Table 4.1 el número de copias presente en organismos con un homólogo clasificado como HisF se comprobó que los *Sulfolobus* sólo poseen una copia en este árbol. Sin embargo, tanto al explorar manualmente los genomas de *Sulfolobus*, como al revisar la literatura sobre metabolismo de histidina en Archaea se encontró que sí se han encontrado homólogos de PriA en genomas de *Sulfolobus* [65]. Por tanto este árbol sugiere que aunque existe tanto PriA como su expansión HisF en el género *Sulfolobus*, la copia HisF es la más parecida a la semilla con que se generó este árbol. Además la similitud de secuencia del homólogo de HisA no fue suficiente para recuperar copias de HisA en este experimento computacional. Algunas de las hojas grises del árbol sin destino metabólico conocido serán exploradas en la siguiente sección.

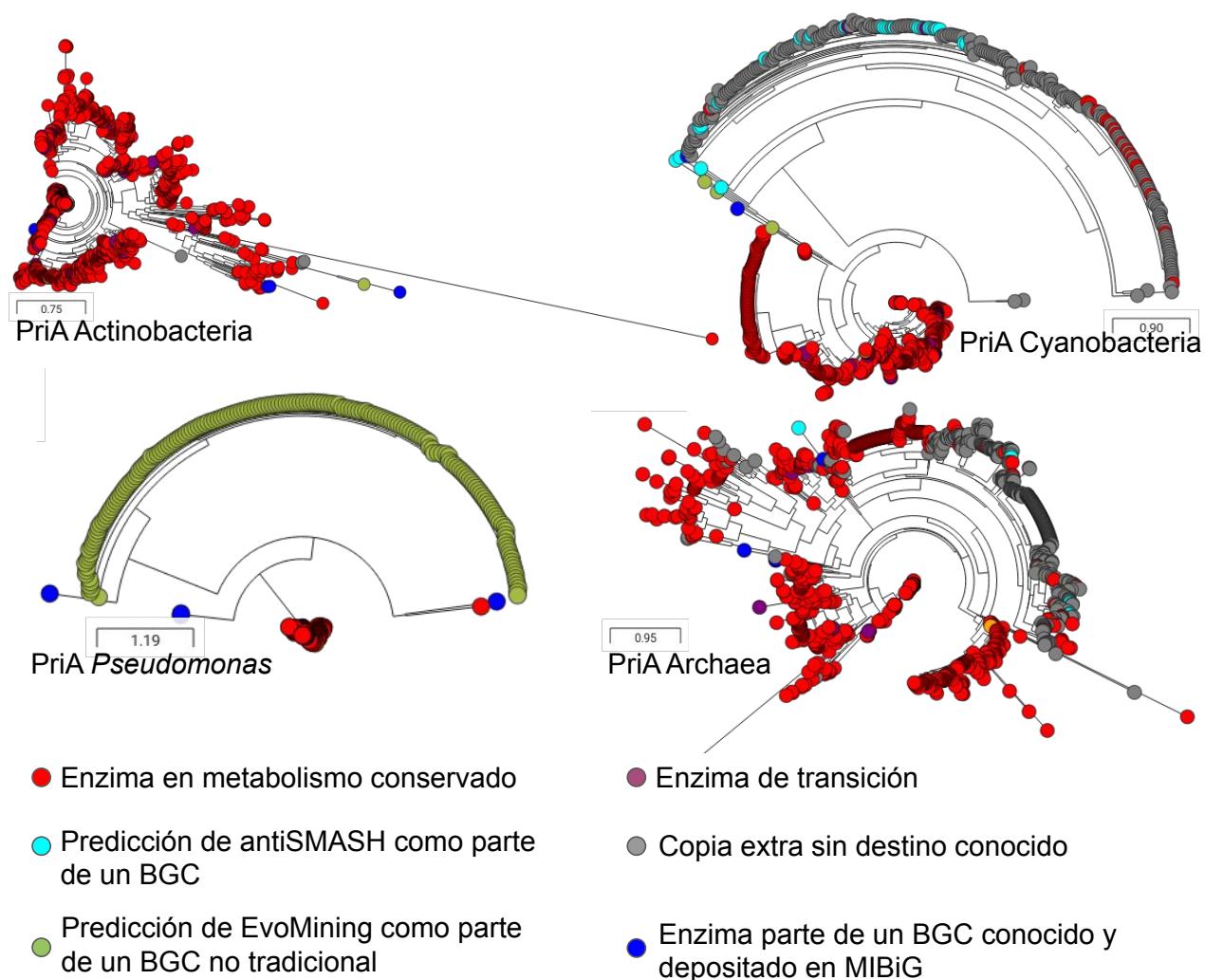


Figura 4.4: Árboles de destino metabólico de PriA en Actinobacteria, Cyanobacteria, Pseudomonas y Archaea según EvoMining

Los reclutamientos conocidos, marcados en azul en el árbol, que tuvo la familia expandida PriA/HisA en estos linajes genómicos están listados en la tabla Figure 4.1. Entre ellos se encuentran dos toxinas de Cyanobacteria [103], un lipopolisacárido producido por una Proteobacteria y un BGC productor de cloro-pentostatina producido en Actinobacteria [174].

Table: Reclutamientos de expansiones de PriA en MIBiG

Compuesto	Organismo	BGC origen	Clase
ada	<i>Actinomadura</i> sp. ATCC 39365	Actinobacteria	Otros
saxitoxin	<i>Cylindrospermopsis raciborskii</i> T3	Cyanobacteria	Alcaloide
toxin	<i>Dolichospermum circinale</i> AWQC131C	Cyanobacteria	Otros
lipopolisacárido	<i>Legionella pneumophila</i>	Proteobacteria	Sacárido

La pentostatina es un antibiótico nucleosídico derivado de adenosina, cuyo cluster es llamado *ada*. La PriA del cluster *ada* es llamada *adaK* y sí parece participar del cluster, ya que muestra una isomerización sobre un sustrato similar a los nativos de PriA, Figure 4.5, con un anillo de 5 carbonos, dos OH, un oxígeno y un grupo fosfato. Esta isomerización es muy parecida a la que realiza PriA sobre ProFAR y PRA. Esta PriA no es una copia extra, sino la copia única de este organismo. Este contexto genómico no se encuentra conservado en las copias vecinas en el árbol. Es relevante mencionar que la mutante de PriA no suprime la producción de este antibiótico en *Actinomadura* sp. ATCC 39365, por lo que los autores especulan que otra enzima podría estar realizando la isomerización redundantemente. *Actinomadura* no posee una copia de TrpF que sería el candidato inmediato.

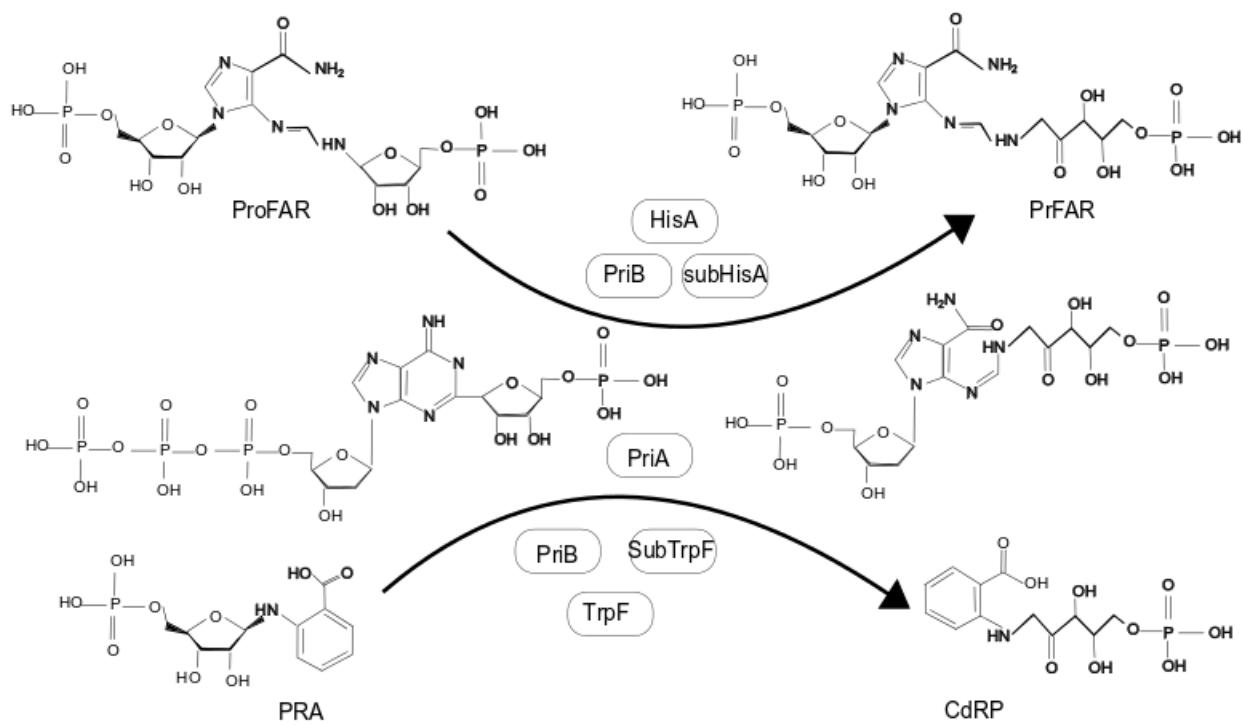


Figura 4.5: PriA participa en la síntesis del antibiótico ada en Actinomadura. Los sustratos nativos de PriA, ProFAR y PRA son isomerizados de manera muy similar a un paso en la ruta de síntesis de ada

## 4.2. Análisis de contextos genómicos de PriA/HisA en distintos linajes utilizando CORASON como herramienta de visualización.

### 4.2.1. Contextos genómicos de HisA en Actinobacteria

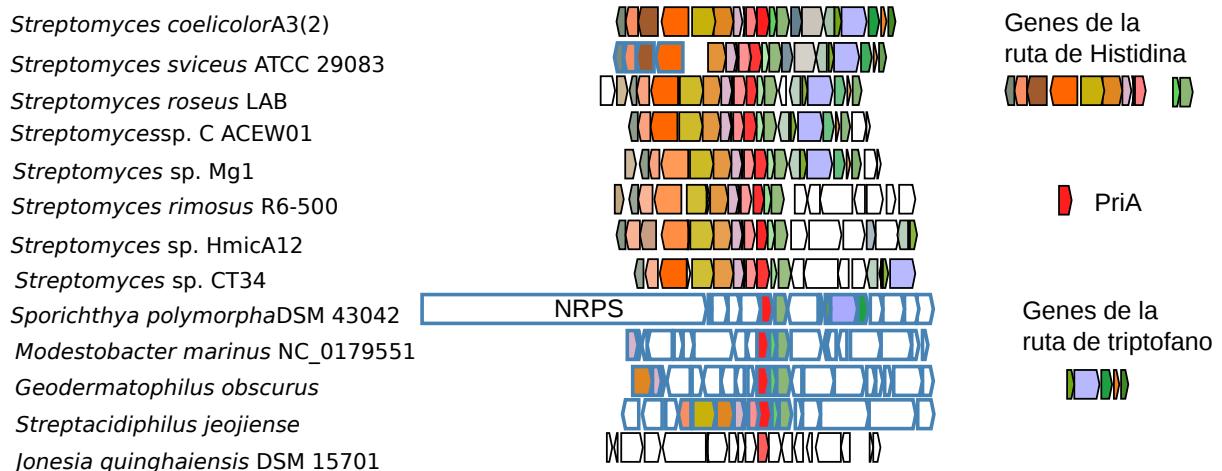


Figura 4.6: Contextos de PriA en Actinobacteria

En Actinobacteria, se observó que todos los *Streptomyces* tienen el cluster de PriA parcialmente conservado con respecto al BGC de *Streptomyces coelicolor*. Ejemplos de ello son *S. roseus*, *S. sviceus*, *S. sp C* y *S. Mg1* donde genes tanto de la ruta de histidina como de triptófano rodean a PriA. Otros como *S rimosus*, *S HmicA12* y *S. sp CT34* tienen los genes de triptófano más alejados Figure 4.6 . Como ya se describió en la sección de EvoMining, el único organismo de este género con una copia extra de PriA es *Streptomyces* CT34. Esta copia parece deberse a transferencia horizontal dado que su mejor hit en NCBI proviene de una *Lentzea*. Aún así parece ser un homólogo lejano ya que tuvo 50 % de identidad en 98 % de cobertura con respecto a la copia de *Lentzea*. Otro caso interesante en Actinobacteria son las Actinomaduras, ya que CORASON muestra que el cluster *ada* no está conservado en ellas (datos no mostrados). Además, también en Actinobacteria CORASON muestra en *Sporichthya polymorpha* DSM 43042 una PriA precedida por una NRPS, una enzima por excelencia de productos naturales. En otros organismos antiSMASH predice que PriA forma parte de clusters putativos, por ejemplo en *Modestobacter marinus* NC\_0179551, *Geodermatophilus obscurus*, y en *Streptacidiphilus jeojiense*. Los contornos de los genes reconocidos por antiSMASH como parte de un BGC son marcados en azul en las figuras de CORASON.

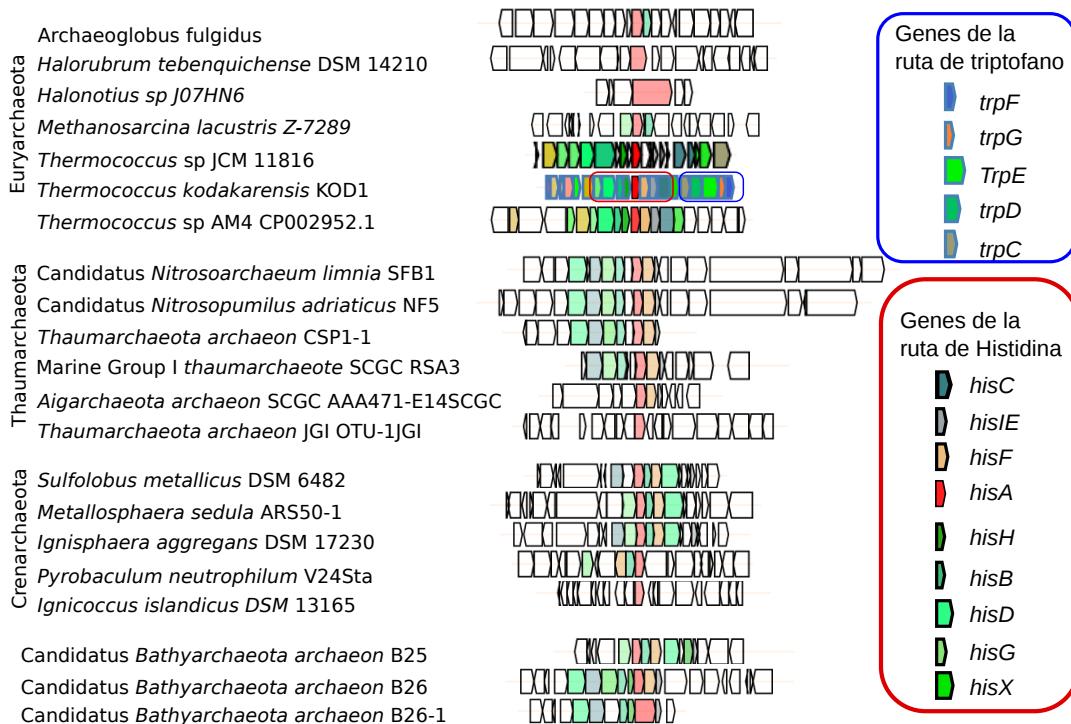


Figura 4.7: Contextos de PriA en Archaea

#### 4.2.2. Contextos genómicos de HisA en Archaea

En cuanto al contexto de HisA en Archaea la Figure 4.7 muestra que existen contextos como los de *Thermococcus kodakarensis KOD1* y *Thermococcus sp JCM 11816* donde *hisA* está rodeada de genes tanto de histidina como de triptofano. Sin embargo esta configuración conjunta no es la generalidad. En Archaea, se conocen tres escenarios para el operón de histidina, el primero es la no existencia del operón, con los genes de síntesis dispersos en el genoma. El segundo escenario es la existencia de suboperones pequeños como *hisHAF* y finalmente el tercer escenario es el operón con la mayoría de los genes juntos aunque no siempre en el mismo orden [65].

En efecto en Euryarchaeota están presentes los tres escenarios, en *Archaeoglobales* y *Halobacteriales* los genes del operón *his* están distribuidos en el genoma, como muestran los contextos de *hisA* en *Halorubrum tebenquichense* y *Halonotius sp.* sin ningún otro gen del operón cerca. En *Thermosarcinales* donde ya se conocía la existencia del suboperón *hisGBA*, se observó que éste se encuentra conservado en la mayoría de los miembros del orden como ejemplifica el contexto correspondiente a *Methanosaeca lacustris*. En *Thermococcales* se conocía la configuración *hisGDBHAFIEC* que es confirmada por los tres *Thermococcus* de la Figure 4.7, donde además se muestran dos contextos acompañados del su operón de triptofano *trpCDEGF*, resultando un contexto total de *hisGDBHAFIECXtrpCDEGF*. En el phylum Thaumarchaeota se habían estudiado sólo dos organismos y estos presentaban la configuración *hisGDCXBHAI*. En contraste en este trabajo encontramos en el orden *Nitrosopumilales* la configuración

*hisDCXBHAFIE*, donde además la *hisX* es una fosfatasa homóloga a la *hisX* presente en *T. kodakarensis*. Este patrón es común en Thaumarchaeota, aunque también se muestra la existencia del suboperón *hisAF* en *Aigarchaeota* así como la configuración de genes dispersos en el genoma en *Thaumarchaeota archaeon JGI OTU-1*. En Crenarchaeota *hisCGABFDEHI* está presente según la anotación funcional de RAST, aunque no todos los genes aparecen en color en la figura debido a la baja identidad de secuencia que conservan con respecto al contexto de referencia de *T. kodakarensis*. En *Pyrobaculum neutrophilum* está parcialmente conservado el operón con la configuración *hisCEDFBA*, nuevamente con algunos genes con baja identidad de secuencia respecto a *T. kodakarensis*. La configuración genes dispersos en el cromosoma en el phylum Crenarchaeota es mostrada en el contexto de *Ignicoccus islandicus*. Finalmente fue analizado el contexto de *hisA* en Bathyarchaeota un phylum donde no se contaba con información previa. En Bathyarchaeota se conserva como en Thaumarchaeota el ordenamiento *hisGDCXBHAI*, aunque existen otras configuraciones como muestra *Bathyarchaeota archaeon B25*.

Otra característica interesante de la visualización de los contextos de HisA en Archaea es que los tamaños de *hisA* son más variados que en Actinobacteria. Esta característica se debe tanto a variación intrínseca del tamaño de *hisA* como a diversas fusiones. Por ejemplo en *Halonotius sp J07HN6* *hisA* se encuentra fusionada con *flaJ* una proteína relacionada con la construcción de flagelos en Archaea. En *Candidatus Bathyarchaeota archaeon B26-1* *hisA* se muestra fusionada con *hisF*.

No existe una configuración conservada que prevalezca en todo Archaea. Esto puede deberse a que Archaea es un dominio, no un phylum como Actinobacteria, y por tanto hay una mayor distancia entre los organismos que la conforman.

#### 4.2.3. Contextos genómicos de HisA en saxitoxin

Como se mostró en la sección de análisis de expansiones de HisA con EvoMining en Cyanobacteria, *hisA* aparece dentro del cluster reportado como productor de saxitoxina en MIBiG. El contexto que rodea al reclutamiento de *hisA* en el cluster de saxitoxina no está conservado en Cyanobacteria según la visualización de contextos producida por CORASON. Tomando como semilla la secuencia de HisA del BGC de saxitoxina y como referencia el cluster de saxitoxina, se muestra que si bien los reguladores del BGC saxitoxina sí están conservados en la vecindad genómica de HisA en otros organismos, este no es el caso para las enzimas biosintéticas del *cluster*. Además en el lado izquierdo de la Figure 4.8 se muestra que la HisA del BGC saxitoxina no está ubicada en una rama de PriA divergente, al contrario está embebida en la parte más conservada. Por estas razones es posible que PriA esté en la orilla del BGC de saxitoxina y más bien no participe en la síntesis de este compuesto o bien que mantenga su función primaria y a la vez tenga una función aún desconocida teniendo un papel en la regulación de la producción de saxitoxina.

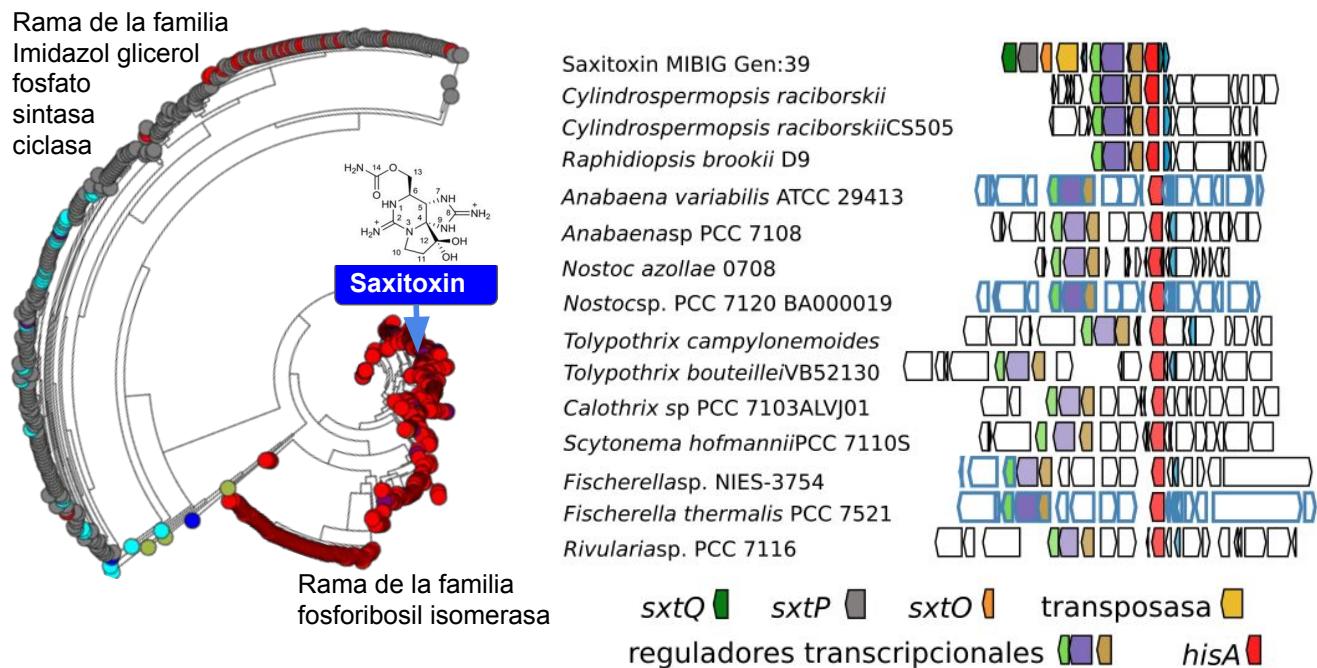


Figura 4.8: HisA en saxitoxin, un cluster de Cyanobacteria

### 4.3. Evolución molecular y estructural de PriA

En esta segunda sección concerniente a la familia PriA se busca información contenida en la secuencia de aminoácidos. En la primera parte se discute cómo en datos de evolución dirigida en el laboatorio no se encontró ninguna trayectoria en donde algún paso incrementara la actividad de PriA en sus dos sustratos nativos. En la segunda parte se muestra una reconstrucción de la estructura tridimensional de PriA basada en la covarianza de sus aminoácidos en secuencias del registro evolutivo.

#### 4.3.1. Al transformar una subHisA en una PriA mediante mutaciones no se observó ninguna trayectoria creciente para ambos sustratos (darwiniana)

En esta sección analizamos cómo cambia la capacidad catalítica de PriA sobre un sustrato mientras se varía la del otro. Para ello se utilizaron datos de mutantes de subHisA de *Corynebacterium diphtheriae*. Estas mediciones de cinéticas enzimáticas fueron obtenidos del trabajo de tesis de Lianet Noda [71]. A partir de la secuencia original que se mostró es una subHisA, se realizaron mutantes con el objetivo de alcanzar la promiscuidad, es decir de convertir la enzima subHisA en una PriA. Se comenzó con diferentes mutantes puntuales adicionando una mutación cada vez, hasta

llegar a una con 11 mutaciones. En esta colección de mutantes varias ganaron la función de PRA isomerasa, a distintos niveles. La que alcanzó mayor actividad PRA isomerasa fue la 9.3, una variante con nueve mutaciones. En estos datos quedaba pendiente la exploración de las rutas, es decir cómo es el camino desde una mutante sencilla hasta una múltiple ¿cuántas rutas son posibles? ¿Existe alguna tendencia en ciertos momentos de la ruta sobre el incremento/decremento de alguna de las dos funciones?

Así pues se desarrolló un programa utilizando recursividad para reconstruir todas las rutas posibles. El total de rutas calculadas hasta la mutación 11 fue de 2928 caminos, las rutas posibles hasta la mutante 9.3 son 790. Estas rutas son mostradas en la Figure 4.9.

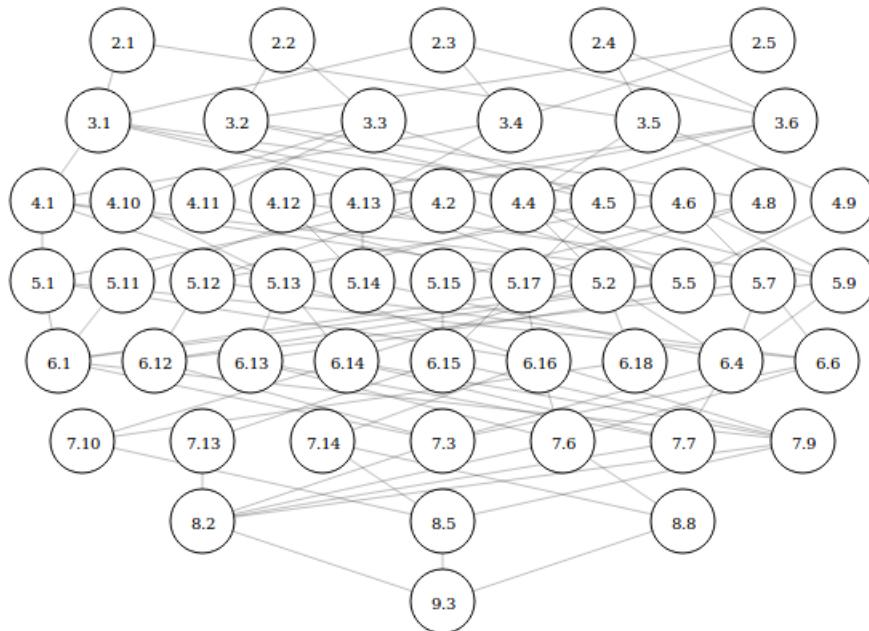


Figura 4.9: Rutas desde una subHisA hasta una variante con 9 mutaciones. En cada círculo el primer dígito indica el número de mutaciones

Al analizar los cambios sufridos en cada actividad en cada paso de cada ruta se descubrió que no existe en ellas una trayectoria no decreciente para ningún sustrato. Como ejemplo, en la Figure 4.9 se muestran las rutas donde cada mutante mantiene un nivel mínimo de actividad de ProFAR isomerasa ( $\frac{K_{cat}}{K_m} \text{PriA}_{\text{ProFAR}} \geq ,004$ ). En azul sólido se ven los incrementos en PRA y en rojo sólido los incrementos en ProFAR. Las líneas punteadas indican que la actividad decreció en ese paso de la ruta. Entre una y cuatro mutaciones el azul sólido es predominante, es decir se incrementa la actividad para PRA, pero entre 4 y 5 mutaciones ningún paso incrementa la actividad de PRA

y en cambio sí se incrementa la actividad para ProFAR, esta figura sugiere que al mejorar una actividad se compromete el mejoramiento de la otra. En este ejemplo, las mutaciones puntuales que llevan a una enzima monofuncional a adquirir promiscuidad no mantienen una tendencia no decreciente de principio a fin sobre ninguna ruta en ninguna de las dos reacciones isomerización de PRA e isomerización de ProFAR. Este tipo de trayectorias se conoce como no darwiniana ya que siempre existe algún paso donde decrece alguna de las actividades.

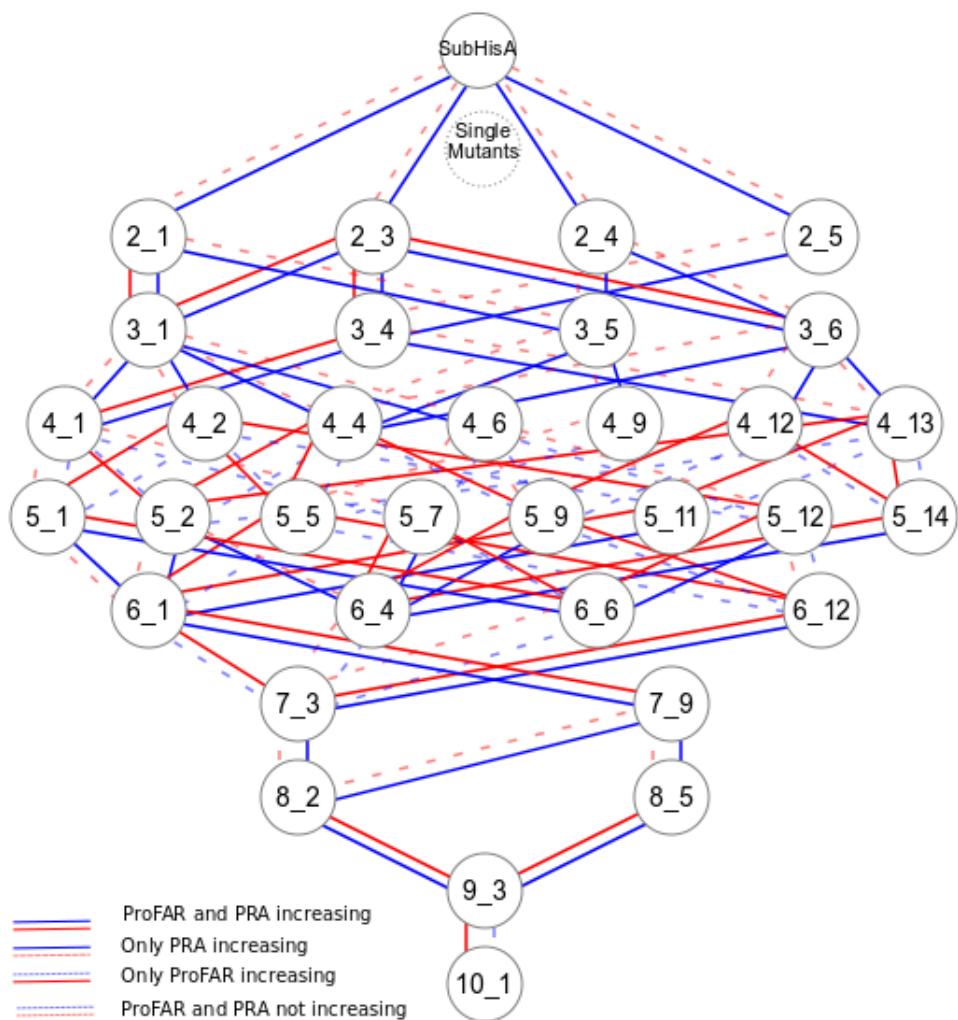


Figura 4.10: Rutas de evolución dirigida para ganar la función PRA. En la figura se muestra como la mayoría de los pasos que incrementan una función hacen decrecer la capacidad catalítica de la otra.

### 4.3.2. Los residuos que covariarán en el registro evolutivo de PriA permiten una reconstrucción aproximada de su estructura tridimensional

El estudio de la evolución de PriA en la sección anterior nos proporcionó el aprendizaje de que para adquirir una actividad el camino no es estrictamente creciente. En cambio, suele haber pasos donde alguna de las dos actividades baja. En esta sección utilizaremos el registro evolutivo resultado de millones de años, tomaremos miles de homólogos de PriA para inferir la estructura tridimensional de una secuencia. EVcouplings es un método que considera las secuencias génicas existentes como experimentos exitosos de la naturaleza. Con esta información obtiene la covariación entre pares de aminoácidos de las secuencias existentes en el registro evolutivo. Los pares fuertemente relacionados se denominan acoplamientos, estos acoplamientos a menudo están cerca físicamente en la estructura terciaria de la proteína. Se ha demostrado que muchas proteínas contienen suficientes acoplamientos distribuidos ampliamente en toda la secuencia, de forma que con ellos es posible la reconstrucción de su estructura tridimensional [175]. En esta sección se aplicará EVcouplings para reconstruir la estructura tridimensional de PriA. El RMSD es una técnica usual de comparación de estructuras.

Las diferencias a nivel estructural pueden amplificar la información proporcionada por variaciones a nivel de secuencia. Una métrica común de distancia entre dos estructuras de proteínas es el *root mean squared deviation* (RMSD). Esta distancia se obtiene después de alinear las estructuras calculando el promedio de las distancias al cuadrado entre los átomos de carbono primarios [176]. La información comprendida en las covariaciones ha sido suficiente en distintas proteínas para que el método EVcouplings genere estructuras con un error de 2.7–4.8 Å respecto a la estructura cristalográfica conocida [175]. Por este motivo se decidió implementar EVcouplings y aplicarlo a la familia PriA. Este método es de difícil instalación ya que requiere varias dependencias. Por ello, desarrollé un contenedor docker donde dependencias, software y base de datos quedan instalados. Este desarrollo fue incluido por los autores de EVcouplings como sugerencia de instalación. El contenedor docker implementa el *EVcouplings python framework* [177] que comprende cinco etapas para estudiar el análisis de coevolución de residuos de una familia de proteínas. Dichas etapas son i) Alineado, ii) análisis de acoplamiento, iii) plegamiento basado en acoplamientos iv) análisis de mutación y v) comparación con estructuras conocidas.

EVcouplings fue aplicado a la secuencia de PriA de *Streptomyces coelicolor* obtenida de la base de datos Uniprot con identificador HIS4\_STRCO. Los parámetros de EVcouplings se dejaron con su configuración inicial incluyendo el umbral de recuperación para las secuencias utilizadas en el alineamiento. Con este alineamiento, utilizando la información de los acoplamientos de sus aminoácidos, se obtuvo el modelo *PriAEV* de la estructura tridimensional de PriA de *S. coelicolor*.

Ahora bien, 1VZW es una estructura de PriA de *S. coelicolor* obtenida mediante cristalográfia [178]. En esta tesis calculé que el RMSD entre PriA (1VZW) y su homólogo lejano HisF (2A0N proveniente de *Thermotoga maritima*) es de 7.34 Å. Así

misma la distancia obtenida entre PriA (1VZW) y el modelo de TrpF 5LHE de *Thermococcus kodakaraensis* es de 12.088 Å. Por tanto cualquier modelo existoso de PriA debe tener un RMSD menor que estas distancias con respecto al modelo 1VZW. La distancia entre la estructura PriAEV y HisF es de 5.603 Å y el RMSD entre PriAEV y TrpF es de 12.554 Å. Además, la distancia entre estas 1VZW y PriAEV medida utilizando el RSMD (pymol 2.2.3) entre ellas es de 3.362 Å. Así pues PriAEV consigue diferenciarse de HisF y TrpF y estar a una distancia de 3.362 Å de 1VZW. Esta distancia es menor que el 3.730 Å el RMSD entre la estructura de PriA 1VZW de *S. coelicolor* y la estructura de PriA *Streptomyces sviceus ATCC 29083* con identificador de PDB 4TX9. Además, las estructuras cristalográficas tienen una resolución de 1-3 Å como puede verse en la Table 4.4, por tanto un RMSD de 3 Å es aceptable para decir que las estructuras son similares. Sin embargo los RMSD entre otras estructuras cristalográficas de *Streptomyces coelicolor* son menores a 2 Å, por ejemplo el (RMSD (2VEP, 1VZW) = 0.449 Å, RMSD (1VZW, 2X30)= 1.929 Å, RMSD (1VZW, 5DN1)=2.802 Å). Así pues EVcouplings obtuvo una estructura de PriA de *S. coelicolor* similar a la cristalográfica 1VZW, pero aparentemente no mejor que otras estructuras cristalográficas de *S. coelicolor*. La comparación de la estructura PriAEV lograda por EVcouplings con respecto a 1VZW es mostrada en la Figure 4.11.

La importancia de haber montado EVcouplings en el Laboratorio de Evolución de la diversidad metabólica es que ahora puede obtenerse un modelo tridimensional para cada secuencia de PriA disponible en horas, en contraste con otros métodos computacionales que pueden tardar días. Con estas secuencias se podría corroborar si diferenciar a nivel familia mediante la información de estructuras tridimensionales es posible para PriA en Actinobacteria. Sin embargo, es posible que el RMSD de la estructura de toda la proteína no sea suficiente para diferenciar por familia, ya sea porque las diferencias estructurales no son suficientes, porque los métodos de construcción de las estructuras necesiten mejorarse, porque se deba refinar la selección de proteínas en el alineamiento diferenciando entre secuencias conocidas de PriA, subHisA, PriB y subTrpF, o bien porque deban considerarse para el RMSD solo las regiones de la estructura con más diferencias conocidas incluyendo todos los átomos de esas regiones en lugar de sólo los átomos de carbono alfa. . En la siguiente sección mediante una colaboración se aborda un poco la dinámica molecular de la familia PriA que es un paso más allá de la comparación estructural.

Finalmente los aminoácidos utilizados en la evolución dirigida de la sección anterior fueron comparados con los provistos por EVcouplings como altamente partícipes en la covariación. Los 10 aminoácidos con más acoplamientos fueron 90L, 117V, 127V, 48W, 208I, 87D, 135T, 21V, 109E. El número indica la posición de la secuencia y la letra es el código de los aminoácidos. Por ejemplo 87D es la posición 87 el aminoácido ácido aspártico. Sólo el 21V es parte los aminoácidos mutados en el estudio previamente descrito. En la Table 4.2 se muestra en la primera columna los aminoácidos variados en el estudio de mutación dirigida en *Corynebacterium*, en la segunda columna el aminoácido correspondiente en la secuencia de *Streptomyces coelicolor* y finalmente su correspondiente acoplamiento más significativo.

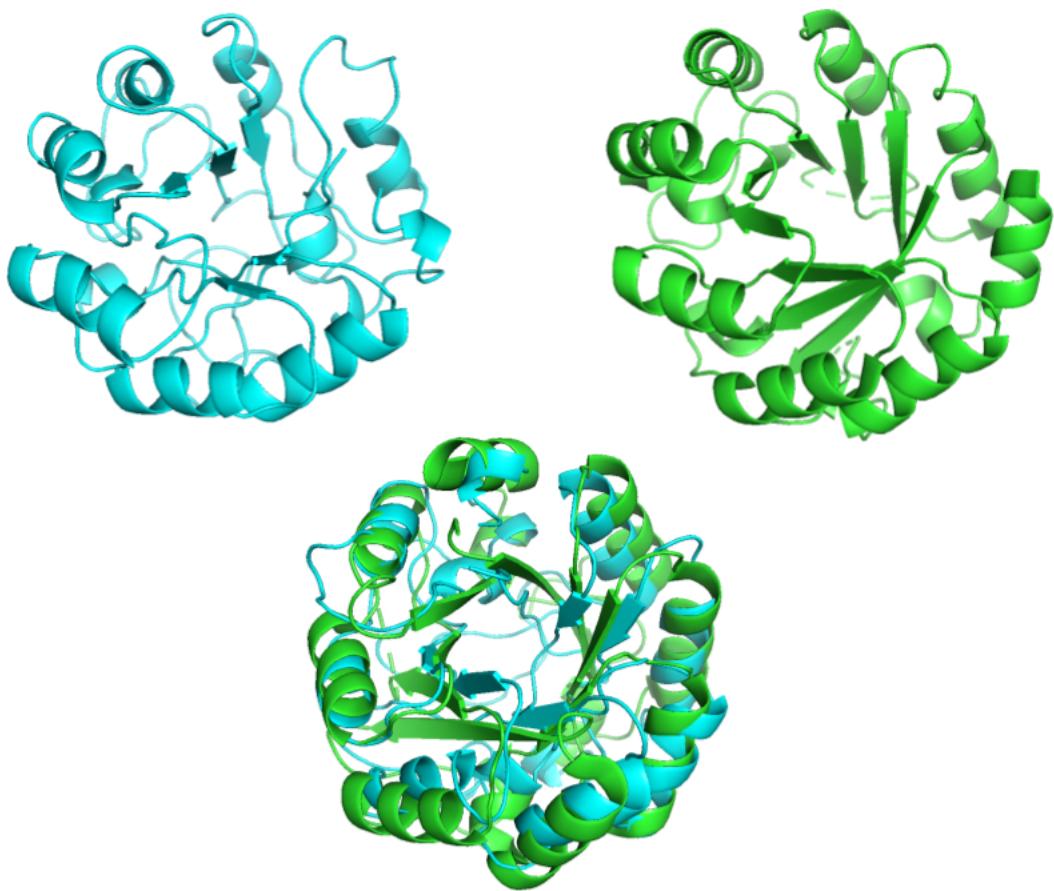


Figura 4.11: Comparación de estructura tridimensional de PriA generada por EVcouplings con una estructura cristalográfica de *Streptomyces coelicolor*. En azul se muestra la estructura generada considerando acoplamientos de los aminoácidos de PriA según el registro evolutivo. En verde está la estructura 1vzw obtenida experimentalmente. En la parte inferior de la figura se muestran las dos estructuras alineadas. Su RMSD es de 3.362 Å

Cuadro 4.2: Árboles EvoMining de PriA en MicroReact

<i>Corynebacterium</i>	<i>Streptomyces</i>	Acoplamiento más relevante
D20V	21V	173T
L48I	49L	76I
F50L	51L	70V
M66I	67I	80L
T80S	81S	102N
A97C	98C	107A
D127A	128G	164V
A129D	130D	168I
T139L	136L	151T
Y214L	211Y	-
E230A	227A	234E

Una forma gráfica de ver esta tabla es la Figure 4.12. En las líneas naranjas se muestra como el aminoácido 21D tiene un acoplamiento con el 173T, que está muy cerca del 175D que ha sido asociado con la actividad de isomerización de PRA, pero no de ProFAR. Además se investigaron los aminoácidos catalíticos D130 [179] y D175 reportados en *Mycobacterium tuberculosis*[67] que corresponden a en *S coelicolor* D11, D131 y D171. En este trabajo se econtró que existe un conjunto de secuencias donde estos posiciones tienen variación en dichos residuos Figure 4.13. Las secuencias más relevantes de PriA/HisA abordadas en este capítulo junto con secuencias seleccionadas divergentes en los residuos mencionados fueron alineadas para mostrar la diversidad a nivel aminoácidos de la familias PriA/HisA.

Secuencias fuera del phylum Actinobacteria, como *Pseudomonas*, Cyanobacteria, Enterobacteria, Proteobacteria y Chloroflexi muestran variantes del residuo 131D. Sin embargo, en Actinobacteria se encontraron 159 *Corynebacterium*, 49 *Streptomyces* y 2 *Actinokineospora* sin 131D. Entre los *Streptomyces* está el ya mencionado *Streptomyces* CT 34 y *Streptomyces rimosus*. En *Corynebacterium* se encuentra localizada la familia subHisA y en *Streptomyces* la familia PriB, la variabilidad mostrada en estos géneros podría estar relacionada con la existencia de estas familias. A futuro, para obtener resultados exclusivos sobre la covariación de residuos en Actinobacteria, se debe proveer un alineamiento exclusivo de Actinobacteria.

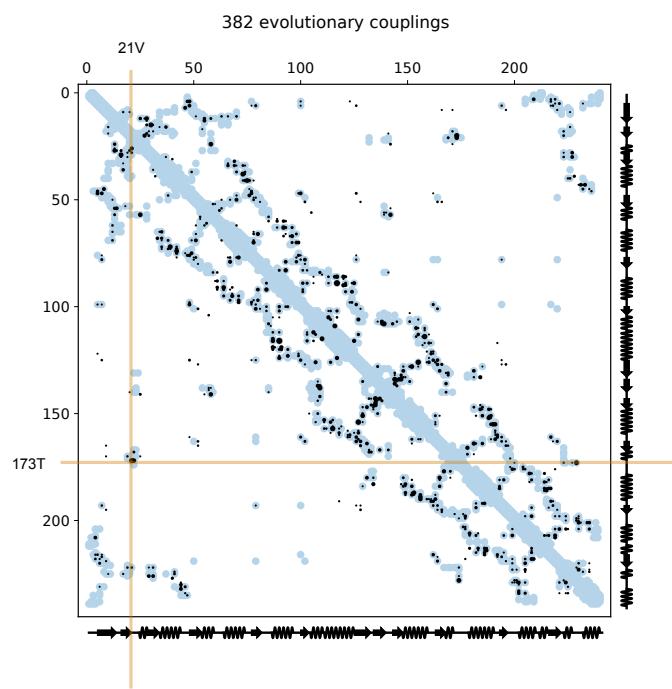


Figura 4.12: Visualización de los acoplamientos en la familia PriA

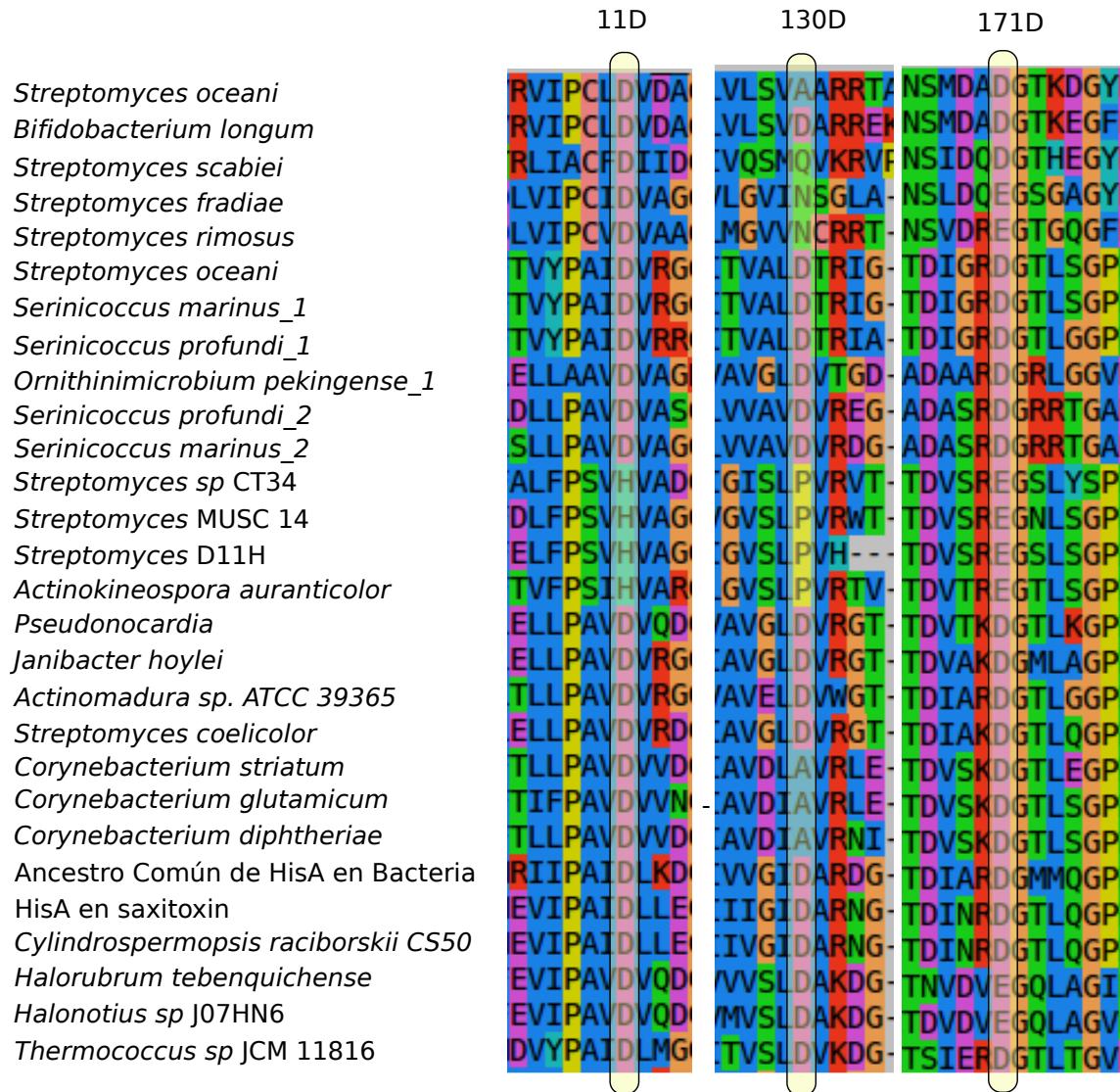
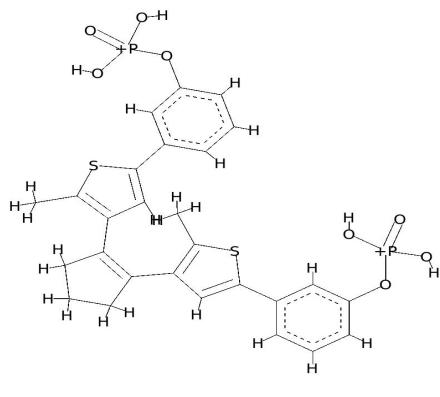


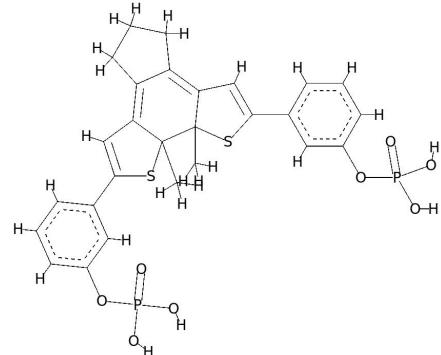
Figura 4.13: Miembros de PriA que poseen variantes los residuos catalíticos D11, D130

#### 4.4. Afinidad de enzimas selectas por sustratos químicamente parecidos a PRA y PROFAR

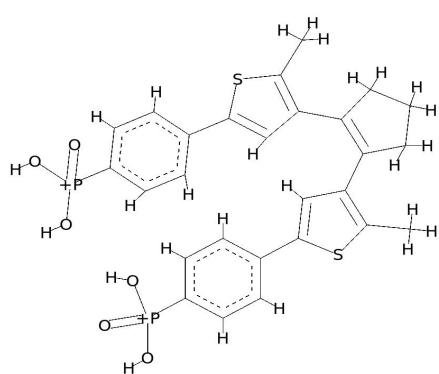
Además de los sustratos conocidos ProFAR y PRA en los que PriA es capaz de realizar una isomerización, es posible que PriA pueda ser promiscua en otros sustratos. De hecho, como se vio en la sección de EvoMining de este capítulo PriA parece participar en la síntesis del antibiótico pentostaina (*ada* BGC). Tomando este ejemplo como inspiración, se buscaron sustratos químicamente parecidos a ProFAR y PRA. Esta sección buscará probar la afinidad de sustratos parecidos a los nativos de PriA para posteriormente seleccionar probar alguno en copias selectas de PriA provenientes de diversos organismos. Veinte sustratos S1, S2, ... S20 sustratos fueron recolectados tanto de la literatura como de predicciones quimioinformáticas. Estos sustratos son mostrados en Figure 4.14, Figure 4.15,Figure 4.16,Figure 4.17. En este set S3 es el sustrato PRA y S7 el sustrato PROFAR, los sustratos nativos. S13-S16 son sustratos activados por luz. Los sustratos S17 (PRAP) y S18 (Compuesto V) se encontraron en la literatura, mientras que S6 (GMP), S11 (GTP) y otros fueron sugeridos por el grupo del Dr Chris Henry debido a sus similitudes quimioinformáticas. En la siguiente subsección, posteriormente a la selección de sustratos seleccionaremos las secuencias de PriA con las cuales se realizará el docking enzima-sustrato.



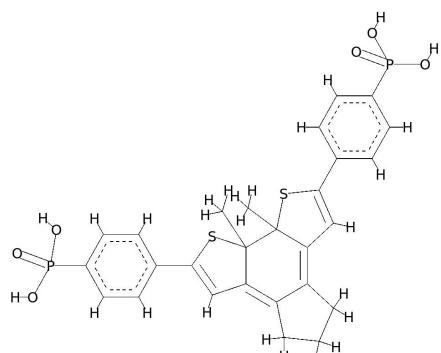
**13**  
DTE-metaphosphate(dte6\_open form)



**14**  
DTE-metaphosphate(dte6\_Closed form)



**15**  
DTE-Para-Phosphonate(dte13\_closed form)



**16**  
DTE-para-phosphonate(dte13\_closed form)

Figura 4.14: Substatos químicamente similares a los sustratos de PriA (parte 1)

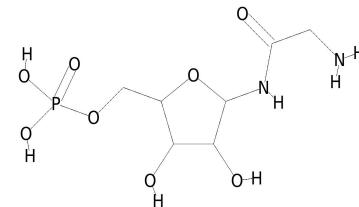
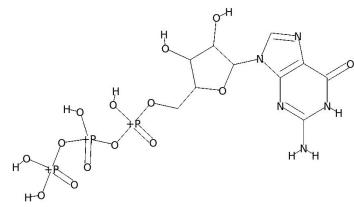
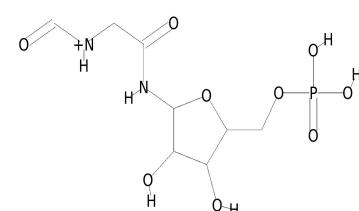
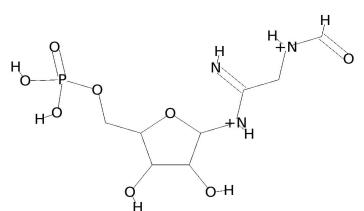
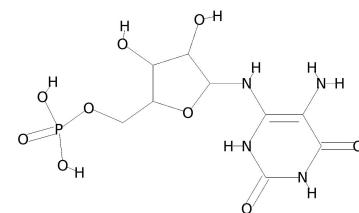
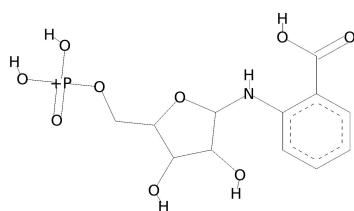
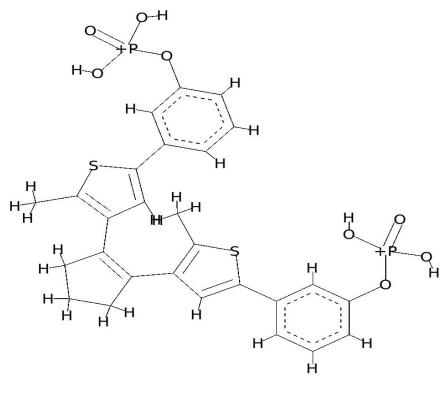
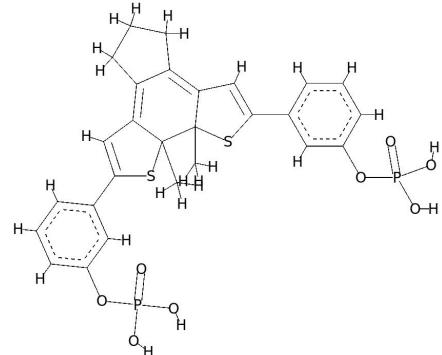


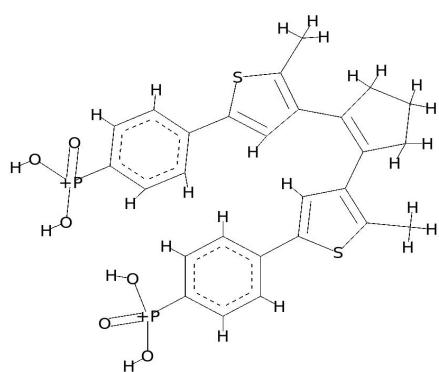
Figura 4.15: Substatos químicamente similares a los sustratos de PriA (parte 2)



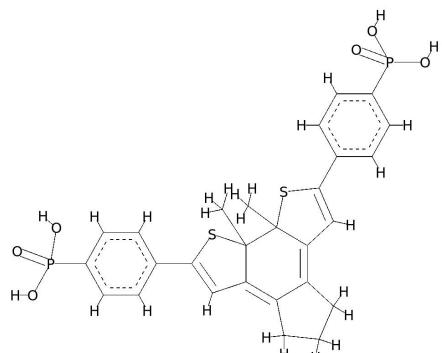
**13**  
DTE-metaphosphate(dte6\_open form)



**14**  
DTE-metaphosphate(dte6\_Closed form)

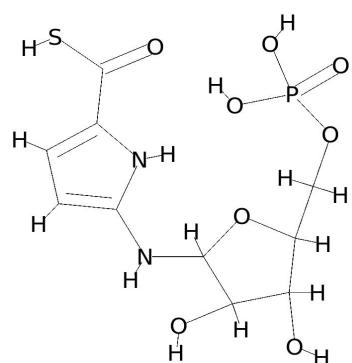


**15**  
DTE-Para-Phosphonate(dte13\_closed form)

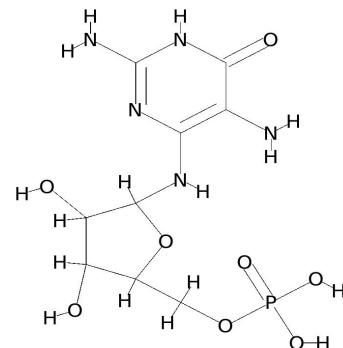


**16**  
DTE-para-phosphonate(dte13\_closed form)

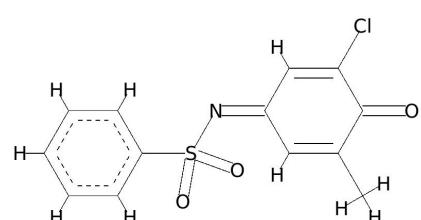
Figura 4.16: Substatos químicamente similares a los sustratos de PriA (parte 3)



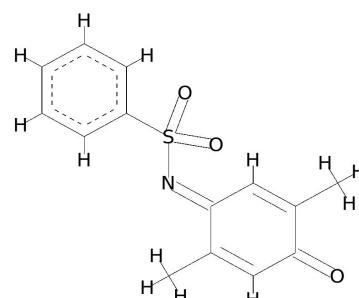
**17**  
4N'-(5'-phosphoribosyl) 4-aminopyrrole  
-2-carboxilate



**18**  
2,5-di-amino-6-ribosylamino-4  
(3H)-pyrimidinone 5'-phosphate



**19**  
(E)-N-(3-chloro-5-methyl-4-oxocyclohexa-  
2,5-dienylidene)benzenesulfonamide



**20**  
2,5 dimethyl-N-(4-oxocyclohexa-  
2,5-dienylidene)benzenesulfonamide

Figura 4.17: Substatos químicamente similares a los sustratos de PriA (parte 4)

#### 4.4.1. Selección de secuencias de PriA o familias relacionadas para docking con sustratos similares a lo nativos

Una vez escogidos los sustratos quedaban por definir las secuencias con las que se llevaría a cabo el análisis bioinformático de acoplamiento enzima-sustrato, en colaboración con el grupo del Dr Carillo-Trip. Para ello seleccioné 39 secuencias de la familia PriA y sus subfamilias, entre ellas varios *Streptomyces* con el objetivo de incluir tanto secuencias pertenecientes a PriA y como a la subfamilia PriB. Estas secuencias seleccionadas de PriA / PriB están uniformemente distribuidas en *Streptomyces* de acuerdo a un árbol de especies de *Streptomyces* basado en la proteína RpoB. Además estos *Streptomyces* tienen variedad en cuanto a la presencia / ausencia de *trpF* en su genoma. Se incluyeron además otros homólogos de PriA que han sido caracterizados químicamente y finalmente se con el mismo criterio de caraterización bioquímica se agregaron secuencias de HisA de *Escherichia coli*, *Arthrobacter Aurescens*, *Salmonella enterica* y *Acidimicrobium ferrooxidans* y secuencias TrpF provenientes *Jonesia denitrificans* y *Streptomyces sp Mg1* para ser utilizadas como controles.

Para realizar el docking se requieren estructuras cristalográficas. Cuando existían estructuras cristalográficas de la secuencia específica de PriA del organismo se utilizó dicha estructura. En caso contrario, el grupo del Dr Carrillo-Tripp generó estructuras homólogas utilizando la técnica de modelado por homología mediante el software Rosetta. Para cada secuencia decidí usar como plantilla la estructura tridimensional del homólogo de PriA más cercano que sí contara estructura con estructura cristalográfica. Los organismos con estructura cristalográfica de alguna familia relacionada a PriA están descritos en la Table 4.4. Para este análisis fueron las secuencias con estructura(s) cristalográfica(s) seleccionadas fueron: para la familia HisA la Enterobacteria *Salmonella enterica* (PDB:5AHE), para la familia subHisA la Actinobacteria *Actinomyces urogenitalis* (PDB:4X2R), la estructura que representa a subTrpF es la de *Arthrobacter aurescens* (PDB:4WD0). En cuanto a la familia PriA, varias estructuras cristalográficas de *Mycobacterium tuberculosis* (Mtub PDB:2Y88,2Y89,2Y85,3ZS4) y *Streptomyces coelicolor* (Scoe PDB:2VEP,2X30,1VZW) fueron incluidas. Las estructuras de PriB consideradas fueron las de *Streptomyces sviceus* (PDB:4U28,4TX9) y las de *Streptomyces sp Mg1* (4W9T, 4X9S). Finalmente la estructura cristalográfica de TrpF corresponde a *Jonesia denitrificans* (PDB:4WUI). Table 4.3.

Se incluyeron también enzimas TrpF provenientes de *Streptomyces Mg1*, *Jonesia denitrificans*. Las estructuras cristalográficas disponibles de PriB provienen de . . *Streptomyces globisporus*, *Actinomyces urogenitalis* (4X2R) y *Corynebacterium jeikeum*. De la familia subHisA se muestran *Corynebacterium efficiens* y . y finalmente, *Chlamidya thrachomatis*, *Streptomyces sp. Mg1* y *Actinomyces odontolyticus*. Las representantes de la familia TrpF son las Actinobacterias *Jonesia denitrificans* y *Streptomyces sp Mg1*. A continuación se muestra una tabla con las estructuras cristalográficas disponibles de PriA y familias relacionadas.

Cuadro 4.3: Secuencias selectas de PriA y familias relacionadas para el análisis de docking

Abreviatura	Organismo
Save	<i>Streptomyces avellaneus</i>
Scar	<i>Streptomyces carneus</i>
Spur	<i>Streptomyces purpeofuscus</i>
Smeg	<i>Streptomyces megasporus</i>
Sfrad	<i>Streptomyces fradiae</i>
Svar	<i>Streptomyces varsoviensis</i>
Satra	<i>Streptomyces atratus</i>
Srim	<i>Streptomyces rimosus</i> R6-500
SCT34	<i>Streptomyces sp.</i> CT34
S1813	<i>Streptomyces sp.</i> NRRL S-1813
Ssul	<i>Streptomyces sulphureus</i> DSM 40104
Scla	<i>Streptomyces clavuligerus</i> ATCC 27064
Stsu	<i>Streptomyces tsukubaensis</i> NRRL18488
Sbik	<i>Streptomyces bikiniensis</i>
Sven	<i>Streptomyces venezuelae</i> ATCC 10712
Sful	<i>Streptomyces fulvoviridis</i>
Scal	<i>Streptomyces californicus</i>
Sbaa	<i>Streptomyces baarnensis</i>
Salb	<i>Streptomyces albus</i> J1074
Siak	<i>Streptomyces iakyruis</i>
Sgha	<i>Streptomyces ghanaensis</i> ATCC 14672
Sbic	<i>Streptomyces bicolor</i>
Sipo	<i>Streptomyces ipomoeae</i> 91-03
Sbot	<i>Streptomyces bottropensis</i> ATCC 25435
SspC	<i>Streptomyces sp.</i> C
SMg1	<i>Streptomyces sp.</i> Mg1
Sxan	<i>Streptomyces xanthophaeus</i>
Skat	<i>Streptomyces katrae</i>
Slav	<i>Streptomyces lavendulae</i> subsp. <i>lavendulae</i>
Sery	<i>Streptomyces erythrochromogenes</i>
Sniv	<i>Streptomyces niveus</i> NCIMB 11891
Ssvi	<i>Streptomyces sviceus</i> ATCC 29083
Sfla	<i>Streptomyces flaveus</i>
Save	<i>Streptomyces avermitilis</i> MA-4680 = NBRC 14893
Scoe	<i>Streptomyces coelicolor</i> A3(2)
ArTC1	<i>Arthrobacter aurescens</i> TC1
Sglob	<i>Streptomyces globisporus</i> C-1027
Sgri	<i>Streptomyces griseolus</i>
S34	<i>Streptomyces sp.</i> CT34 (paralogo)

Cuadro 4.4: Estructuras cristalográficas disponibles de PriA y familias relacionadas

Organismo	PDB	Familia	Resolución Å	Fecha
<i>Salmonella enterica</i>	5AHE	HisA	1.7	2015
<i>Salmonella enterica</i>	5AB3	HisA	1.8	2016
<i>Salmonella enterica</i>	5ABT	HisA	1.65	2016
<i>Salmonella enterica</i>	5AC7	HisA	1.9	2016
<i>Salmonella enterica</i>	5AC8	HisA	1.7	2016
<i>Salmonella enterica</i>	5AC6	HisA	1.99	2016
<i>Salmonella enterica</i>	5A5W	HisA	1.6	2015
<i>Salmonella enterica</i>	5AHF	HisA	2.2	2012
<i>Thermotoga maritima</i>	2W79	HisA	1.85	2008
<i>Thermotoga maritima</i>	1QO2	HisA	1.85	2000
<i>Streptomyces sp.</i> Mg1	4X9S	PriB	1.6	2014
<i>Streptomyces sviceus</i>	4TX9	PriB	1.6	2014
<i>Streptomyces sviceus</i>	4U28	PriB	1.33	2014
<i>Streptomyces sp.</i> Mg1	4W9T	PriB	1.57	2014
<i>Arthrobacter aurescens</i>	4WD0	subTrpF	1.5	2014
<i>Streptomyces coelicolor</i>	5DN1	PriA	1.95	2015
<i>Streptomyces coelicolor</i>	1VZW	PriA	1.8	2004
<i>Streptomyces coelicolor</i>	2VEP	PriA	1.8	2007
<i>Streptomyces coelicolor</i>	2X30	PriA	1.95	2010
<i>Mycobacterium tuberculosis</i>	2Y85	PriA	2.4	2011
<i>Mycobacterium tuberculosis</i>	2Y88	PriA	1.33	2011
<i>Mycobacterium tuberculosis</i>	2Y89	PriA	2.5	2011
<i>Mycobacterium tuberculosis</i>	3ZS4	PriA	1.9	2012
<i>Actinomyces urogenitalis</i>	4X2R	SubHisA	1.05	2014
<i>Corynebacterium efficiens</i>	4AXK	SubHisA	2.25	2013
<i>Thermococcus kodakaraensis</i>	5LHE	TrpF	1.85	2016
<i>Thermococcus kodakaraensis</i>	5LHF	TrpF	1.75	2016
<i>Thermus thermophilus</i>	1V5X	TrpF	2	2003
<i>Thermotoga maritima</i>	1DL3	TrpF	2.7	1999
<i>Thermotoga maritima</i>	1LBM	TrpF	2.8	2002
<i>Thermotoga maritima</i>	1NSJ	TrpF	2	1996
<i>Jonesia denitrificans</i>	4WUI	TrpF	1.09	2014
<i>Pyrococcus furiosus</i>	4AAJ	TrpF	1.75	2012

#### 4.4.2. El análisis de PriA a nivel estructural sugiere que GTP es el sustrato más afín

Con las estructuras de las enzimas seleccionadas de PriA se realizaron simulaciones de docking respecto a los veinte sustratos descritos. El procedimiento puede ser consultado

en Docking Protocols. En la Figure 4.18 presento una visualización de los resultados obtenidos. En esta figura se marcan en color los organismos de los que se tiene una caracterización bioquímica, en azul están las secuencias correspondientes a TrpF, en rojo las de HisA, en rosa las secuencias de PriA y subHisA y en verde las de PriB y subTrpF. Con un asterisco fueron marcadas las secuencias de las que se cuenta con estructura cristalográfica. En el caso de las secuencias que no contaban con estructura cristalográfica, la estructura que sirvió como base está anotada después de la abreviatura del organismo. Las abreviaturas corresponden a los organismos de la Table 4.3. Los cuadrados grises son datos que faltaron de simular. Entre más azul es mayor la afinidad del sustrato por la secuencia, entre más rojo menor la afinidad de la enzima por este sustrato. Se requieren más controles para este experimento, pero una tendencia general es que la columna correspondiente a S11 (GTP) es el sustrato que tiene mayor afinidad. Por ello, aunque afinidad no tiene porque corresponder con capacidad catalítica, se decidió investigar la capacidad de secuencias de PriA para catalizar alguna reacción sobre el GTP, esa investigación es el objeto de la última sección de este capítulo.

#### 4.4.3. GTP es el sustrato por el que PriA muestra mayor afinidad

```
## boxplot de los sustratos
ggplot(docking.m, aes(x=variable, y=value)) + labs(x = "Substrates", y = "Affinity", text
```

Warning: Removed 100 rows containing non-finite values (stat\_boxplot).

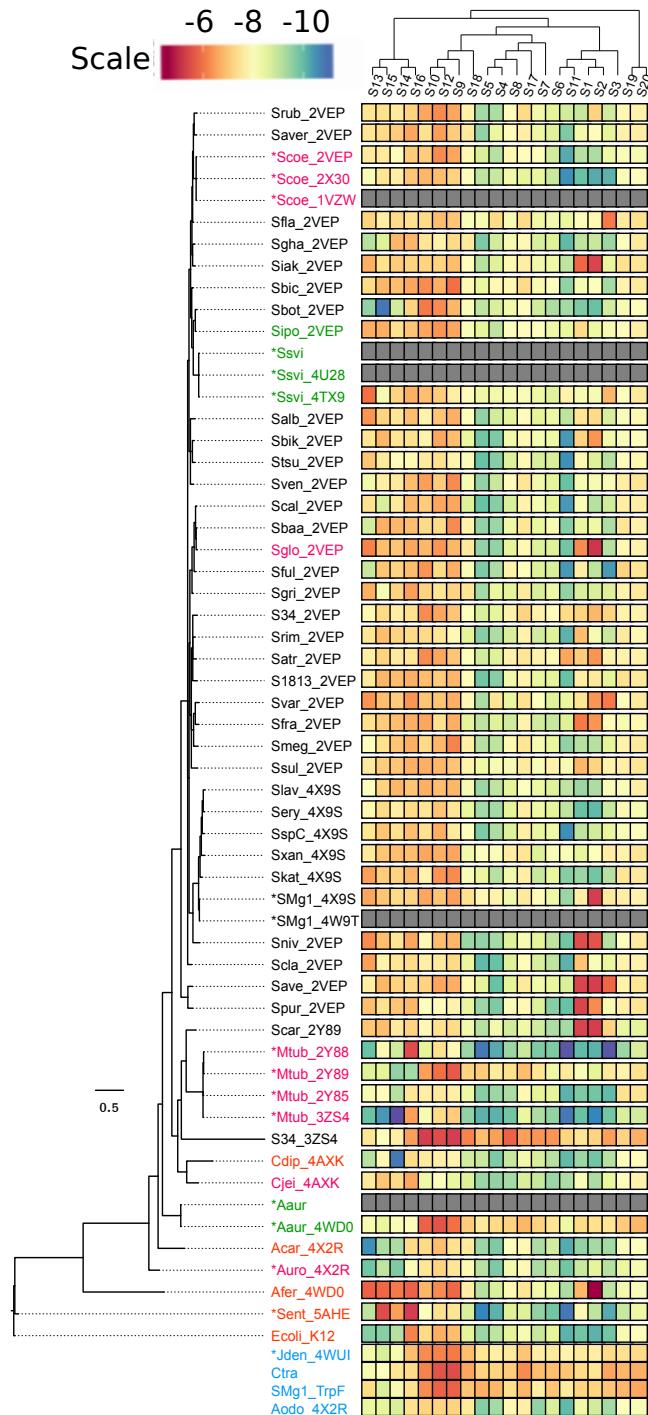
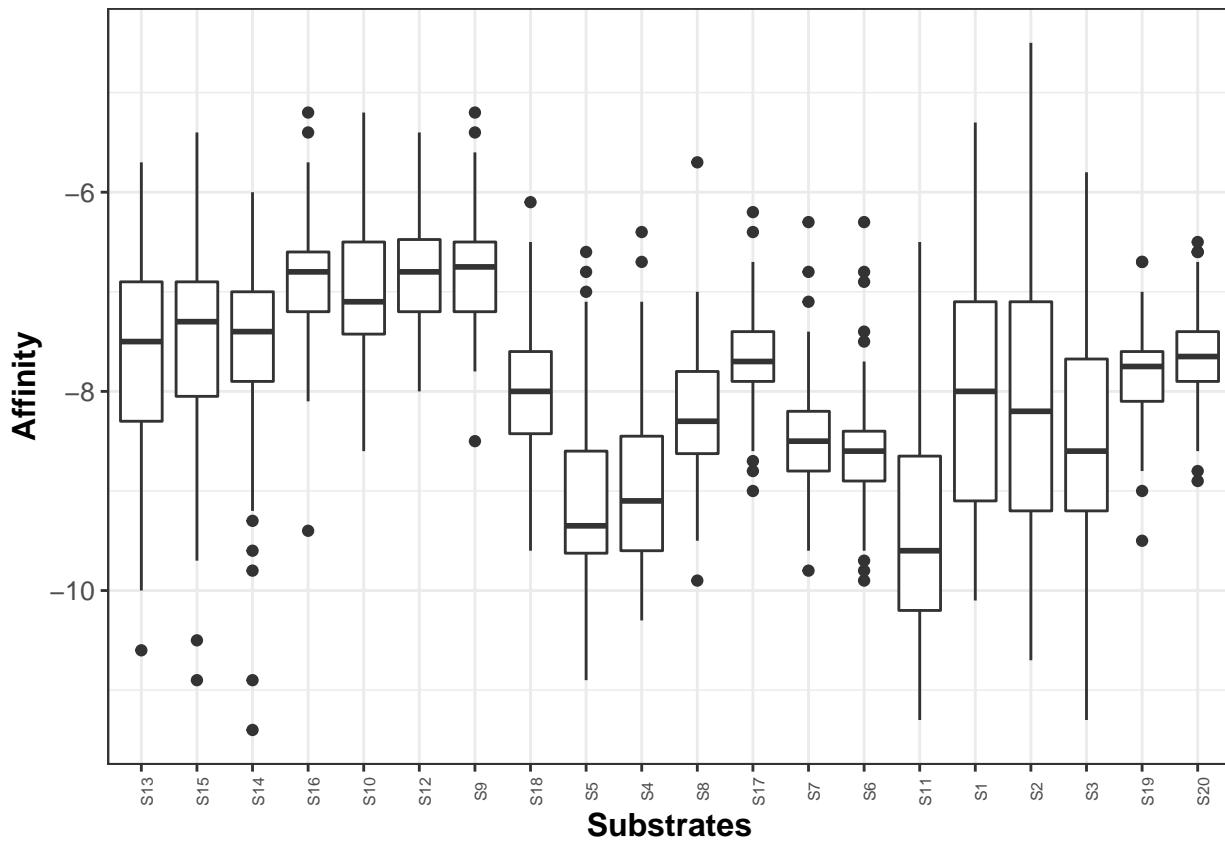


Figura 4.18: Heatplot del docking de enzimas relacionadas a PriA vs posibles substratos



#### 4.4.4. Dinámica molecular vs datos experimentales

En las tablas Subsection 4.4.4 , Subsection 4.4.4 se muestra a detalle la comparación entre las dinámicas moleculares de PriA respecto a PRA y ProFAR. Estos datos sugirieron que GTp es el sustrato con mayor afinidad y por ello en la última sección se avanzó en la realización de cinéticas enzimáticas con GTP.

Table: Actividad de PriA en S3 PRA

organism	Family	$K_M$	$k_{cat}$	$\frac{k_{cat}}{K_M}$	Pre MD	Pos MD	Reference
Afer	HisA	$1,1 \pm 0,2$	$0,05 \pm 0,001$	0.045	-10.1	-12.3	Noda-García L et al 2015
Ecoli	HisA	1,6	4,9	3.1	-9.9	-16	Henn-Sax et al. (2002)
Sent	HisA	$17,0 \pm 0,1$	$7,8 \pm 2,4$	$4,5 \times 10^5$	-10.3	-20.1	Söderholm A et al (2015)
Aaur	PriB	$2,1 \pm 0,5$	$1,8 \pm 0,2$	0.9	-7.4		verduzco-castro 2016
Sipo	PriB	$3,8 \pm 0,2$	$0,82 \pm 0,02$	0.21	-8.2	-14.7	verduzco-castro 2016
SspC	PriB	$11,4 \pm 3,4$	$2,53 \pm 0,74$	0.22	-8.5	-12.7	verduzco-castro 2016
SMg1	PriB	$13,2 \pm 3,4$	$0,92 \pm 0,19$	0.069	-8	-15.2	verduzco-castro 2016
Ssvi	PriB	$3,9 \pm 0,89$	$0,69 \pm 0,04$	0.18	-8.2	-16.7	verduzco-castro 2016
Scoe	PriA	$3,6 \pm 0,7$	$1,3 \pm 0,2$	0.4	-8.4	-15	Noda-García et al (2010)
Sglob	PriA	$4,2 \pm 0,8$	$0,74 \pm 0,03$	0.18	-9.2	-16.7	verduzco-castro
Mtub 2Y85	priA	190,23	0,012 – 9,7				Due et al 2011
Mtub 3ZS4	priA	?	-9,9				Due et al 2011 (To be published)
Auro	priA	$4,0 \pm 0,9$	$0,2 \pm 0,03$	0.04	-9.2		Vazquez-Juarez (2016)
Cjei	PriA	$2,3 \pm 0,2$	$0,9 \pm 0,08$	0.39	-8.5		Noda-García et al (2013)
Cdip	subHisA	$4,4 \pm 0,5$	$2,6 \pm 0,3$	0.59	-9.2		Noda-García et al (2013)
SMg1 TrpF	TrpF3	-	-	-	-6.9	-9.6	verduzco-castro 2016
Jden	TrpF3	-	-	-	-7.2	-9.4	$16,8 \pm 3,3$
Acar	SubHisA	0.02					Verduzco-Castro E et al 2016
Aodo	SubTrpF	-	-	-			

La siguiente tabla es análoga a la anterior pero con el sustrato ProFAR

Table: Actividad de PriA en S7 ProFAR

organism	Family	$K_M$	$k_{cat}$	$\frac{k_{cat}}{K_M}$	Pre MD	Pos MD	Reference
Afer	HisA	- -	-	-9.2	-9		Noda-García L et al. (2015)
Ecoli	HisA	- -	-	-9	-11.1		Henn-Sax et al. (2002)
Sent	HisA	- -	-	-9.6	-10.2		Söderholm A et al (2015)
Aaur	PriB	$26,3 \pm 6,3$	$0,37 \pm 0,09$	0.014	-7.1	-	verduzco-castro 2016
Sipo	PriB	$60,8 \pm 1,1$	$8,25 \pm 0,4$	0.14	-8	-8.5	verduzco-castro 2016
SspC	PriB	$149,9 \pm 29$	$1,4 \pm 0,12$	0.009	-8.5	-10.8	verduzco-castro 2016
SMg1	PriB	$129,6 \pm 34$	$0,29 \pm 0,04$	0.0022	-7.5	-11	verduzco-castro 2016
Ssvi	PriB	$24,5 \pm 4,0$	$1,6 \pm 0,29$	0.067	-8	-9.7	verduzco-castro 2016
Scoe	PriA	$5,0 \pm 0,08$	$3,4 \pm 0,09$	0.7	-8	-9.4	Noda-García et al (2010)
Sglob	PriA	$11 \pm 1,0$	$3,8 \pm 0,2$	0.34	-8.7	-9.4	verduzco-castro 2016
Mtub2Y85	priA	21	3.6	0.17	-8.6		Due et al 2011
Mtub3ZS4	priA				-9.3		Due et al 2011 (To be published)
Auro	priA	$23 \pm 6,5$	$0,5 \pm 0,05$	0.02	-9.3		Vazquez-Juarez (2016)
Cjei	PriA	$5,1 \pm 1,0$	$1,6 \pm 0,16$	0.31	-9		Noda-García et al (2013)
Cdip	subHisA	- -	-	-8.8			Noda-García et al (2013)
SMg1 TrpF	TrpF3	$8,4 \pm 1,7$	$10,5 \pm 2,4$	1.25	-7.6	-9	verduzco-castro
Jden	TrpF3	$16,8 \pm 3,3$	$27 \pm 1,6$	1.6	-7.6	-7.7	verduzco-castro
Acar	SubHisA	Na	Na	0.02	Na	Na	Na
Aodo	SubTrpF	- -	-	-	Na	Na	Na

Con actividad de FolE i.e activa para el compuesto V Adams et al (2014)

Cuadro 4.5: Enzymes docking

Enzima	S13	S15	S14	S16	S10	S12	S9	S18	S5	S4	S8	S17	S7	S6	S11	S1	S2	S3	S19	S20
Subr_2VEP	-7.4	-7.3	-7.5	-7.1	-6.5	-6.2	-6.5	-7.7	-9.4	-9.3	-7.9	-7.2	-8.3	-8.6	-8.9	-9.0	-7.1	-8.9	-7.8	-7.7
Saver_2VEP	-7.4	-7.2	-7.0	-6.5	-7.3	-6.4	-7.0	-7.5	-9.6	-8.5	-7.9	-7.6	-8.4	-8.7	-9.8	-8.3	-7.9	-8.6	-7.7	-7.6
Scoe_2VEP	-7.5	-7.5	-7.9	-7.0	-7.0	-6.2	-6.5	-7.8	-8.8	-9.2	-7.8	-7.9	-8.0	-8.9	-10.3	-9.2	-9.3	-8.4	-8.1	-8.2
Scoe_2X30	-8.1	-7.4	-7.6	-6.9	-6.7	-6.8	-7.1	-7.9	-9.1	-9.0	-8.3	-8.6	-8.5	-9.0	-10.6	-10.0	-10.3	-10.2	-8.1	-7.9
Scoe_1VZW	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Sfla_2VEP	-7.1	-7.6	-7.2	-7.3	-7.2	-6.8	-6.9	-8.1	-8.2	-7.2	-8.2	-7.2	-8.4	-8.3	-8.5	-7.9	-7.8	-6.0	-7.5	-7.3
Sgha_2VEP	-9.2	-8.7	-6.7	-6.7	-7.4	-7.7	-7.2	-7.5	-9.8	-8.9	-8.2	-7.8	-8.8	-8.7	-10.1	-9.1	-9.0	-9.5	-8.0	-7.5
Siak_2VEP	-6.6	-7.3	-7.0	-7.1	-7.1	-6.8	-8.0	-9.2	-8.7	-7.8	-7.6	-8.3	-8.4	-9.1	-5.8	-5.3	-8.5	-7.3	-7.4	
Sbic_2VEP	-7.2	-6.7	-6.8	-6.5	-6.2	-6.6	-5.9	-7.8	-8.5	-7.8	-7.8	-7.2	-8.2	-8.0	-9.6	-8.2	-8.0	-9.4	-7.7	-7.5
Sbot_2VEP	-9.6	-10.9	-8.9	-7.1	-6.0	-6.2	-6.7	-8.1	-9.1	-8.8	-8.3	-7.9	-8.9	-9.3	-9.4	-9.8	-10.0	-8.9	-8.2	-8.0
Sipo_2VEP	-6.6	-6.6	-7.3	-6.9	-6.5	-6.3	-6.5	-8.1	-8.6	-9.1	-8.0	-7.9	-8.0	-8.4	-8.5	-7.2	-8.4	-8.2	-7.9	-7.6
Ssvi	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ssvi_4U28	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ssvi_4TX9	-5.9	-8.2	-7.1	-6.6	-6.8	-7.0	-7.6	-7.9	-8.4	-8.3	-8.2	-8.1	-8.4	-8.7	-7.7	-8.2	-8.2	-6.7	-8.0	-7.5
Salb_2VEP	-6.3	-7.1	-7.4	-6.8	-7.7	-6.9	-6.6	-8.0	-9.6	-8.9	-8.6	-7.9	-8.6	-8.4	-9.1	-7.4	-7.2	-8.8	-8.1	-7.8
Sbik_2VEP	-7.4	-6.7	-7.4	-7.3	-7.7	-6.5	-6.8	-8.5	-9.9	-9.8	-8.3	-7.8	-8.2	-8.4	-10.5	-7.1	-6.3	-8.3	-8.1	-7.9
Stsu_2VEP	-6.9	-7.5	-7.9	-7.8	-7.5	-7.3	-7.7	-8.1	-10.0	-9.7	-8.7	-7.8	-8.8	-8.8	-10.6	-7.8	-8.7	-9.2	-7.9	-7.8
Sven_2VEP	-8.3	-7.6	-7.5	-6.8	-6.4	-7.0	-6.2	-8.2	-9.6	-9.0	-7.9	-7.4	-8.3	-8.6	-10.0	-8.2	-8.2	-8.1	-7.6	-7.4
Scal_2VEP	-7.3	-8.8	-7.5	-6.7	-7.0	-6.5	-6.5	-8.7	-10.0	-9.7	-8.8	-7.7	-8.2	-8.6	-10.5	-8.1	-9.2	-8.9	-7.6	-7.6
Sbaa_2VEP	-8.9	-6.6	-6.7	-6.8	-7.2	-7.2	-6.2	-7.7	-9.6	-9.5	-8.4	-7.4	-8.5	-8.2	-9.7	-8.5	-8.4	-9.3	-7.4	-7.7
Sglo_2VEP	-6.1	-6.8	-6.9	-6.5	-6.7	-6.3	-6.7	-7.5	-9.6	-9.6	-8.6	-7.8	-8.7	-8.7	-9.9	-6.2	-5.1	-9.2	-7.8	-7.7
Sful_2VEP	-9.0	-6.9	-7.1	-6.8	-6.3	-7.2	-6.6	-8.1	-9.8	-9.3	-7.7	-8.0	-8.7	-8.8	-10.5	-7.6	-9.1	-10.5	-7.1	-7.4
Sgri_2VEP	-6.6	-8.2	-7.1	-6.4	-7.1	-7.4	-7.1	-7.4	-9.4	-8.5	-8.6	-7.5	-8.8	-8.4	-9.6	-8.8	-8.8	-8.7	-7.7	-7.5
S34_2VEP	-7.8	-7.2	-7.6	-7.3	-6.2	-6.5	-6.9	-8.0	-8.8	-8.6	-7.7	-7.1	-7.7	-7.9	-7.2	-7.1	-6.7	-7.1	-7.7	-7.7
Srim_2VEP	-7.4	-6.8	-7.1	-7.2	-7.2	-7.4	-7.8	-8.6	-9.7	-9.3	-8.8	-7.7	-8.9	-8.7	-10.2	-6.8	-8.2	-9.0	-7.5	-7.8
Satr_2VEP	-7.6	-7.1	-7.5	-7.4	-6.2	-6.5	-6.9	-8.0	-8.9	-8.7	-7.7	-7.4	-7.7	-7.8	-6.5	-7.0	-6.3	-7.7	-7.7	-7.7
S1813_2VEP	-7.5	-6.7	-6.8	-6.6	-6.8	-7.0	-6.7	-7.9	-9.9	-9.7	-8.3	-7.7	-8.5	-8.6	-10.0	-7.6	-7.5	-8.6	-7.3	-7.5
Svar_2VEP	-6.3	-6.8	-7.0	-6.3	-6.9	-6.9	-6.6	-7.8	-9.3	-7.4	-8.5	-8.5	-7.3	-8.0	-8.7	-8.5	-7.4	-6.3	-6.0	-7.5
Sfra_2VEP	-7.3	-7.0	-6.8	-6.6	-6.5	-7.3	-6.6	-8.6	-9.1	-8.7	-8.9	-7.7	-8.9	-9.0	-10.6	-6.0	-6.3	-7.9	-8.2	-7.8
Smeg_2VEP	-8.0	-7.3	-6.9	-6.6	-7.2	-6.7	-6.1	-7.6	-9.2	-9.1	-7.8	-7.4	-8.2	-8.5	-9.6	-9.2	-9.2	-8.6	-7.7	-7.5
Ssul_2VEP	-7.5	-7.0	-6.9	-6.6	-6.6	-6.7	-6.9	-7.7	-8.4	-8.2	-8.2	-7.5	-8.1	-7.7	-8.2	-6.7	-7.1	-7.7	-7.6	-7.4
Slav_4X9S	-8.0	-7.1	-6.7	-6.9	-6.6	-6.8	-6.7	-8.2	-9.5	-9.1	-8.6	-8.0	-8.4	-8.7	-9.1	-9.5	-9.3	-8.1	-8.1	-7.6
Sery_4X9S	-7.8	-7.2	-7.6	-7.2	-7.4	-6.9	-7.4	-8.5	-9.4	-9.4	-8.6	-7.6	-8.4	-8.6	-9.1	-10.0	-9.0	-7.8	-8.2	-8.2
SspC_4X9S	-7.4	-6.9	-7.3	-6.8	-7.3	-6.6	-7.5	-8.0	-9.8	-9.3	-8.7	-7.6	-8.5	-8.4	-10.6	-9.1	-8.9	-8.5	-8.1	-8.0
Sxan_4X9S	-7.5	-6.9	-6.8	-6.8	-6.5	-6.6	-6.5	-8.3	-8.0	-8.1	-7.6	-7.4	-8.7	-8.1	-8.1	-8.5	-8.0	-7.6	-7.7	-7.4
Skat_4X9S	-6.4	-7.0	-7.1	-6.7	-7.7	-6.3	-6.2	-8.1	-8.5	-9.2	-8.3	-7.6	-9.0	-8.5	-9.7	-9.5	-9.9	-9.2	-7.4	-7.5
SMg1_4X9S	-6.5	-6.9	-7.2	-7.1	-6.5	-6.9	-6.4	-7.3	-7.8	-7.7	-8.3	-7.5	-7.9	-8.4	-9.5	-7.6	-5.2	-7.5	-7.6	-7.7
SMg1_W9T	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Sniv_2VEP	-6.2	-6.8	-7.4	-6.5	-7.8	-6.8	-6.7	-9.5	-9.6	-9.4	-8.7	-8.2	-8.6	-9.1	-9.9	-5.4	-5.5	-9.2	-8.0	-7.5
Scla_2VEP	-6.4	-7.7	-7.4	-7.2	-7.6	-7.5	-7.2	-8.5	-10.1	-10.0	-8.7	-7.9	-8.8	-9.1	-10.2	-6.9	-8.1	-8.5	-7.7	-7.7
Save_2VEP	-7.6	-6.7	-7.1	-7.2	-7.1	-6.5	-6.6	-7.8	-8.6	-10.0	-8.6	-7.5	-8.3	-8.5	-8.8	-5.3	-5.2	-5.8	-7.7	-7.4
Spur_2VEP	-6.8	-7.3	-6.9	-6.7	-8.0	-7.9	-7.7	-8.5	-9.5	-9.8	-8.1	-7.8	-8.7	-9.7	-10.0	-5.3	-6.2	-8.2	-7.6	-7.4
Scar_2Y89	-6.9	-6.8	-7.5	-7.1	-7.8	-7.3	-7.2	-8.3	-9.2	-8.3	-8.9	-8.4	-8.9	-9.3	-9.4	-5.3	-5.2	-7.1	-8.5	-8.6
Mtub_2Y88	-10.0	-7.8	-8.9	-5.4	-8.6	-7.3	-7.8	-9.5	-10.9	-10.3	-9.5	-9.0	-9.8	-9.8	-11.3	-10.1	-10.2	-11.3	-9.5	-8.8
Mtub_2Y89	-8.7	-8.6	-9.6	-9.4	-6.4	-5.9	-5.6	-7.1	-7.0	-7.5	-7.3	-6.8	-7.4	-8.4	-7.5	-8.1	-8.6	-7.5	-7.7	-7.3
Mtub_2Y85	-8.2	-7.9	-9.2	-7.4	-7.6	-7.5	-7.6	-8.4	-9.7	-9.5	-9.3	-7.8	-8.6	-8.6	-10.2	-9.8	-9.9	-10.1	-7.3	-7.4
Mtub_3ZS4	-10.0	-10.5	-11.4	-6.4	-8.2	-7.2	-7.0	-9.6	-10.2	-10.2	-9.9	-8.5	-9.3	-9.6	-10.9	-10.0	-10.7	-9.9	-8.8	-8.9
S34_3ZS4	-7.4	-8.0	-7.6	-6.4	-5.2	-5.4	-5.2	-6.1	-6.8	-6.4	-5.7	-6.4	-6.3	-6.3	-7.1	-7.4	-6.4	-7.1	-6.6	-6.6
Cdip_4AXK	-9.2	-7.8	-10.9	-7.2	-7.5	-7.6	-7.7	-8.9	-9.0	-9.8	-9.0	-8.3	-8.8	-9.2	-10.1	-9.5	-9.9	-9.2	-8.0	-7.9
Cjei_4AXK	-7.5	-6.9	-7.2	-6.5	-8.4	-8.0	-8.5	-8.7	-9.5	-9.6	-9.4	-8.8	-9.0	-9.5	-9.3	-8.1	-9.3	-8.5	-7.9	-7.7
Aaur	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Aaur_4WD0	-8.2	-8.5	-8.1	-7.9	-5.7	-5.5	-5.9	-7.0	-7.5	-7.2	-7.1	-6.7	-7.1	-7.4	-8.4	-7.2	-7.3	-7.4	-7.0	-6.7
Acar_4X2R	-10.6	-9.3	-9.3	-7.2	-7.2	-6.8	-7.3	-7.5	-9.5	-9.8	-8.0	-7.8	-9.3	-8.9	-10.3	-9.6	-9.4	-9.4	-8.1	-8.0
Auro_4X2R	-9.9	-9.1	-9.8	-8.1	-7.4	-7.1	-7.4	-7.8	-9.2	-9.8	-8.3	-7.8	-9.3	-9.0	-9.9	-9.0	-8.6	-9.2	-8.5	-8.2
Afer_4WD0	-5.7	-5.8	-6.0	-5.7	-6.7	-6.2	-5.8	-7.6	-9.2	-8.8	-7.8	-7.4	-8.3	-8.4	-9.3	-6.7	-4.5	-9.1	-8.3	-8.1
Sent_5AHE	-9.1	-5.4	-6.4	-5.2	-8.0	-7.3	-7.4	-8.8	-10.7	-10.2	-8.7	-8.7	-9.6	-9.9	-10.9	-7.8	-9.1	-10.3	-9.0	-8.4
Ecoli_K12	-9.7	-9.7	-9.2	-6.1	-7.2	-6.6	-6.8	-8.6	-9.5	-9.1	-8.6	-8.2	-9.0	-8.6	-10.2	-9.9	-10.2	-9.9	-8.2	-7.9
Jden_4WUI	-8.3	-8.8	-8.2	-6.8	-6.2	-6.1	-6.0	-6.8	-7.4	-7.6	-7.5	-6.9	-7.6	-7.5	-7.7	-7.5	-7.6	-7.2	-7.3	-7.2
Ctra	-8.2	-8.0	-7.4	-7.2	-6.1	-5.5	-5.4	-6.5	-7.1	-7.1	-7.0	-6.2	-6.8	-6.8	-6.9	-7.2	-7.4	-6.5	-6.7	-6.6
SMg1_trpF	-7.2	-8.8	-8.2	-7.3	-6.2	-5.7	-5.7	-6.9	-6.6	-6.7	-7.3	-6.7	-7.6	-6.9	-7.5	-7.1	-7.1	-6.9	-6.7	-6.5
Aodo_4X2R	-8.5	-8.6	-8.8	-7.3	-7.1	-7.1	-7.5	-9.7	-9.4	-9.4	-7.8	-7.6	-9.6	-8.8	-10.1	-8.4	-8.9	-9.4	-8.0	-8.1

## 4.5. PriA en cinéticas enzimáticas no tradicionales.

Además de la exploración genómica de PriA se realizaron caracterizaciones experimentales. No se tuvo el tiempo ni la experiencia para tener réplicas de los resultados descritos a continuación, sin embargo pueden ser un buen comienzo para futuros estudiantes que deseen retomarlos. Todos los protocolos fueron cuidadosamente descritos y se encuentran en los anexos. Este capítulo se enfoca en el montaje de dos experimentos. *i) Cinéticas de PriA en GTP . ii) Medición simultánea de la actividad de PriA sobre ProFAR y PRA.*

Lo primero que se realizó fue la sobre expresión heteróloga de PriA y su respectiva purificación para poder realizar *in vitro* cinéticas enzimáticas. Las enzimas fueron clonadas en cepas de *E. coli* V68 se indujo sobre expresión y se purifico la proteína. Se corroboró el éxito de esta actividad mediante un gel de proteína en la que puede verse expresión en la barra correspondiente al tamaño de PriA Figure 4.19.



Figura 4.19: Gel de proteína donde se muestra la banda correspondiente a PriA después de ser sobre-expresada y purificada

Con la proteína purificada y almacenada a -80° en esferas de proteína selladas con nitrógeno líquido como se describe en el anexo de los procedimientos se procedió a realizar cinéticas enzimáticas.

#### 4.5.1. PriA puede tener actividad en GTP

La actividad fue medida fluorometricamente en placas de 96 (Nuc 96-Well Optical Botto Plates) en un lector de placas TECAN infinite M1000 (con exitación a 286 nm y emisión a 386 nm). Ensayos preliminares de actividad se realizaron en una PriA activa respecto a los otros substratos y proveniente de *Streptomyces coelicolor* y en la mutante inactiva D11A.

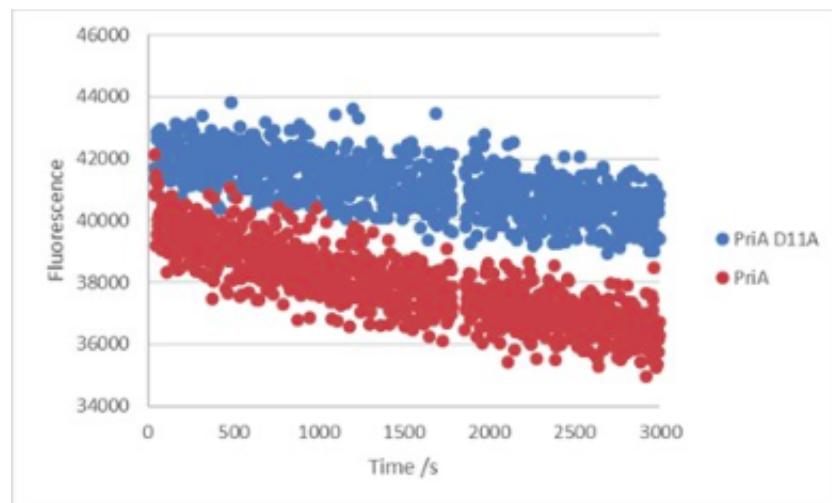


Figura 4.20:

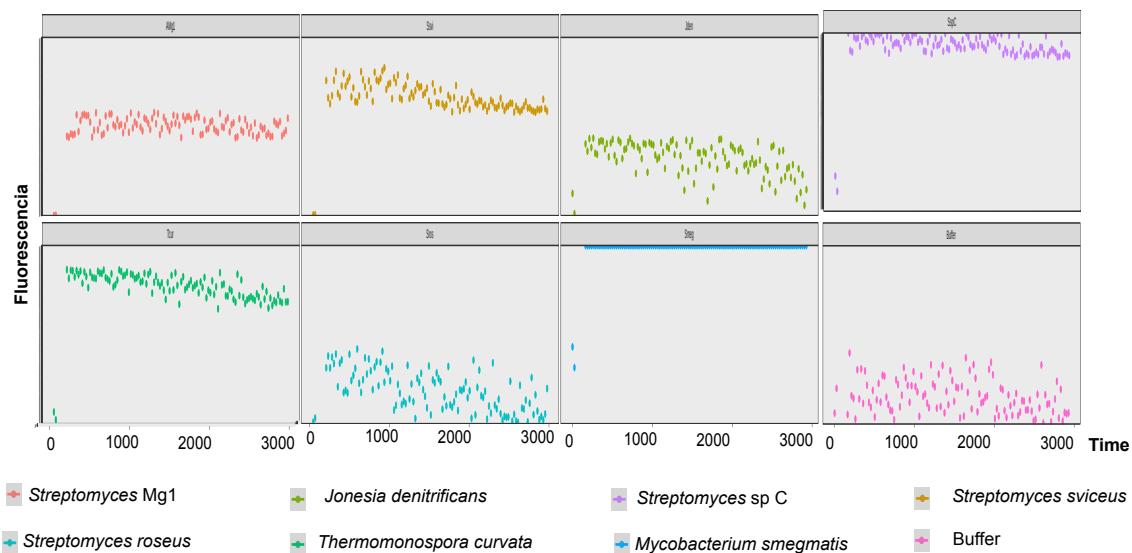


Figura 4.21:

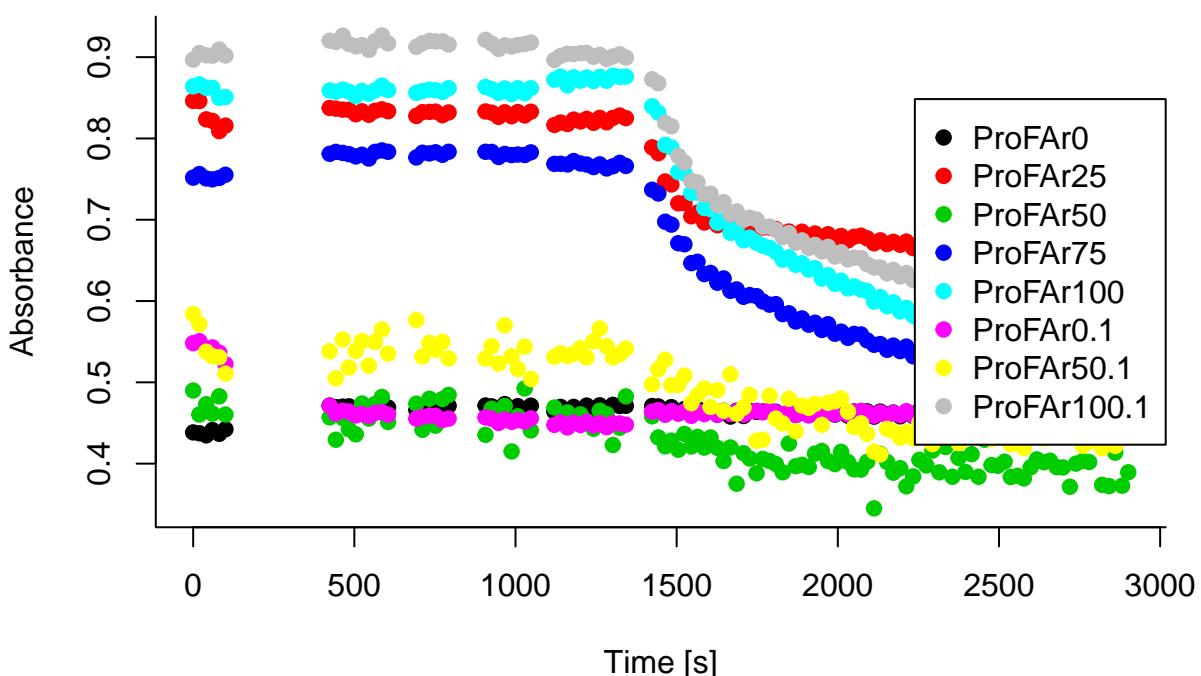
Entre las enzimas donde se aprecia actividad está *Thermomonospora curvata* que es una Actinobacteria termófila de la familia *Thermomonosporaceae* y que puede ser encontrada en compostas ya que participa en la degradación de celulosa [180]. La trpF de *Jonesia denitrificans* no mostró actividad. Esta bacteria está clasificada como un organismo patogénico para animales, su genoma fue originalmente aislado de sangre [181].

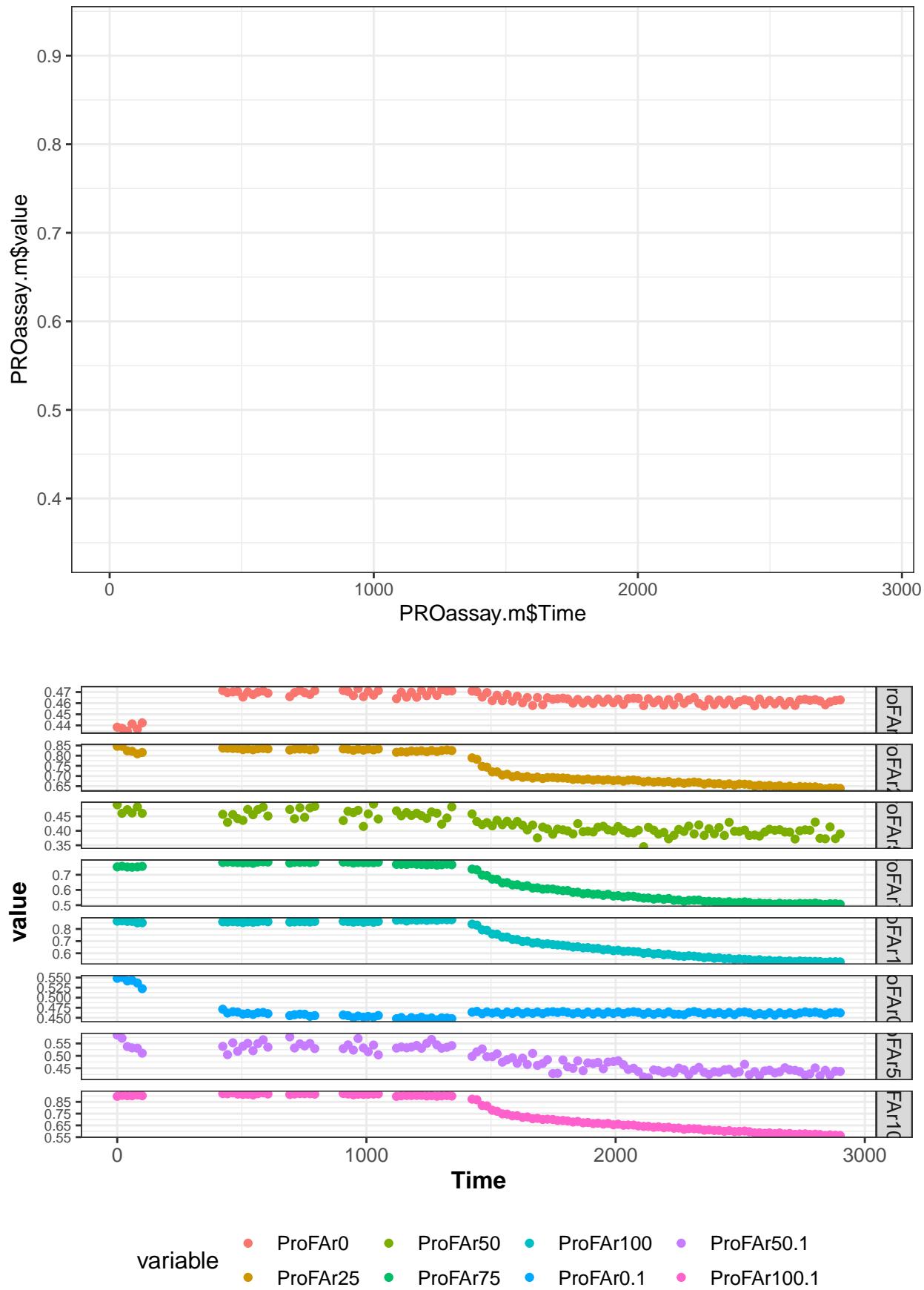
### 4.5.2. Cinéticas simultáneas para PRA y ProFAR

En esta sección se trata de medir simultáneamente las actividades de Pra y ProFAR.

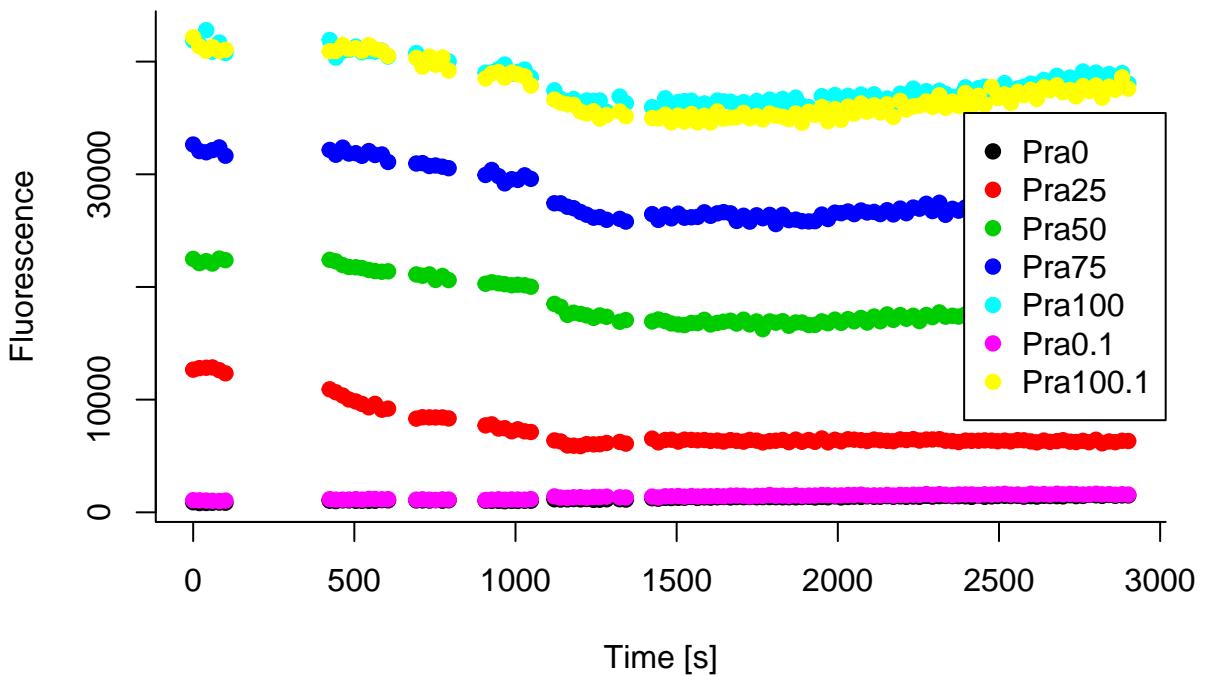
## 4.6. Cinéticas

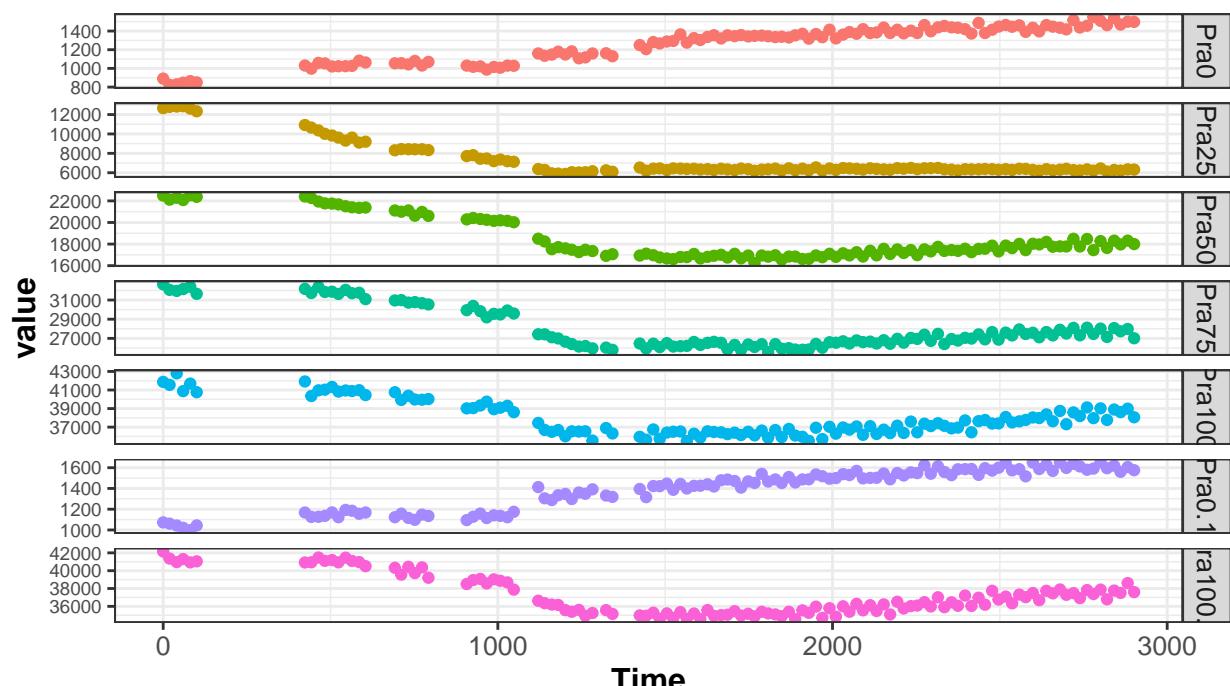
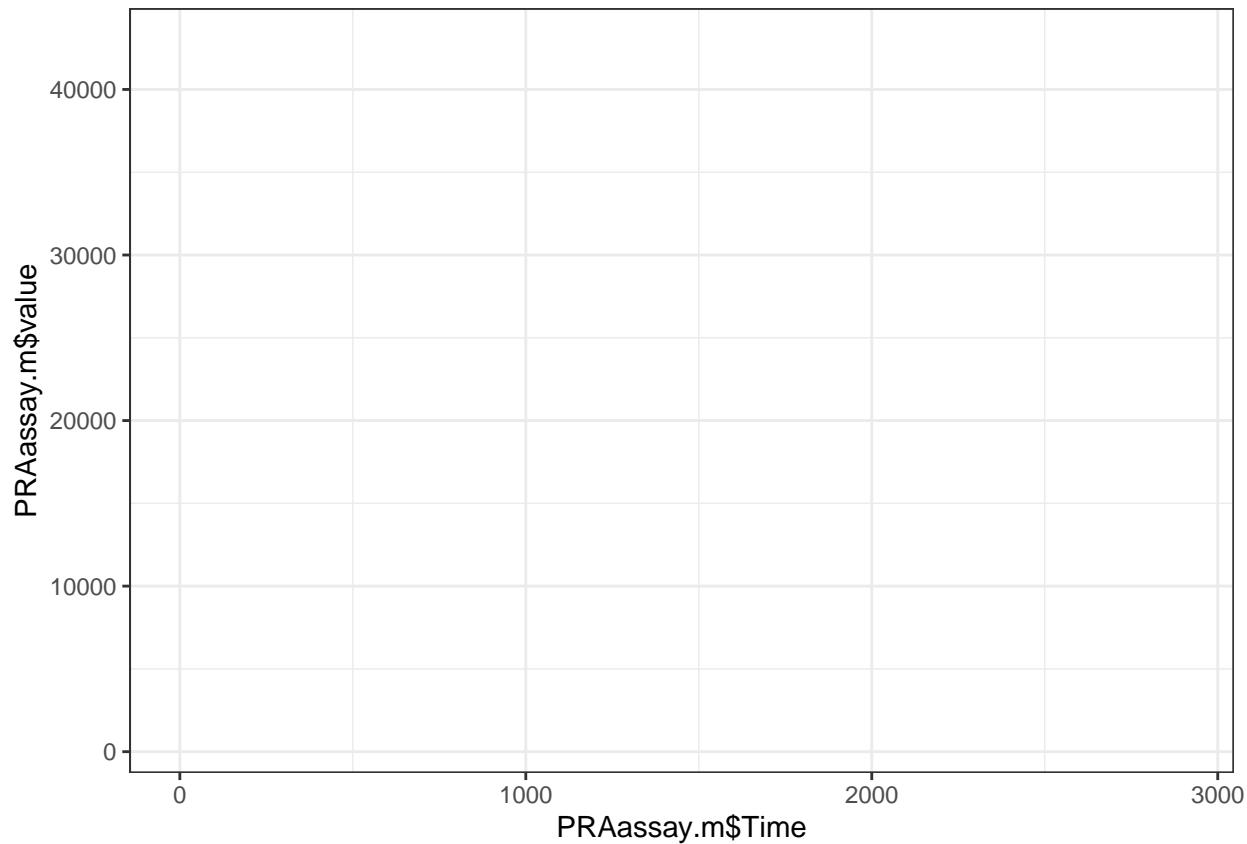
[1]	0.00	19.15	40.50	59.65	81.00	100.15	423.50	442.65
[9]	464.00	483.15	504.50	523.65	545.00	564.15	585.50	604.65
[17]	692.40	711.55	732.90	752.05	773.40	792.55	907.00	926.15
[25]	947.50	966.65	988.00	1007.15	1028.50	1047.65	1120.70	1139.85
[33]	1161.20	1180.35	1201.70	1220.85	1242.20	1261.35	NA	NA
[41]	1282.80	1301.95	1323.30	1342.45	1424.00	1443.15	1464.50	1483.65
[49]	1505.00	1524.15	1545.50	1564.65	1586.00	1605.15	1626.50	1645.65
[57]	1667.00	1686.15	1707.50	1726.65	1748.00	1767.15	1788.50	1807.65
[65]	1829.00	1848.15	1869.50	1888.65	1910.00	1929.15	1950.50	1969.65
[73]	1991.00	2010.15	2031.40	2050.55	2071.90	2091.05	2112.50	2131.65
[81]	2153.00	2172.15	2193.50	2212.65	2234.00	2253.15	2274.50	2293.65
[89]	2315.00	2334.15	2355.50	2374.65	2396.00	2415.15	2436.50	2455.65
[97]	2477.00	2496.15	2517.50	2536.65	2558.00	2577.15	2598.50	2617.65
[105]	2639.00	2658.15	2679.50	2698.65	2720.00	2739.15	2760.50	2779.65
[113]	2801.00	2820.15	2841.50	2860.65	2882.00	2901.15		





[1]	0.00	19.15	40.50	59.65	81.00	100.15	423.50	442.65
[9]	464.00	483.15	504.50	523.65	545.00	564.15	585.50	604.65
[17]	692.30	711.45	732.80	751.95	773.40	792.55	907.00	926.15
[25]	947.50	966.65	988.00	1007.15	1028.50	1047.65	1120.70	1139.85
[33]	1161.20	1180.35	1201.70	1220.85	1242.20	1261.35	1282.80	1301.95
[41]	1323.30	1342.45	1424.00	1443.15	1464.50	1483.65	1505.00	1524.15
[49]	1545.50	1564.65	1586.00	1605.15	1626.50	1645.65	1667.00	1686.15
[57]	1707.50	1726.65	1748.00	1767.15	1788.50	1807.65	1829.00	1848.15
[65]	1869.50	1888.65	1910.00	1929.15	1950.40	1969.55	1991.00	2010.15
[73]	2031.50	2050.65	2071.90	2091.05	2112.50	2131.65	2153.00	2172.15
[81]	2193.50	2212.65	2234.00	2253.15	2274.50	2293.65	2315.00	2334.15
[89]	2355.50	2374.65	2396.00	2415.15	2436.50	2455.65	2477.00	2496.15
[97]	2517.50	2536.65	2558.00	2577.15	2598.50	2617.65	2639.00	2658.15
[105]	2679.50	2698.65	2720.00	2739.15	2760.50	2779.65	2801.00	2820.15
[113]	2841.50	2860.65	2882.00	2901.15				

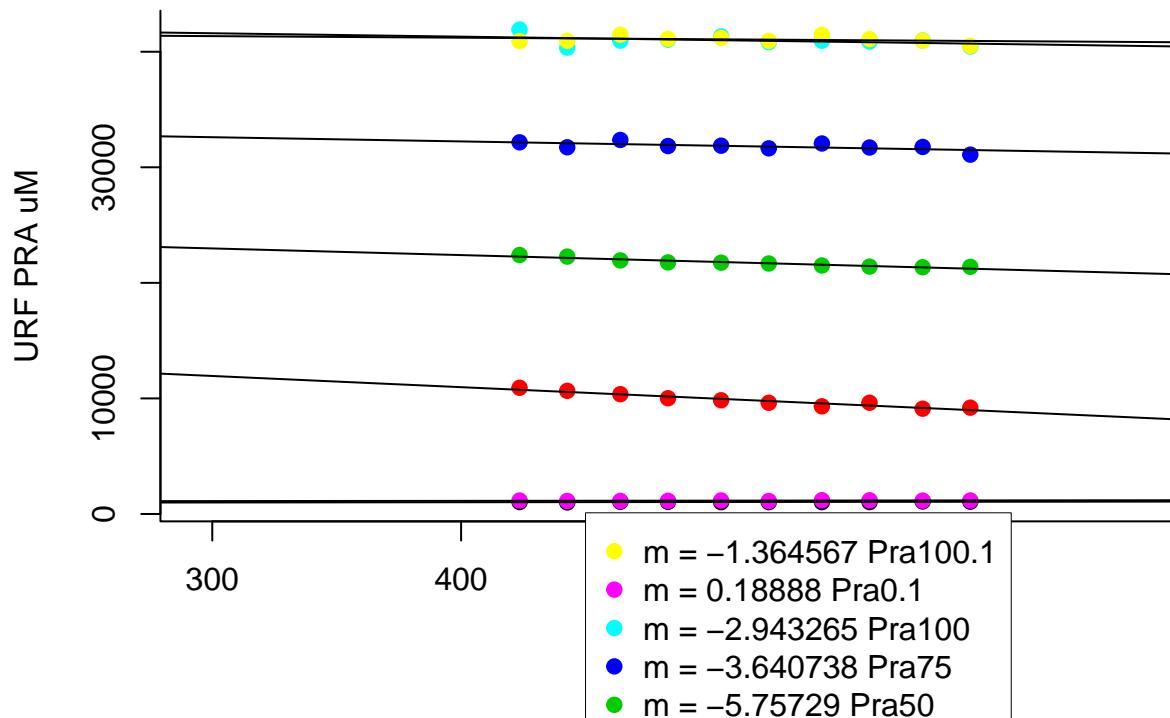




variable	● Pra0	● Pra50	● Pra100	● Pra100.1
	● Pra25	● Pra75	● Pra0.1	

## 4.7. Gráfico de pendientes

NULL



```
[1] 0.196328 -9.695510 -5.757290 -3.640738 -2.943265  0.188880 -1.364567  
[1] 0.196328 -9.695510 -5.757290 -3.640738 -2.943265  0.188880 -1.364567  
[1] 1031 10920 22410 32161 41918 1168 40922
```

## 4.8. Interpolación con Michaelis Menden

```
#sessionInfo()
```



# Metodología

## 4.1. Protocolos para usar Orthocore, myRAST, fastOrtho, Clavigenomics, y BPGA

Anotación genómica con el docker myRAST

Esta es una distribución de myRAST en un contenedor de docker. Para usarla se necesita una cuenta del anotador genómico RAST. El docker myRAST permite hacer anotación genómica y funcional masiva mediante el uso de la terminal en el anotador RAST. Después de anotar los resultados pueden descargarse y procesarse en una terminal.

Descargar myrast docker distribution

Una vez con docker instalado en la computadora, se hace pull al docker myrast.

docker pull nselem/myrast

Abrir myRast en la terminal

docker run -i -t -v \$(pwd):/home nselem/myrast /bin/bash

Usar myRast

-Ejemplo subir un archivo fasta

svr\_submit\_RAST\_job -user -passwd -fasta -domain Bacteria -bioname "Organism name" -genetic\_code 11 -gene\_caller rast

-Para bajar un archivo de anotación genómica funcional:

svr\_retrieve\_RAST\_job table\_txt > \$ID.txt

Una lista completa de archivos puede ser procesada usando bash. Por ejemplo para bajar una lista de archivos de RAST se deben guardar los identificadores de RAST en una columna de un archivo, (Rast\_ID en este ejemplo) y usar un while para obtenerlos:

En este caso la variable “line” contendrá el identificador RAST Id, y cada archivo fasta de aminoácidos podrá ser obtenido mediante su identificador de RAST y será guardado en el archivo “\$line.faa”

```
cut -f1 Rast_ID | while read line; do svr_retrieve_RAST_job $line amino_acid >
$line.faa ; done
```

Formatos de RAST para descargar archivos

Puedes cambiar el formato table\_txt por el que tú necesites.

Cuadro 4.6: Formatos de descarga disponibles en myRAST

Atributo	Descripción
genbank	GenBank (con funciones y enriquecimiento de SEED)
genbank_stripped	Genbank con EC-numbers removidos de las funciones
embl	EMBL (con funciones y enriquecimiento de SEED)
embl_stripped	EMBL con EC-numbers removidos de las funciones
gff3	GFF3
gff3_stripped	GFF3 con EC-numbers removidos de las funciones
gtf	GTF
gtf_stripped	GTF con EC-numbers removidos de las funciones
rast_tarball	Archivo comprimido (gzipped) con todo el directorio de las anotaciones de RAST sobre el genoma
nucleic_acid	Fasta de DNA de genes
amino_acid	Fasta de DNA de aminoácidos
table_txt	Gene data in tab-separated format
table_xls	Preserve the original gene calls and use RAST

Referencia Table 4.6.

## 4.2. Orthocore

El umbral de e-value de Orthocore es por default 1e-6. Todas las secuencias son alineadas usando MUSCLE v3.8.31 con los parámetros default y curadas utilizando Gblocks con 5 posiciones como longitud mínima del bloque, 10 como máximo número de posiciones contiguas no conservadas y sólo considerando posiciones con gaps menores que el 50 % de las secuencias en el alineamiento final. Después de esta curación las secuencias son concatenadas en una matriz final.

# Capítulo 5

## Conclusión

```
\setcounter{section}{1}
```

Idea de Rosario -ver dell cluster de saxitoxin cuantos pasos se necesitron para llegar ahi.

- A donde se iria el resultado de abrir el GMP
- Otra vez, que Actinos tienen FolE

### 5.1. Discussion

Finalmente, una pregunta abierta es aún dada una enzima promiscua es la población promiscua, y son sólo unos confórmeros o todas y cada una de las enzimas son promiscuas

If we don't want Conclusion to have a chapter number next to it, we can add the `{.unnumbered}` attribute. This has an unintended consequence of the sections being labeled as 3.6 for example though instead of 4.1. The L<sup>A</sup>T<sub>E</sub>X commands immediately following the Conclusion declaration get things back on track.

#### More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.



# Apéndice A

## The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readability and/or setup.

In the main Rmd file:

```
# This chunk ensures that the reedtemplates package is
# installed and loaded. This reedtemplates package includes
# the template files for the thesis and also two functions
# used for labeling and referencing
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(reedtemplates)){
  library(devtools)
  devtools::install_github("ismayc/reedtemplates")
}
library(reedtemplates)
library(kableExtra)
```

In :

```
# This chunk ensures that the reedtemplates package is
# installed and loaded. This reedtemplates package includes
# the template files for the thesis and also two functions
# used for labeling and referencing
if(!require(devtools))
  install.packages("devtools", repos = "http://cran.rstudio.com")
if(!require(plyr))
```

```
install.packages("plyr", repos = "http://cran.rstudio.com") ## this shoul always go
if(!require(dplyr))
  install.packages("dplyr", repos = "http://cran.rstudio.com")
if(!require(ggplot2))
  install.packages("ggplot2", repos = "http://cran.rstudio.com")
if(!require(reedtemplates)){
  library(devtools)
  devtools::install_github("ismayc/reedtemplates")
}
library(reedtemplates)
```

# Apéndice B

## The Second Appendix, Open source code on this document

### B.1. R markdown

Thanks to Rmакdown Thesis  
Apendix one Useful docker commands  
-Create a new repository  
`docker build . -t evomining`  
`docker push nselemevomining`

### B.2. Docker

Restart docker and free all ports  
`sudo service docker restart`

list containers  
`docker ps -a`

ssh or bash into a running docker container  
`sudo docker exec -i -t romantic_brahmagupta /bin/bash`   `docker exec -it <mycontainer> bash`

Stop all containers  
`docker rm $(docker ps -a -q)`

Remove stopped containers  
`docker rm $(docker ps -q -f status=exited)`

Remove all images  
`docker rmi $(docker images -q)`

uninstall docker from ubuntu (Fresh start)

```
sudo apt-get purge docker-engine
```

```
sudo apt-get autoremove --purge docker-engine
```

```
rm -rf /var/lib/docker # This deletes all images, containers, and volumes
```

Run Evomining container using nselem/newevomining image

```
docker run -i -t -v /home/nelly/GIT/EvoMining:/var/www/html/EvoMining/exchange
-p 80:80 nselem/newevomining /bin/bash
```

Start evomining inside this container

```
perl startevomining
```

Vizualice a tree

```
http://10.10.100.234/EvoMining/cgi-bin/color\_tree.pl?9&&/var/www/html/EvoMining/exchange
file 9.new must be on folder volume CyanosBBH_MiBIG_DB.faa_CYANOS
```

Find a perl module

```
perl -MList::Util -e'print $_ . " => " . $INC{$_} . "\n" for keys %INC'
```

EvoMining notes

Gblocks only runs inside folder /var/www/html/EvoMining

### B.3. Git

```
git add --all
git commit -m "Some message"
git push -u origin master
git clone
```

### B.4. Connect GitHub and DockerHub

automated builds The Dockerfile is available to anyone with access to your Docker Hub repository. Your repository is kept up-to-date with code changes automatically.

### B.5. Additional resources

- *Markdown Cheatsheet* - <https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet>
- *R Markdown Reference Guide* - <https://www.rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf>
- Introduction to dplyr - <https://cran.rstudio.com/web/packages/dplyr/vignettes/introduction.html>

- `ggplot2` Documentation - <http://docs.ggplot2.org/current/>
- Docker antiSMASH run\_antismash /.gbk ~/as\_results -knownclusterblast -inclusive -cf\_threshold 0.7



# **Apéndice C**

## **El tercer Apéndice, Protocolos experimentales utilizados en el estudio de PriA**

### **C.1. Materiales**

- Incubadora
- Centrifugadora refrigerada
- Sonicador
- Columna HisTrapHP
- Columna AKTA-FPLC
- Columna Vivapure Mini
- Fluorímetro

### **C.2. Purification**

TrpC and TrpD will be purified by Nickel chromatography using AKTA-FPLC while ScoePriA and D11A purification will be done by Nickel chromatography using vivapure columns

#### **C.2.1. Centrifugation**

1. Centrifuge on ,5L cans (Cans as big as possible) rotor that suits for cans (On ,5L cans rotor code is 03)  
Karina: velocity *6000rpm* Time *10min* Ana: *7000rpm* for *10min* At this point experiment may be interrupted and pellet can be stored at  $-80^{\circ}$  on Falcon tubes or better run the essay the same day.

To restart this protocol re suspend pellet on 25ml lysis buffer keeping on ice all the following steps.

### C.2.2. Sonication

We sonicate pellet to disrupt cells and get the proteins. All reactives must be cooled previously and kept on ice. The sonicator must touch the falcon tube. Falcon must be kept on ice.

Add fresh components to Lysis buffer:

- Add proteasas inhibitor mix  $2\mu l$  (Protein Box)
- Add DTT , $125\mu l$
- Karina: Add Lysozyme  $\frac{2mg}{ml}$   $100\mu l$
- Ana: Add Lysozyme  $\frac{10mg}{ml}$  ( $1000X$ )  $100\mu l$

2. Resuspension. Add lysis Buffer 25ml to resuspend pellet and put test tubes on an small glass ware with ice
3. Sonication. After buffer addition sonicate:  
7 times sonication 20s with 1min rest between every sonication cycle.  
Only total time of sonication must be entered on sonication device (140s).
4. Centrifugation. After sonication centrifuge on refrigerated centrifuge at 4°C 25min at 8000rpm
5. Inject into AKTA-FPL previously balanced with Buffer A (or vivapure balanced with bufferA)  
When vivapure is being used, the supernatant must be filtered on .45  $\mu l$  filters to avoid column damage.
6. Wash with 5 % Buffer B and subsequently with an imidazol gradient (Buffer B) from 5 % to 60 % on 16 vol of the column with a flux of 1ml/min. Conserve each elution tube properly labeled. 2000rpm 5 min  
What we did to purify Scoe PriA was an imidazol gradient:

Flask each one with a 100ml were prepared. columns were washed with 20ml of this tubes at 2000rpm 5 min

Imidazol concentration=x

v: (x=)100ml/500mM

Flask	BufferB (mM)	BufferB (ml)	Buffer A ml	Buffer B
1	50	10	90	10
2	100	20	80	20
3	150	30	70	30
4	300	60	40	60
5	500	100	0	100

Flask must be kept cold.

7. Recover the fractions that contains the protein that elutes approximately at 150-200 mM imidazol. (TrpC, TrpD) or 300mM imidazol (must PriA). Collect the relevant fractions. At this point a protein gel can be run to check protein elution.
8. Concentrate the protein using AMICON tubes 10kDa (all molecules below this weight will be discarded, molecules with weight above 10kDa such as PriA will be kept inside the tube) until an approximate concentration of  $90 \frac{\mu g}{\mu l}$  (1ml vol approx). Because imidazol>200mM reacts with Bradford, protein on amicon must be washed with 10 ml of buffer A to eliminate imidazol (4ml each time on amicon tube). Finally, after protein concentration, 100 ul sample must be recovered with 900 ul Buffer A (free of imidazol) (8000 rpm 10 min) when AMICON are stored for long periods leave on ethanol at 20 %, to reuse them throw ethanol out and wash many times with miliQ water.
9. Check purity by SDS-PAGE gel  
on *S coelicolor* PriA molecular weight is around 25.7 kDa (length 240 aminoacids),  
on *T maritima* TrpC molecular weight is around 27.5 kDa (length 242 aminoacids)
10. Add glycerol at a final concentration of 15 % and quantify using Bradford. Store at -80C

*TrpC important before step 5*

After centrifuge step 4 :

Heat supernatant by 15 min at 65C

Centrifuge at 10000rpm by 20min. Recover the supernatant.

Continue with steps 5-10 as in trpD purification.

On protein activity is more important to be pure than to have protein abundance.

## C.3. Protein Quantification

### C.3.1. Bradford calibration curve

Mix well while pipeting. Holes with protein content must become blue.  
Incubate 5 min.

Measure at 595 abs (TECAN). clear box.

5 ul y 50ul of protein. Take measures inside the curve range.

BSA on ug/ml

\$\$

BSA $\mu l$ por ml	BSA $\mu l$	Bradford $\mu l$	Other
0	0	40	$H_2O$ 160 $\mu l$
10	160	40	-
20	160	40	-
40	160	40	-
60	160	40	-
100	160	40	-
0	0	40	PriA 5 $\mu l$ $H_2O$ 155 $\mu l$
0	0	40	PriA 20 $\mu l$ $H_2O$ 130 $\mu l$

\$\$

```
## 160 ul of BSA on this concentrations on ug/ml were added to a 40 ul of Bradford
## BSA concentration on ug/ml
BSAConcentrationStock<-c(0,10,20,40,60,100)

#adding .16 ml on plates
vol<-160
Bradford_df = data.frame(BSAConcentrationStock,vol)

## quantity of microgrammes on .16 ml
Bradford_df["Quantity"]<-BSAConcentrationStock*.160
## concentration after added 40 ul bradford (total 200 ul)
Bradford_df["Concentration200"]<-Bradford_df["Quantity"]/2

## This are the absorbance measures for the different BSA concentrations
Abs1BSA<-c(0.2649999857,0.4329000115,0.6754000187,0.9449999928,1.24240005,1.432000041)
Abs2BSA<-c(0.2948000133,0.4298999906,0.7348999977,0.9965000153,1.225999951,1.494899988)
```

```

Bradford_df["Abs1BSA"]<-Abs1BSA
Bradford_df["Abs2BSA"]<-Abs2BSA

## mean value of measures
Bradford_df$mean<-rowMeans(subset(Bradford_df, select = c(Abs1BSA, Abs2BSA)), na.rm=TRUE)

#adjusting blank
Bradford_df$meanAdjusted<-Bradford_df$mean-Bradford_df$mean[1]
BSA.lm = lm(meanAdjusted ~ Concentration200, data=Bradford_df)
coeffs = coefficients(BSA.lm); coeffs

```

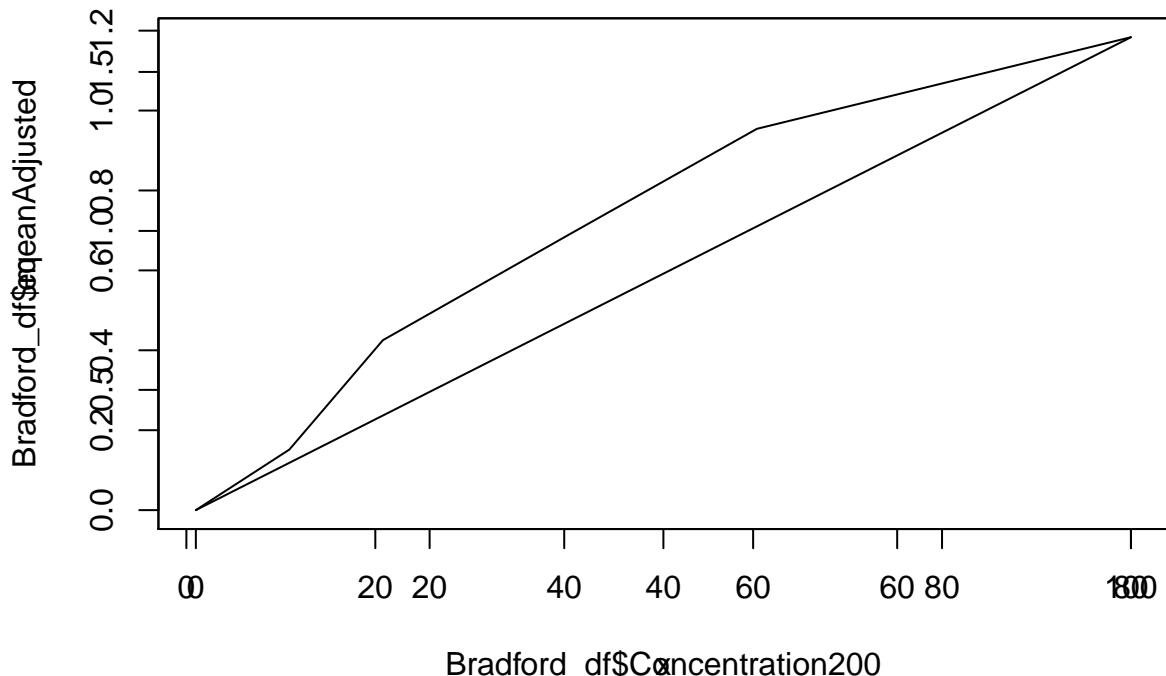
```

(Intercept) Concentration200
0.10688027      0.01502265

b=coeffs[1]
m=coeffs[2]

eq = function(x){x*m+b}
plot(Bradford_df$Concentration200,Bradford_df$meanAdjusted,type="l")
par(new=TRUE)
plot(eq, 1,100)

```



### C.3.2. Enzyme quantification

For Bradford to work properly After protein concentration, 900 ul Buffer A (free of imidazol) must be added to 100  $\mu$ l sample

```
##  ul of
vol=5
type="Scoe"
ABS1=0.932099998
ABS2=1.048400044
Pria_df=data.frame(vol,type,ABS1,ABS2)
Pria_df$type<-as.character(Pria_df$type)
data2=c(20,"Scoe",1.639400005,1.569700003)
data3=c(5,"D11A",0.575600028, 0.576099992)
data4=c(20,"D11A",1.039499998, 1.067399979)
data5=c(160,"test",.69085, .69085)

## add data to dataframe Pria_df
Pria_df=rbind(Pria_df,data2)
Pria_df=rbind(Pria_df,data3)
Pria_df=rbind(Pria_df,data4)
Pria_df=rbind(Pria_df,data5)

Pria_df$mean<-rowMeans(cbind(as.numeric(Pria_df$ABS1),as.numeric(Pria_df$ABS2)))

## Pasando absorbancia a concentration de proteina en 200 ul (ug/ml)
## kari values
#m=0.004679534923
#b=0.103674687

inter = function(y){(y-b)/m}
Pria_df$concentration200<-inter(Pria_df$mean)

## protein quantity on ug on 200 ul
Pria_df["Quantity"]=Pria_df$concentration200*.2

### Inicial concentration C200_vol200=Cinicial_Vinicial
Pria_df["ConcentrationInitial"]=Pria_df$Quantity/(.001*as.numeric(Pria_df$vol))

kable(Bradford_df)
```

BSAConcentrationStock	vol	Quantity	Concentration200	Abs1BSA	Abs2BSA	mean
0	160	0.0	0	0.2650	0.2948	0.27990
10	160	1.6	8	0.4329	0.4299	0.43140
20	160	3.2	16	0.6754	0.7349	0.70515
40	160	6.4	32	0.9450	0.9965	0.97075
60	160	9.6	48	1.2424	1.2260	1.23420
100	160	16.0	80	1.4320	1.4949	1.46345

```
kable(Pria_df)
```

vol	type	ABS1	ABS2	mean	concentration200	Quantity	Concentration
5	Scoe	0.932099998	1.048400044	0.99025	58.80251	11.760502	2352
20	Scoe	1.639400005	1.569700003	1.60455	99.69408	19.938816	990
5	D11A	0.575600028	0.576099992	0.57585	31.21750	6.243500	1248
20	D11A	1.039499998	1.067399979	1.05345	63.00948	12.601897	630
160	test	0.69085	0.69085	0.69085	38.87261	7.774521	48

Since  $1\text{kDa} = 1\text{Kg/mol}$  and PriA Scoe molecular weight is  $25.7\text{kDa}$  we have that

$$M \times V = \frac{g}{mw}$$

$$M = \frac{g}{V} \times \frac{1}{mw}$$

```
## mas on Kda = kg/mol
mw<-25.7
# converting to g/mol
mwgram<-mw*1000
## concentration on g/L (after convert ug/ml)
concentScoe<-2.352
concentD11A<-1.248
## Concentration on uM (10**6M)
molar = function(x){(10**6)*x/mwgram}
molar(concentScoe)
```

```
[1] 91.51751
```

```
molar(concentD11A)
```

```
[1] 48.56031
```

## C.4. Concentration on microMolar

So Scoe concentration is  $91.5175097$  and D11A is  $48.5603113 \mu M$

```
## we want final concentration 2.4uM on a final volume of 160 ul
## ciVi=CfVf
## Vi =CfVf/Vi
## Cf= 80uM Vf=5ul Vi=?ul Ci=x
```

```
adjustMol = function(x){2.5*160/x}
## this are the ul need from stock cristal enzyme that will be diluted on buffer to comp
adjustMol(molar(concentScoe))
```

```
[1] 4.370748
```

```
adjustMol(molar(concentD11A))
```

```
[1] 8.237179
```

# References

1. Khersonsky O, Tawfik DS. Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annual Review of Biochemistry*. 2010;79:471–505.
2. Copley SD. Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. *Current Opinion in Chemical Biology* [Internet]. 2003 Apr [cited 2017 Feb 8];7(2):265–72. Available from: <https://www.sciencedirect.com/science/article/pii/S1367593103000322>
3. Hult K, Berglund P. Enzyme promiscuity: Mechanism and applications. *Trends in Biotechnology* [Internet]. 2007 May [cited 2017 Feb 8];25(5):231–8. Available from: <https://www.sciencedirect.com/science/article/pii/S016777990700073X>
4. O'Brien PJ, Herschlag D. Catalytic promiscuity and the evolution of new enzymatic activities. *Chemistry & Biology* [Internet]. 1999 Apr [cited 2017 Feb 9];6(4):R91–R105. Available from: <http://www.sciencedirect.com/science/article/pii/S1074552199800337>
5. Barona Gómez F, Hodgson DA. Occurrence of a putative ancient like isomerase involved in histidine and tryptophan biosynthesis. *EMBO reports* [Internet]. 2003 Mar [cited 2017 Feb 8];4(3):296–300. Available from: <http://embor.embopress.org/content/4/3/296>
6. Risso VA, Gavira JA, Gaucher EA, Sanchez Ruiz JM. Phenotypic comparisons of consensus variants versus laboratory resurrections of precambrian proteins. *Proteins: Structure, Function, and Bioinformatics* [Internet]. 2014 Jun [cited 2017 Feb 9];82(6):887–96. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/prot.24575/abstract>
7. Kumari V, Shah S, Gupta MN. Preparation of Biodiesel by Lipase-Catalyzed Transesterification of High Free Fatty Acid Containing Oil from Madhuca indica. *Energy & Fuels* [Internet]. 2007 Jan [cited 2017 Feb 8];21(1):368–72. Available from: <http://dx.doi.org/10.1021/ef0602168>
8. Li C, Henry CS, Jankowski MD, Ionita JA, Hatzimanikatis V, Broadbelt LJ. Computational discovery of biochemical routes to specialty chemicals. *Chemical Engineering Science* [Internet]. 2004 Nov [cited 2017 Feb 8];59(22–23):5051–60.

- Available from: <https://www.sciencedirect.com/science/article/pii/S0009250904006669>
9. Glasner ME, Gerlt JA, Babbitt PC. Evolution of enzyme superfamilies. *Current Opinion in Chemical Biology* [Internet]. 2006 Oct [cited 2017 Feb 9];10(5):492–7. Available from: <https://www.sciencedirect.com/science/article/pii/S1367593106001177>
  10. Baier F, Copp JN, Tokuriki N. Evolution of Enzyme Superfamilies: Comprehensive Exploration of Sequence–Function Relationships. *Biochemistry* [Internet]. 2016 Nov [cited 2017 Feb 8];55(46):6375–88. Available from: <http://dx.doi.org/10.1021/acs.biochem.6b00723>
  11. Bloom JD, Romero PA, Lu Z, Arnold FH. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. *Biology Direct* [Internet]. 2007 [cited 2017 Feb 8];2:17. Available from: <http://dx.doi.org/10.1186/1745-6150-2-17>
  12. Nath A, Atkins WM. A Quantitative Index of Substrate Promiscuity. *Biochemistry* [Internet]. 2008 Jan [cited 2017 Feb 9];47(1):157–66. Available from: <http://dx.doi.org/10.1021/bi701448p>
  13. Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. *Molecular Biology and Evolution* [Internet]. 2015 Jan [cited 2017 Feb 9];32(1):132–43. Available from: <https://academic.oup.com/mbe/article/32/1/132/2925568/Evolution-of-Conformational-Dynamics-Determines>
  14. Firn RD, Jones CG. A Darwinian view of metabolism: Molecular properties determine fitness. *Journal of Experimental Botany* [Internet]. 2009 Mar [cited 2017 Feb 8];60(3):719–26. Available from: <https://academic.oup.com/jxb/article/60/3/719/452667/A-Darwinian-view-of-metabolism-molecular>
  15. Jia B, Cheong G-W, Zhang S. Multifunctional enzymes in archaea: Promiscuity and moonlight. *Extremophiles* [Internet]. 2013 Mar [cited 2017 Feb 8];17(2):193–203. Available from: <http://link.springer.com/article/10.1007/s00792-012-0509-1>
  16. Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, Tawfik DS. The 'evolvability' of promiscuous protein functions. *Nature Genetics* [Internet]. 2005 Jan [cited 2017 Feb 8];37(1):73–6. Available from: <http://www.nature.com/ng/journal/v37/n1/full/ng1482.html>
  17. Jensen. Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology* [Internet]. 1976 [cited 2017 Feb 8];30(1):409–25. Available from: <http://dx.doi.org/10.1146/annurev.mi.30.100176.002205>
  18. Pandya C, Farelli JD, Dunaway-Mariano D, Allen KN. Enzyme Promiscuity: Engine of Evolutionary Innovation. *Journal of Biological Chemistry* [Internet]. 2014

- Oct [cited 2017 Feb 9];289(44):30229–36. Available from: <http://www.jbc.org/content/289/44/30229>
19. Dean AM, Thornton JW. Mechanistic approaches to the study of evolution. *Nature reviews Genetics* [Internet]. 2007 Sep [cited 2017 Feb 8];8(9):675–88. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2488205/>
  20. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. *Nature Biotechnology* [Internet]. 2009 Feb [cited 2017 Feb 9];27(2):157–67. Available from: <http://www.nature.com/nbt/journal/v27/n2/full/nbt1519.html>
  21. Hopkins AL. Drug discovery: Predicting promiscuity. *Nature* [Internet]. 2009 Nov [cited 2017 Feb 8];462(7270):167–8. Available from: <http://www.nature.com/nature/journal/v462/n7270/full/462167a.html>
  22. Nath A, Zientek MA, Burke BJ, Jiang Y, Atkins WM. Quantifying and Predicting the Promiscuity and Isoform Specificity of Small-Molecule Cytochrome P450 Inhibitors. *Drug Metabolism and Disposition* [Internet]. 2010 Dec [cited 2017 Feb 9];38(12):2195–203. Available from: <http://dmd.aspetjournals.org/content/38/12/2195>
  23. Eichborn J von, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R. PROMISCUOUS: A database for network-based drug-repositioning. *Nucleic Acids Research* [Internet]. 2011 Jan [cited 2017 Feb 9];39(suppl\_1):D1060–6. Available from: [https://academic.oup.com/nar/article/39/suppl\\_1/D1060/2506056/PROMISCUOUS-a-database-for-network-based-drug](https://academic.oup.com/nar/article/39/suppl_1/D1060/2506056/PROMISCUOUS-a-database-for-network-based-drug)
  24. Zhang W, Dourado DFAR, Fernandes PA, Ramos MJ, Mannervik B. Multidimensional epistasis and fitness landscapes in enzyme evolution. *Biochemical Journal* [Internet]. 2012 Jul [cited 2017 Feb 9];445(1):39–46. Available from: <http://www.biochemj.org/content/445/1/39>
  25. Sanchez-Ruiz JM. On promiscuity, changing environments and the possibility of replaying the molecular tape of life. *Biochemical Journal* [Internet]. 2012 Jul [cited 2017 Feb 9];445(1):e1–3. Available from: <http://www.biochemj.org/content/445/1/e1>
  26. Martínez-Núñez MA, Rodríguez-Vázquez K, Pérez-Rueda E. The lifestyle of prokaryotic organisms influences the repertoire of promiscuous enzymes. *Proteins: Structure, Function, and Bioinformatics* [Internet]. 2015 Sep [cited 2017 Feb 8];83(9):1625–31. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/prot.24847/abstract>
  27. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. Multicopy Suppression Underpins Metabolic Evolvability. *Molecular Biology and Evolution* [Internet]. 2007 Dec [cited 2017 Feb 9];24(12):2716–22. Available from: <https://academic.oup.com/mbe/article/24/12/2716/978039/Multicopy-Suppression-Underpins-Metabolic-Evolvability>

**Suppression-Underpins-Metabolic**

28. Notebaart RA, Szappanos B, Kintses B, Pál F, Györkei Á, Bogos B, et al. Network-level architecture and the evolutionary potential of underground metabolism. *Proceedings of the National Academy of Sciences* [Internet]. 2014 Aug [cited 2017 Feb 9];111(32):11762–7. Available from: <http://www.pnas.org/content/111/32/11762>
29. Linster CL, Van Schaftingen E, Hanson AD. Metabolite damage and its repair or pre-emption. *Nature Chemical Biology* [Internet]. 2013 Feb [cited 2017 Feb 8];9(2):72–80. Available from: <http://www.nature.com/nchembio/journal/v9/n2/full/nchembio.1141.html>
30. Khanal A, Yu McLoughlin S, Kershner JP, Copley SD. Differential Effects of a Mutation on the Normal and Promiscuous Activities of Orthologs: Implications for Natural and Directed Evolution. *Molecular Biology and Evolution* [Internet]. 2015 Jan [cited 2017 Feb 8];32(1):100–8. Available from: <https://academic.oup.com/mbe/article/32/1/100/2925554/Differential-Effects-of-a-Mutation-on-the-Normal>
31. Ma H-M, Zhou Q, Tang Y-M, Zhang Z, Chen Y-S, He H-Y, et al. Unconventional Origin and Hybrid System for Construction of Pyrrolopyrrole Moiety in Kosinostatin Biosynthesis. *Chemistry & Biology* [Internet]. 2013 Jun [cited 2017 Feb 8];20(6):796–805. Available from: <https://www.sciencedirect.com/science/article/pii/S1074552113001701>
32. Adams NE, Thiaville JJ, Proestos J, Juárez-Vázquez AL, McCoy AJ, Barona-Gómez F, et al. Promiscuous and Adaptable Enzymes Fill “Holes” in the Tetrahydrofolate Pathway in Chlamydia Species. *mBio* [Internet]. 2014 Jul [cited 2017 Jan 31];5(4). Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4161248/>
33. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics* [Internet]. 2010 Aug [cited 2017 Feb 9];11(8):572–82. Available from: <http://www.nature.com/nrg/journal/v11/n8/full/nrg2808.html>
34. Halachev MR, Loman NJ, Pallen MJ. Calculating Orthologs in Bacteria and Archaea: A Divide and Conquer Approach. *PLOS ONE* [Internet]. 2011 Dec [cited 2016 Sep 16];6(12):e28388. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028388>
35. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: An integrative view of gene diversity within microbial populations. *BMC Genomics* [Internet]. 2011 [cited 2017 Jan 26];12:32. Available from: <http://dx.doi.org/10.1186/1471-2164-12-32>
36. Carbonell P, Faulon J-L. Molecular signatures-based prediction of enzyme promiscuity. *Bioinformatics* [Internet]. 2010 Aug [cited 2017 Feb 8];26(16):2012–9.

- Available from: <https://academic.oup.com/bioinformatics/article/26/16/2012/215921/Molecular-signatures-based-prediction-of-enzyme>
37. Cheng X-Y, Huang W-J, Hu S-C, Zhang H-L, Wang H, Zhang J-X, et al. A Global Characterization and Identification of Multifunctional Enzymes. PLoS ONE [Internet]. 2012 Jun [cited 2017 Feb 8];7(6). Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3377604/>
  38. Nagao C, Nagano N, Mizuguchi K. Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests. PLOS ONE [Internet]. 2014 Jan [cited 2017 Feb 8];9(1):e84623. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0084623>
  39. Noda-García L, Juárez-Vázquez AL, Ávila-Arcos MC, Verduzco-Castro EA, Montero-Morán G, Gaytán P, et al. Insights into the evolution of enzyme substrate promiscuity after the discovery of  $\beta\alpha_8$  isomerase evolutionary intermediates from a diverse metagenome. BMC Evolutionary Biology [Internet]. 2015 Jun [cited 2017 Jan 31];15. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4462073/>
  40. Garcia-Seisdedos H, Ibarra-Molero B, Sanchez-Ruiz JM. Probing the Mutational Interplay between Primary and Promiscuous Protein Functions: A Computational-Experimental Approach. PLOS Computational Biology [Internet]. 2012 Jun [cited 2017 Feb 8];8(6):e1002558. Available from: <http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002558>
  41. Noda-García L, Camacho-Zarco AR, Medina-Ruiz S, Gaytán P, Carrillo-Tripp M, Fülöp V, et al. Evolution of Substrate Specificity in a Recipient's Enzyme Following Horizontal Gene Transfer. Molecular Biology and Evolution [Internet]. 2013 Sep [cited 2017 Jan 31];30(9):2024–34. Available from: <https://academic.oup.com/mbe/article/30/9/2024/1000280/Evolution-of-Substrate-Specificity-in-a-Recipient>
  42. Verdel-Aranda K, López-Cortina ST, Hodgson DA, Barona-Gómez F. Molecular annotation of ketol-acid reductoisomerases from Streptomyces reveals a novel amino acid biosynthesis interlock mediated by enzyme promiscuity. Microbial Biotechnology [Internet]. 2015 Mar [cited 2017 Feb 9];8(2):239–52. Available from: <http://onlinelibrary.wiley.com/doi/10.1111/1751-7915.12175/abstract>
  43. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nature Methods [Internet]. 2007 Oct [cited 2017 Feb 9];4(10):787–97. Available from: <http://www.nature.com/nmeth/journal/v4/n10/abs/nmeth1088.html>
  44. Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nature Chemical Biology [Internet]. 2015 Sep [cited 2017 Jan 24];11(9):639–48. Available from: <http://www.nature.com/nchembio/journal/v11/n9/full/>

- nchembio.1884.html
45. Campbell I. Biophysical Techniques - Paperback - Iain D. Campbell - Oxford University Press. 2012.
  46. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, et al. Molecular Networking as a Dereplication Strategy. *Journal of Natural Products* [Internet]. 2013 Sep [cited 2017 Feb 9];76(9):1686–99. Available from: <http://dx.doi.org/10.1021/np400413s>
  47. Köcher T, Superti-Furga G. Mass spectrometry-based functional proteomics: From molecular machines to protein networks. *Nature Methods* [Internet]. 2007 Oct [cited 2017 Feb 10];4(10):807–15. Available from: <http://www.nature.com/nmeth/journal/v4/n10/full/nmeth1093.html>
  48. James LC, Tawfik DS. Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. *Trends in Biochemical Sciences* [Internet]. 2003 Jul;28(7):361–8. Available from: <https://www.sciencedirect.com/science/article/pii/S096800040300135X>
  49. Parisi G, Zea DJ, Monzon AM, Marino-Buslje C. Conformational diversity and the emergence of sequence signatures during evolution. *Current Opinion in Structural Biology* [Internet]. 2015 Jun [cited 2017 Feb 9];32:58–65. Available from: <https://www.sciencedirect.com/science/article/pii/S09594440X15000147>
  50. Javier Zea D, Miguel Monzon A, Fornasari MS, Marino-Buslje C, Parisi G. Protein Conformational Diversity Correlates with Evolutionary Rate. *Molecular Biology and Evolution* [Internet]. 2013 Jul;30(7):1500–3. Available from: <https://academic.oup.com/mbe/article/30/7/1500/972515/Protein-Conformational-Diversity-Correlates-with>
  51. Gatti-Lafranconi P, Hollfelder F. Flexibility and Reactivity in Promiscuous Enzymes. *ChemBioChem* [Internet]. 2013 Feb [cited 2017 Feb 8];14(3):285–92. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/cbic.201200628/abstract>
  52. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, Ramos-Aboites H, et al. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. *Genome Biology and Evolution* [Internet]. 2016 Jun [cited 2017 Jan 24];8(6):1906–16. Available from: <http://gbe.oxfordjournals.org/content/8/6/1906>
  53. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research* [Internet]. 2003 Sep [cited 2017 Feb 8];13(9):2178–89. Available from: <http://genome.cshlp.org/content/13/9/2178>
  54. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB:

- A hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Research [Internet]. 2013 Jan [cited 2017 Feb 9];41(D1):D358–65. Available from: <https://academic.oup.com/nar/article/41/D1/D358/1060216/OrthoDB-a-hierarchical-catalog-of-animal-fungal>
55. Gao B, Gupta RS. Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria. *Microbiology and Molecular Biology Reviews : MMBR* [Internet]. 2012 Mar [cited 2017 Feb 8];76(1):66–112. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3294427/>
56. Sen A, Daubin V, Abrouk D, Gifford I, Berry AM, Normand P. Phylogeny of the class Actinobacteria revisited in the light of complete genomes. The orders “Frankiales” and Micrococcales should be split into coherent entities: Proposal of Frankiales ord. nov., Geodermatophilales ord. nov., Acidothermales ord. nov. and Nakamurellales ord. nov. *International Journal of Systematic and Evolutionary Microbiology* [Internet]. 2014 [cited 2017 Feb 9];64(11):3821–32. Available from: <http://ijs.microbiologyresearch.org/content/journal/ijsem/10.1099/ijs.0.063966-0>
57. Zhou Z, Gu J, Li Y-Q, Wang Y. Genome plasticity and systems evolution in Streptomyces. *BMC Bioinformatics* [Internet]. 2012 [cited 2017 Feb 9];13(10):S8. Available from: <http://dx.doi.org/10.1186/1471-2105-13-S10-S8>
58. Kim J-N, Kim Y, Jeong Y, Roe J-H, Kim B-G, Cho B-K. Comparative Genomics Reveals the Core and Accessory Genomes of Streptomyces Species. *Journal of Microbiology and Biotechnology*. 2015 Oct;25(10):1599–605.
59. Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, et al. Network Context and Selection in the Evolution to Enzyme Specificity. *Science* [Internet]. 2012 Aug [cited 2017 Feb 9];337(6098):1101–4. Available from: <http://science.sciencemag.org/content/337/6098/1101>
60. Copley SD. An Evolutionary Biochemist’s Perspective on Promiscuity. *Trends in biochemical sciences* [Internet]. 2015 Feb [cited 2017 Feb 8];40(2):72–8. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4836852/>
61. Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, et al. Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature* [Internet]. 2013 Oct [cited 2017 Feb 9];502(7473):698–702. Available from: <http://www.nature.com/nature/journal/v502/n7473/full/nature12576.html>
62. Hughes AL. The Evolution of Functionally Novel Proteins after Gene Duplication. *Proceedings of the Royal Society of London B: Biological Sciences* [Internet]. 1994 May;256(1346):119–24. Available from: <http://rspb.royalsocietypublishing.org/content/256/1346/119>
63. Divergent Evolution of Enzymatic Function: Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies. *Annual Review of Biochemistry* [Internet]. 2001 [cited 2017 Feb 8];70(1):209–46. Available from: <http://dx.doi.org/10.1146/annurev.biochem.70.1.209>

- 1146/annrev.biochem.70.1.209
64. Huang R, Hippauf F, Rohrbeck D, Haustein M, Wenke K, Feike J, et al. Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. *Proceedings of the National Academy of Sciences* [Internet]. 2012 Feb;109(8):2966–71. Available from: <http://www.pnas.org/content/109/8/2966>
  65. Fondi M, Emiliani G, Liò P, Gribaldo S, Fani R. The evolution of histidine biosynthesis in archaea: Insights into the his genes structure and organization in LUCA. *Journal of Molecular Evolution*. 2009 Nov;69(5):512–26.
  66. Merino E, Jensen RA, Yanofsky C. Evolution of bacterial trp operons and their regulation. *Current opinion in microbiology* [Internet]. 2008 Apr [cited 2017 Feb 10];11(2):78–86. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2387123/>
  67. Verduzco-Castro EA, Michalska K, Endres M, Juárez-Vazquez AL, Noda-García L, Chang C, et al. Co-occurrence of analogous enzymes determines evolution of a novel  $\beta\alpha_8$ -isomerase sub-family after non-conserved mutations in flexible loop. *Biochemical Journal* [Internet]. 2016 May;473(9):1141–52. Available from: <http://www.biochemj.org/content/473/9/1141>
  68. Lamble HJ, Heyer NI, Bull SD, Hough DW, Danson MJ. Metabolic Pathway Promiscuity in the Archaeon *Sulfolobus solfataricus* Revealed by Studies on Glucose Dehydrogenase and 2-Keto-3-deoxygluconate Aldolase. *Journal of Biological Chemistry* [Internet]. 2003 Sep [cited 2019 Jan 25];278(36):34066–72. Available from: <http://www.jbc.org/content/278/36/34066>
  69. Weng J-K, Noel JP. The Remarkable Pliability and Promiscuity of Specialized Metabolism. *Cold Spring Harbor Symposia on Quantitative Biology* [Internet]. 2012 Jan [cited 2019 Jan 24];77:309–20. Available from: <http://symposium.cshlp.org/content/77/309>
  70. Juárez-Vázquez AL, Edirisinghe JN, Verduzco-Castro EA, Michalska K, Wu C, Noda-García L, et al. Evolution of substrate specificity in a retained enzyme driven by gene loss. *eLife* [Internet]. [cited 2018 Jan 16];6. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5404923/>
  71. Noda-García L. Estudio de la evolución molecular de la función enzimática usando como modelo una enzima con características ancestrales [PhD thesis]. [Irapuato, GTO]: Langebio, CINVESTAV; 2012.
  72. Zhao S, Sakai A, Zhang X, Vetting MW, Kumar R, Hillerich B, et al. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. *eLife* [Internet]. 2014 Jun [cited 2017 Feb 9];3:e03275. Available from: <https://elifesciences.org/content/3/e03275v2>
  73. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein–protein interaction networks, integrated over the tree of

- life. Nucleic Acids Research [Internet]. 2015 Jan [cited 2017 Feb 9];43(D1):D447–52. Available from: <https://academic.oup.com/nar/article/43/D1/D447/2435295/STRING-v10-protein-protein-interaction-networks>
74. Segata N, Börnigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nature Communications [Internet]. 2013 Aug [cited 2019 Jan 24];4:2304. Available from: <https://www.nature.com/articles/ncomms3304>
75. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, et al. Insights from 20 years of bacterial genome sequencing. Functional & Integrative Genomics [Internet]. 2015 [cited 2017 Jan 28];15(2):141–61. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4361730/>
76. Koonin EV. The Turbulent Network Dynamics of Microbial Evolution and the Statistical Tree of Life. Journal of Molecular Evolution [Internet]. 2015 [cited 2017 Jan 28];80(5-6):244–50. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4472940/>
77. Veen BE van der, Harris HM, O'Toole PW, Claesson MJ. Metaphor: Finding Bi-directional Best Hit homology relationships in (meta)genomic datasets. Genomics [Internet]. 2014 Dec [cited 2018 Jul 3];104(6, Part B):459–63. Available from: <http://www.sciencedirect.com/science/article/pii/S0888754314002092>
78. Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mittenhal JE. The origin and evolution of modern metabolism. The International Journal of Biochemistry & Cell Biology [Internet]. 2009 Feb [cited 2018 Jul 3];41(2):285–97. Available from: <http://www.sciencedirect.com/science/article/pii/S1357272508003373>
79. Schniete JK, Cruz-Morales P, Selem-Mojica N, Fernández-Martínez LT, Hunter IS, Barona-Gómez F, et al. Expanding Primary Metabolism Helps Generate the Metabolic Robustness To Facilitate Antibiotic Biosynthesis in Streptomyces. mBio [Internet]. 2018 Mar [cited 2018 Aug 10];9(1):e02283–17. Available from: <http://mbio.asm.org/content/9/1/e02283-17>
80. Alanjary M, Kronmiller B, Adamek M, Blin K, Weber T, Huson D, et al. The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. Nucleic Acids Research [Internet]. 2017 Jul [cited 2018 Jan 16];45(W1):W42–8. Available from: <https://academic.oup.com/nar/article/45/W1/W42/3787867>
81. Martínez-Núñez MA, Rodríguez-Escamilla Z, Rodríguez-Vázquez K, Pérez-Rueda E. Tracing the Repertoire of Promiscuous Enzymes along the Metabolic Pathways in Archaeal Organisms. Life [Internet]. 2017 Jul [cited 2019 Jan 24];7(3). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5617955/>
82. Pearson H. Prehistoric proteins: Raising the dead. Nature News [Internet]. 2012 Mar [cited 2017 Feb 9];483(7390):390. Available from: <http://www.nature.com/>

- news/prehistoric-proteins-raising-the-dead-1.10261
83. Treangen TJ, Rocha EPC. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLOS Genetics* [Internet]. 2011 Jan [cited 2017 Feb 9];7(1):e1001284. Available from: <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1001284>
  84. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences* [Internet]. 1999 Mar [cited 2017 Feb 9];96(6):2896–901. Available from: <http://www.pnas.org/content/96/6/2896>
  85. Snel B, Lehmann G, Bork P, Huynen MA. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Research* [Internet]. 2000 Sep [cited 2017 Feb 9];28(18):3442–4. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC110752/>
  86. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* [Internet]. 2008 Feb [cited 2017 Feb 7];9:75. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2265698/>
  87. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). *Nucleic Acids Research* [Internet]. 2014 Jan [cited 2017 Feb 7];42(Database issue):D206–14. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965101/>
  88. Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery* [Internet]. 2015 Feb [cited 2016 Sep 16];14(2):111–29. Available from: <http://www.nature.com/nrd/journal/v14/n2/full/nrd4510.html>
  89. Petrenko R, Meller J. Molecular Dynamics. In: eLS [Internet]. John Wiley & Sons, Ltd; 2001. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0003048.pub2/abstract>
  90. Molecular Modeling of Proteins Andreas Kukol Springer [Internet]. [cited 2017 Feb 8]. Available from: <http://www.springer.com/us/book/9781588298645>
  91. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of The Royal Society Interface* [Internet]. 2014 Nov [cited 2017 Feb 9];11(100):20140419. Available from: <http://rsif.royalsocietypublishing.org/content/11/100/20140419>
  92. Zhou R. Replica Exchange Molecular Dynamics Method for Protein Folding Simulation. In: Bai Y, Nussinov R, editors. *Protein Folding Protocols* [Internet]. Humana Press; 2006 [cited 2017 Feb 9]. pp. 205–23. (Methods in Molecular Biology<sup>TM</sup>).

- Available from: <http://dx.doi.org/10.1385/1-59745-189-4%3A205>
93. Bisswanger H. General Aspects of Enzyme Analysis. In: Practical Enzymology [Internet]. Wiley-VCH Verlag GmbH & Co. KGaA; 2011 [cited 2017 Feb 8]. pp. 5–91. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/9783527659227.ch2/summary>
  94. Hommel U, Eberhard M, Kirschner K. Phosphoribosyl Anthranilate Isomerase Catalyzes a Reversible Amadori Reaction. Biochemistry [Internet]. 1995 Apr [cited 2017 Feb 8];34(16):5429–39. Available from: <http://dx.doi.org/10.1021/bi00016a014>
  95. Scheer M, Grote A, Chang A, Schomburg I, Munaretto C, Rother M, et al. BRENDA, the enzyme information system in 2011. Nucleic Acids Research [Internet]. 2011 Jan [cited 2017 Feb 9];39(suppl\_1):D670–6. Available from: [https://academic.oup.com/nar/article/39/suppl\\_1/D670/2506543/BRENDA-the-enzyme-information-system-in-2011](https://academic.oup.com/nar/article/39/suppl_1/D670/2506543/BRENDA-the-enzyme-information-system-in-2011)
  96. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. Journal of Computational Chemistry [Internet]. 2005 Dec [cited 2017 Feb 9];26(16):1701–18. Available from: <http://onlinelibrary.wiley.com/doi/10.1002/jcc.20291/abstract>
  97. Odokonyero D, Sakai A, Patkovsky Y, Malashkevich VN, Fedorov AA, Bonanno JB, et al. Loss of quaternary structure is associated with rapid sequence divergence in the OSBS family. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 2014 Jun [cited 2017 Feb 9];111(23):8535–40. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4060685/>
  98. Osbourn A. Gene Clusters for Secondary Metabolic Pathways: An Emerging Theme in Plant Biology. Plant Physiology [Internet]. 2010 Oct [cited 2016 Sep 13];154(2):531–5. Available from: <http://www.plantphysiol.org/content/154/2/531>
  99. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, et al. Comparative Genomics of the Archaea (Euryarchaeota): Evolution of Conserved Protein Families, the Stable Core, and the Variable Shell. Genome Research [Internet]. 1999 Jul [cited 2017 Feb 12];9(7):608–28. Available from: <http://genome.cshlp.org/content/9/7/608>
  100. Benedict MN, Gonnerman MC, Metcalf WW, Price ND. Genome-Scale Metabolic Reconstruction and Hypothesis Testing in the Methanogenic Archaeon Methanosaeca acetivorans C2A. Journal of Bacteriology [Internet]. 2012 Feb [cited 2016 Sep 16];194(4):855–65. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3272958/>
  101. Seitz KW, Lazar CS, Hinrichs K-U, Teske AP, Baker BJ. Genomic reconstruction of a novel, deeply branched sediment archaeal phylum with pathways for acetogenesis and sulfur reduction. The ISME Journal [Internet]. 2016 Jul [cited

- 2016 Sep 16];10(7):1696–705. Available from: <http://www.nature.com/ismej/journal/v10/n7/full/ismej2015233a.html>
102. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, et al. MINEs: Open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. *Journal of Cheminformatics [Internet]*. 2015 Dec [cited 2016 Sep 16];7(1). Available from: <http://www.jcheminf.com/content/7/1/44>
103. Moustafa A, Loram JE, Hackett JD, Anderson DM, Plumley FG, Bhattacharya D. Origin of Saxitoxin Biosynthetic Genes in Cyanobacteria. *PLOS ONE [Internet]*. 2009 Jun [cited 2016 Sep 16];4(6):e5758. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0005758>
104. Medema MH, Osbourn A. Computational genomic identification and functional reconstitution of plant natural product biosynthetic pathways. *Natural Product Reports*. 2016 Aug;33(8):951–62.
105. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology [Internet]*. 2015 Sep [cited 2017 Feb 12];11(9):625–31. Available from: <http://www.nature.com/nchembio/journal/v11/n9/full/nchembio.1890.html>
106. Iqbal HA, Low-Beinart L, Obiajulu JU, Brady SF. Natural Product Discovery through Improved Functional Metagenomics in Streptomyces. *Journal of the American Chemical Society [Internet]*. 2016 Aug [cited 2017 Feb 12];138(30):9341–4. Available from: <http://dx.doi.org/10.1021/jacs.6b02921>
107. Ulas T, Riemer SA, Zaparty M, Siebers B, Schomburg D. Genome-Scale Reconstruction and Analysis of the Metabolic Network in the Hyperthermophilic Archaeon *Sulfolobus Solfataricus*. *PLoS ONE [Internet]*. 2012 Aug [cited 2016 Sep 22];7(8). Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3432047/>
108. Charlesworth JC, Burns BP. Untapped Resources: Biotechnological Potential of Peptides and Secondary Metabolites in Archaea. *Archaea [Internet]*. 2015 Oct [cited 2016 Sep 27];2015:e282035. Available from: <http://www.hindawi.com/journals/archaea/2015/282035/abs/>
109. Computational Pan-Genomics Consortium. Computational pan-genomics: Status, promises and challenges. *Briefings in Bioinformatics*. 2016 Oct;
110. Chan C, Jayasekera S, Kao B, Páramo M, Grotthuss M von, Ranz JM. Remodelling of a homeobox gene cluster by multiple independent gene reunions in *Drosophila*. *Nature Communications [Internet]*. 2015 Mar [cited 2016 Dec 7];6:6509. Available from: <http://www.nature.com/ncomms/2015/150305/ncomms7509/full/ncomms7509.html>
111. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et

- al. Insights into the phylogeny and coding potential of microbial dark matter. *Nature* [Internet]. 2013 Jul [cited 2016 Dec 7];499(7459):431–7. Available from: <http://www.nature.com/nature/journal/v499/n7459/full/nature12352.html?cookies=accepted>
112. Castelle CJ, Wrighton KC, Thomas BC, Hug LA, Brown CT, Wilkins MJ, et al. Genomic Expansion of Domain Archaea Highlights Roles for Organisms from New Phyla in Anaerobic Carbon Cycling. *Current Biology* [Internet]. 2015 Mar [cited 2016 Dec 7];25(6):690–701. Available from: <http://www.sciencedirect.com/science/article/pii/S0960982215000160>
113. Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, et al. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* [Internet]. 2015 May [cited 2016 Dec 7];521(7551):173–9. Available from: <http://www.nature.com/nature/journal/v521/n7551/full/nature14447.html>
114. Koonin EV. Archaeal ancestors of eukaryotes: Not so elusive any more. *BMC Biology* [Internet]. 2015 Oct [cited 2016 Dec 7];13. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4594999/>
115. Chaudhari NM, Gupta VK, Dutta C. BPGA- an ultra-fast pan-genome analysis pipeline. *Scientific Reports* [Internet]. 2016 Apr [cited 2016 Dec 7];6:24373. Available from: <http://www.nature.com/srep/2016/160413/srep24373/full/srep24373.html>
116. Glass JI, Assad-Garcia N, Alperovich N, Yooseph S, Lewis MR, Maruf M, et al. Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America* [Internet]. 2006 Jan [cited 2016 Dec 7];103(2):425–30. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1324956/>
117. Narechania A, Baker RH, Sit R, Kolokotronics S-O, DeSalle R, Planet PJ. Random Addition Concatenation Analysis: A Novel Approach to the Exploration of Phylogenomic Signal Reveals Strong Agreement between Core and Shell Genomic Partitions in the Cyanobacteria. *Genome Biology and Evolution* [Internet]. 2012 Jan [cited 2017 Jan 23];4(1):30–43. Available from: <http://gbe.oxfordjournals.org/content/4/1/30>
118. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* [Internet]. 2014 May [cited 2016 Dec 22];30(9):1312–3. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998144/>
119. Powerful tree graphics with ggplot2 [Internet]. [cited 2017 Jan 28]. Available from: [http://joey711.github.io/phyloseq/plot\\_tree-examples.html](http://joey711.github.io/phyloseq/plot_tree-examples.html)
120. Zacharia VM, Traxler MF. Exploring new horizons. *eLife* [Internet]. 2017 Jan [cited 2017 Jan 23];6:e23624. Available from: <https://elifesciences.org/content/6/>

e23624v1

121. Woese C. The universal ancestor. *Proceedings of the National Academy of Sciences* [Internet]. 1998 Jun [cited 2017 Feb 13];95(12):6854–9. Available from: <http://www.pnas.org/content/95/12/6854>
122. Woese CR, Gupta R. Are archaebacteria merely derived “prokaryotes”? *Nature* [Internet]. 1981 Jan [cited 2017 Feb 13];289(5793):95–6. Available from: <http://www.nature.com/nature/journal/v289/n5793/abs/289095a0.html>
123. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences of the United States of America*. 1990 Jun;87(12):4576–9.
124. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America* [Internet]. 1977 Nov [cited 2017 Jan 23];74(11):5088–90. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC432104/>
125. Woese CR. There must be a prokaryote somewhere: Microbiology’s search for itself. *Microbiological Reviews* [Internet]. 1994 Mar [cited 2017 Jan 23];58(1):1–9. Available from: <http://mmbrr.asm.org/content/58/1/1>
126. Graham DE, Overbeek R, Olsen GJ, Woese CR. An archaeal genomic signature. *Proceedings of the National Academy of Sciences* [Internet]. 2000 Mar [cited 2017 Feb 13];97(7):3304–8. Available from: <http://www.pnas.org/content/97/7/3304>
127. Howland JL. *The surprising archaea: Discovering another domain of life*. New York: Oxford University; 2000.
128. Xu Y, Gogarten JP. *Computational Methods for Understanding Bacterial and Archaeal Genomes*. World Scientific; 2008.
129. Garrett RA, Klenk H-P. *Archaea: Evolution, Physiology, and Molecular Biology*. John Wiley & Sons; 2008.
130. Nishida H. Evolution of genome base composition and genome size in bacteria. *Frontiers in Microbiology* [Internet]. 2012 Dec [cited 2017 Jan 28];3. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3515811/>
131. Coyle M, Hu J, Gartner Z. Mysteries in a Minimal Genome. *ACS Central Science* [Internet]. 2016 May [cited 2017 Feb 13];2(5):274–7. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4882734/>
132. O’Meara B. CRAN Task View: Phylogenetics, Especially Comparative Methods. 2016 Dec [cited 2017 Jan 28]; Available from: <https://CRAN.R-project.org/>

- view=Phylogenetics
133. Larsson J, Nylander JA, Bergman B. Genome fluctuations in cyanobacteria reflect evolutionary, developmental and adaptive traits. *BMC Evolutionary Biology* [Internet]. 2011 [cited 2017 Feb 2];11:187. Available from: <http://dx.doi.org/10.1186/1471-2148-11-187>
  134. Whitton BA. *Ecology of Cyanobacteria II: Their Diversity in Space and Time*. Springer Science & Business Media; 2012.
  135. Cohen GN. The biosynthesis of histidine and its regulation. In: *Microbial Biochemistry* [Internet]. Springer Netherlands; 2004 [cited 2017 Jan 31]. pp. 225–30. Available from: [http://link.springer.com/chapter/10.1007/978-1-4020-2237-1\\_29](http://link.springer.com/chapter/10.1007/978-1-4020-2237-1_29)
  136. Plach MG, Reisinger B, Sterner R, Merkl R. Long-Term Persistence of Bi-functionality Contributes to the Robustness of Microbial Life through Exaptation. *PLOS Genetics* [Internet]. 2016 Jan [cited 2017 Feb 13];12(1):e1005836. Available from: <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1005836>
  137. Battistuzzi FU, Feijao A, Hedges SB. A genomic timescale of prokaryote evolution: Insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology* [Internet]. 2004 [cited 2017 Feb 2];4:44. Available from: <http://dx.doi.org/10.1186/1471-2148-4-44>
  138. Lapierre P, Gogarten JP. Estimating the size of the bacterial pan-genome. *Trends in Genetics* [Internet]. 2009 Mar [cited 2017 Feb 3];25(3):107–10. Available from: [http://www.cell.com/trends/genetics/abstract/S0168-9525\(09\)00005-5](http://www.cell.com/trends/genetics/abstract/S0168-9525(09)00005-5)
  139. Větrovský T, Baldrian P. The Variability of the 16S rRNA Gene in Bacterial Genomes and Its Consequences for Bacterial Community Analyses. *PLoS ONE* [Internet]. 2013 Feb [cited 2017 Feb 8];8(2). Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3583900/>
  140. Rodríguez-Orduña L. Microbioma del jitomate: Determinante en la infección por *Clavibacter michiganensis* subsp. *michiganensis* [PhD thesis]. [Irapuato, GTO]: Langebio, CINVESTAV; 2016.
  141. Gutiérrez-García K, Bustos-Díaz ED, Corona-Gómez JA, Ramos-Aboites HE, Sélem-Mojica N, Cruz-Morales P, et al. Cycad Coralloid Roots Contain Bacterial Communities Including Cyanobacteria and Caulobacter spp. That Encode Niche-Specific Biosynthetic Gene Clusters. *Genome Biology and Evolution* [Internet]. 2019 Jan;11(1):319–34. Available from: <https://academic.oup.com/gbe/article/11/1/319/5238076>
  142. Delgado-Suárez EJ, Selem-Mojica N, Ortiz-López R, Gebreyes WA, Allard MW, Barona-Gómez F, et al. Whole genome sequencing reveals widespread distribution

- of typhoidal toxin genes and VirB/D4 plasmids in bovine-associated nontyphoidal *Salmonella*. *Scientific Reports* [Internet]. 2018 Jun;8(1):9864. Available from: <https://doi.org/10.1038/s41598-018-28169-4>
143. Labeda DP, Dunlap CA, Rong X, Huang Y, Doroghazi JR, Ju K-S, et al. Phylogenetic relationships in the family Streptomycetaceae using multi-locus sequence analysis. *Antonie van Leeuwenhoek* [Internet]. 2017 Apr [cited 2019 Mar 13];110(4):563–83. Available from: <https://doi.org/10.1007/s10482-016-0824-0>
144. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Research* [Internet]. 2014 Jan;42(Database issue):D581–91. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965095/>
145. Contreras-Moreira B, Vinuesa P. GET\_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. *Applied and Environmental Microbiology* [Internet]. 2013 Dec;79(24):7696–701. Available from: <https://aem.asm.org/content/79/24/7696>
146. Yasuhara-Bell J, Marrero G, Alvarez AM. Genes clvA, clvF and clvG are unique to *Clavibacter michiganensis* subsp. *michiganensis* and highly conserved. *European Journal of Plant Pathology* [Internet]. 2014 Dec [cited 2019 Mar 13];140(4):655–64. Available from: <https://doi.org/10.1007/s10658-014-0495-5>
147. Garcia-Pichel F, Sherry ND, Castenholz RW. Evidence for an ultraviolet sunscreen role of the extra cellular pigment scytonemin in the terrestrial cyanobacterium *Chlorogloeopsis* sp. *Photochemistry and Photobiology* [Internet]. 1992 Jul;56(1):17–23. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-1097.1992.tb09596.x>
148. Balskus EP, Walsh CT. Investigating the Initial Steps in the Biosynthesis of Cyanobacterial Sunscreen Scytonemin. *Journal of the American Chemical Society* [Internet]. 2008 Nov;130(46):15260–1. Available from: <https://doi.org/10.1021/ja807192u>
149. Soule T, Palmer K, Gao Q, Potrafka RM, Stout V, Garcia-Pichel F. A comparative genomics approach to understanding the biosynthesis of the sunscreen scytonemin in cyanobacteria. *BMC Genomics* [Internet]. 2009 Jul [cited 2018 Jun 27];10:336. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2726228/>
150. Engel PC. Glutamate Dehydrogenases: The Why and How of Coenzyme Specificity. *Neurochemical Research* [Internet]. 2014 Mar;39(3):426–32. Available from: <https://doi.org/10.1007/s11064-013-1089-x>
151. Liu Y, Li Y, Wang X. Acetohydroxyacid synthases: Evolution, structure, and function. *Applied Microbiology and Biotechnology* [Internet]. 2016 Oct;100(20):8633–49. Available from: <https://doi.org/10.1007/s00253-016-7809-9>
152. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH

- 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research [Internet]. 2015 Jul [cited 2017 Feb 7];43(W1):W237–43. Available from: <https://academic.oup.com/nar/article/43/W1/W237/2467910/antiSMASH-3-0-a-comprehensive-resource-for-the>
153. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE [Internet]. 2010 Mar [cited 2018 Jul 3];5(3):e9490. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009490>
154. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, et al. Miicroreact: Visualizing and sharing data for genomic epidemiology and phylogeography. Microbial Genomics. 2016 Nov;2(11).
155. Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV. Lineage-Specific Gene Expansions in Bacterial and Archaeal Genomes. Genome Research [Internet]. 2001 Apr;11(4):555–65. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC311027/>
156. Charlesworth JC, Burns BP. Untapped Resources: Biotechnological Potential of Peptides and Secondary Metabolites in Archaea. Archaea [Internet]. 2015 Oct [cited 2016 Sep 27];2015:e282035. Available from: <http://www.hindawi.com/journals/archaea/2015/282035/abs/>
157. Koonin EV, Wolf YI. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. Nucleic Acids Research [Internet]. 2008 Dec [cited 2017 Jan 26];36(21):6688–719. Available from: <https://academic.oup.com/nar/article/36/21/6688/2410005/Genomics-of-bacteria-and-archaea-the-emerging>
158. Kudo F, Kasama Y, Hirayama T, Eguchi T. Cloning of the Pactamycin Biosynthetic Gene Cluster and Characterization of a Crucial Glycosyltransferase Prior to a Unique Cyclopentane Ring Formation. The Journal of Antibiotics [Internet]. 2007 Aug;60(8):492–503. Available from: <https://www.nature.com/articles/ja200763>
159. Navarro-Muñoz J, Selem-Mojica N, Mullowney M, Kautsar S, Tryon J, Parkinson E, et al. A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. bioRxiv [Internet]. 2018 Oct;445270. Available from: <https://www.biorxiv.org/content/early/2018/10/17/445270>
160. Lilley KS, Baker PJ, Linda Britton K, Stillman TJ, Brown PE, Moir AJG, et al. The partial amino acid sequence of the NAD -dependent glutamate dehydrogenase of Clostridium symbiosum: Implications for the evolution and structural basis of coenzyme specificity. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology [Internet]. 1991 Nov;1080(3):191–7. Available from: <http://www.sciencedirect.com/science/article/pii/016748389190001G>
161. Consalvi V, Chiarylce R, Politi L, Gambacorta A, Rosa MD, Scandurra R.

- Glutamate dehydrogenase from the thermoacidophilic archaebacterium *Sulfolobus solfataricus*. European Journal of Biochemistry [Internet]. 1991 Mar;196(2):459–67. Available from: <https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1432-1033.1991.tb15837.x>
162. Ferrer J, Pérez-Pomares F, Bonete MJ. NADP-glutamate dehydrogenase from the halophilic archaeon *Haloferax mediterranei*: Enzyme purification, N-terminal sequence and stability. FEMS microbiology letters. 1996 Jul;141(1):59–63.
163. Tholl D. Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. Current Opinion in Plant Biology [Internet]. 2006 Jun;9(3):297–304. Available from: <http://www.sciencedirect.com/science/article/pii/S1369526606000537>
164. Balskus EP, Walsh CT. The genetic and molecular basis for sunscreen biosynthesis in cyanobacteria. Science (New York, NY) [Internet]. 2010 Sep;329(5999):1653–6. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3116657/>
165. Hillwig ML, Fuhrman HA, Ittiamornkul K, Sevco TJ, Kwak DH, Liu X. Identification and Characterization of A Welwitindolinone Alkaloid Biosynthetic Gene Cluster in Stigonematalean Cyanobacterium *Hapalosiphon welwitschii*. Chembiochem : a European journal of chemical biology [Internet]. 2014 Mar;15(5):665–9. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4382313/>
166. Li S, Lowell AN, Yu F, Raveh A, Newmister SA, Bair N, et al. Hapalindole/Ambiguine Biogenesis Is Mediated by a Cope Rearrangement, C–C Bond-Forming Cascade. Journal of the American Chemical Society [Internet]. 2015 Dec;137(49):15366–9. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4681624/>
167. Li S, Lowell AN, Newmister SA, Yu F, Williams RM, Sherman DH. Decoding cyclase-dependent assembly of hapalindole and fischerindole alkaloids. Nature chemical biology [Internet]. 2017 May;13(5):467–9. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5391265/>
168. McClure RA, Goering AW, Ju K-S, Baccile JA, Schroeder FC, Metcalf WW, et al. Elucidating the Rimosamide-Detoxin Natural Product Families and Their Biosynthesis Using Metabolite/Gene Cluster Correlations. ACS Chemical Biology [Internet]. 2016 Dec;11(12):3452–60. Available from: <http://dx.doi.org/10.1021/acscchembio.6b00779>
169. Pasek S, Risler J-L, Brézellec P. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics [Internet]. 2006 Jun;22(12):1418–23. Available from: <https://academic.oup.com/bioinformatics/article/22/12/1418/207642>
170. Fani R, Brilli M, Liò P. The Origin and Evolution of Operons: The Piecewise Building of the Proteobacterial Histidine Operon. Journal of Molecular Evolution

- [Internet]. 2005 Mar;60(3):378–90. Available from: <https://doi.org/10.1007/s00239-004-0198-1>
171. Fani R, Liò P, Chiarelli I, Bazzicalupo M. The evolution of the histidine biosynthetic genes in prokaryotes: A common ancestor for the hisA and hisF genes. *Journal of Molecular Evolution [Internet]*. 1994 May;38(5):489–95. Available from: <https://doi.org/10.1007/BF00178849>
172. Fani R, Tamburini E, Mori E, Lazcano A, Liò P, Barberio C, et al. Paralogous histidine biosynthetic genes: Evolutionary analysis of the *Saccharomyces cerevisiae* HIS6 and HIS7 genes. *Gene [Internet]*. 1997 Sep;197(1):9–17. Available from: <http://www.sciencedirect.com/science/article/pii/S0378111997001467>
173. Merkl R, Sterner R. Reconstruction of ancestral enzymes. *Perspectives in Science [Internet]*. 2016 Dec;9:17–23. Available from: <http://www.sciencedirect.com/science/article/pii/S2213020916302336>
174. Gao Y, Xu G, Wu P, Liu J, Cai Y-s, Deng Z, et al. Biosynthesis of 2*I*-Chloropentostatin and 2*I*-Amino-2*I*-Deoxyadenosine Highlights a Single Gene Cluster Responsible for Two Independent Pathways in *Actinomadura* sp. Strain ATCC 39365. *Applied and Environmental Microbiology [Internet]*. 2017 May;83(10). Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5411499/>
175. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D Structure Computed from Evolutionary Sequence Variation. *PLOS ONE [Internet]*. 2011 Dec;6(12):e28766. Available from: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0028766>
176. Kufareva I, Abagyan R. Methods of protein structure comparison. *Methods in molecular biology (Clifton, NJ) [Internet]*. 2012;857:231–57. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321859/>
177. Hopf TA, Green AG, Schubert B, Mersmann S, Schärfe CPI, Ingraham JB, et al. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics [Internet]*. 2019 May;35(9):1582–4. Available from: <https://academic.oup.com/bioinformatics/article/35/9/1582/5124274>
178. Kuper J, Doenges C, Wilmanns M. Two fold repeated  $(\beta\alpha)_4$  half barrels may provide a molecular tool for dual substrate specificity. *EMBO reports [Internet]*. 2005 Feb;6(2):134–9. Available from: <https://www.embopress.org/doi/abs/10.1038/sj.embor.7400330>
179. Due AV, Kuper J, Geerlof A, Kries JP von, Wilmanns M. Bisubstrate specificity in histidine/tryptophan biosynthesis isomerase from *Mycobacterium tuberculosis* by active site metamorphosis. *Proceedings of the National Academy of Sciences of the United States of America [Internet]*. 2011 Mar [cited 2017 Jan 31];108(9):3554–9. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3048130/>
180. Chertkov O, Sikorski J, Nolan M, Lapidus A, Lucas S, Del Rio TG, et al. Complete

- genome sequence of *Thermomonospora curvata* type strain (B9T). Standards in Genomic Sciences [Internet]. 2011 Feb [cited 2017 Mar 3];4(1):13–22. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3072092/>
181. Pukall R, Gehrich-Schröter G, Lapidus A, Nolan M, Glavina Del Rio T, Lucas S, et al. Complete genome sequence of *Jonesia denitrificans* type strain (Prevot 55134T). Standards in Genomic Sciences [Internet]. 2009 Nov [cited 2017 Mar 3];1(3):262–9. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3035236/>
182. Angel E. Interactive computer graphics : A top-down approach with opengl. Boston, MA: Addison Wesley Longman; 2000.
183. Angel E. Batch-file computer graphics : A bottom-up approach with quicktime. Boston, MA: Wesley Addison Longman; 2001.
184. Angel E. Test second book by angel. Boston, MA: Wesley Addison Longman; 2001.