EvoMining

Introduction

La promiscuidad enzimática puede buscarse en familias envueltas en procesos de divergencia funcional. Uno de dichos procesos es la expansión y posterior reclutamiento de familias pertenecientes a rutas metabólicas conservadas hacia el metabolismo especializado. Los productos naturales o metabolitos especializados son sintetizados generalmente por clusters de genes distribuidos en un pequeño porcentaje de un linaje taxonómico. Estos clusters, conocidos como BGCs (Biosynthetical Gene Clusters), contienen reclutamientos de las copias extras de familias pertenecientes al metabolismo conservado. La similitud de secuencia de los genes que pertenecen a los BGCs así como su relativa sintenia en diversos organismos de un linaje hacen que genómica copmarativa sea de utilidad para intentar localizarlos. Finalmente, el auge en la cantidad de genomas disponibles públicamente así como la facilidad por sequenciar nuevos hace posible que los métodos bioinformáticos ayuden a encontrar nuevos BGCs. EvoMining es un método de para sugerir la formación de nuevos BGCs y en consecuencia encontrar zonas donde puede estar ocurriendo la capcidad de adquirir un nuevo sustrato cambio en promiscuidad enzimátcica.

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 enzimátocas provenientes del emtabolismo conservado. Se discutirán también cuatro linajes genómicos Actinobacteria Cyanobacteria, Pseudomonas y Archaea y finalmente se analizará un BGCs scytonemin.

EvoMining es un método para encontrar BGCs no tradicionales

Consideraciones evolutivas en minería genómica permiten identificar clusters no clasificados y dentro de ellos sugerir enzimas con cambios en promiscuidad.

En antiSMASH 2.0 existen estos clusters y ____ son no conocidos

Los clusters biosintéticos no clasificados son difíciles de identificar.

Los no clasificados suelen pasar desapercibidos, porque no hay conocimiento previo

Localizar alguna enzima de un cluster no clasificado es un camino para identificar el BGC.

Como ya se mostró en EvoMining 1.0 encontramos una enzima que

Copias extra de familias enzimáticas que son reclutadas para una nueva función están relacionadas con la promiscuidad.

Introducción de como funciona

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

Algoritmos y bases de datos de EvoMining 2.0

EvoMining está compuesto de dos algoritmos, el de búsqueda de patrones de expansión - reclutamiento y el de visualización de todas las copias de una familia expandida clasificadas según sus posibles destinos metabólicos.

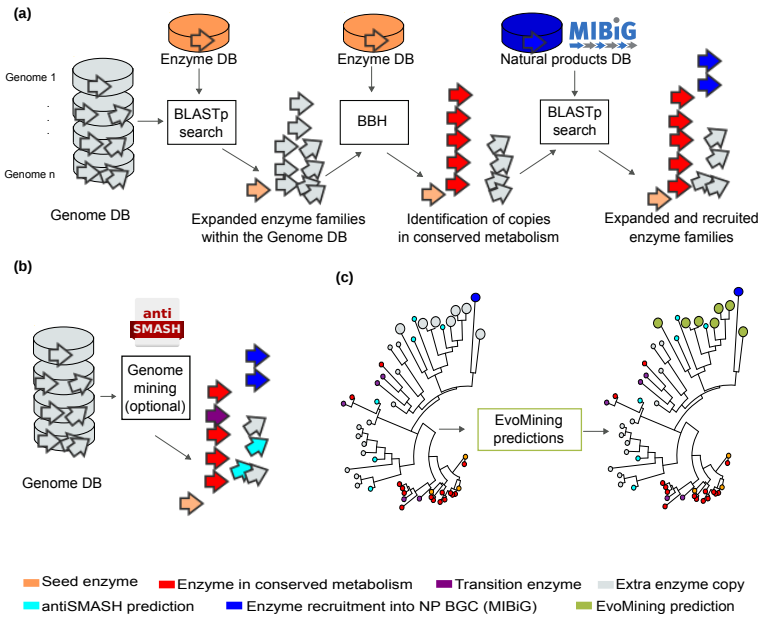

Figure 1:

Además de los algortimos EvoMining necesita tres bases de datos: i) la genómica, ii) las semillas enzimáticas, y iii) la de productos naturales.

Algoritmo de expansión y reclutamiento

La primera parte de la minería genómica evolutiva de EvoMining consiste en detectar las expansiones y los reclutamientos de las familias enzimáticas semilla. Una familia expandida 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 de un linaje taxonómico (Genome DB). Un reclutamiento es una copia extra de una familia de metabolismo conservado que ahora participa en metabolismo especializado. Ejemplos de reclutamientos conocidos han sido observados en BGCs reportados en MIBiG. Ejemplos de reclutamientos predichos se encuentran en enzimas pertenecientes a BGCs reconocidos por antiSMASH que tienen un origen en metabolismo conservado.

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.

En este sentido, no todas ls copias son expansiones, pero aún así pueden ser predicciones de EvoMining, por ello para futuros trabajos es posible que estar una unidad arriba de la moda, sea suficiente para ser un reclutamiento.

Los patrones de reclutamiento sólo pueden ser apreciados tras ordenar los organismos filogenéticamente, por ejemplo *Streptomyces* tiene dos enolasas. Se recomienda al usuario reordenar los organismos del heatplot de acuerdo a un árbol de especies, mismo que puede ser calculado utilizando Orthocore.

En este paso, en cada familia expandida los ortólogos más parecidos a las enzimas semilla de la enzyme DB son identificados por BBH. Estos ortólogos serán considerados parte del metabolismo conservado y serán posteriormente identificados con color rojo en la visualización. Las copias extra parte de la familia expandida

son enzimas de las que no se conoce el destino metabólico, es posible que en los pasos subsequentes de EvoMining sean reconocidas por antiSMASH o EvoMining como posibles reclutamientos a metabolismo especializado. En otro caso permanecerán como copias extra con destino metabólico desconocido.

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 BGCs. Otras bases de datos que contengan enzimas con un destino metabólico conocido podrían utilizarse en lugar de NP DB. Los reclutamientos son la

En conclusión el algoritmo de expansión - reclutamiento trabaja para identificar tres clases de copias en las familias enzimáticas expandidas (i) copias altamente conservadas con algún miembro en el metabolismo conservado; (ii) reclutamientos conocidos en BGCs de productos naturales; y (iii) copias extra que no son reclutamientos conocidos ni parte obvia del metabolismo conservado, quedando por definir su destino metabólico. Reconstrucciones filogenéticas de las familias enzimáticas serán utilizadas en la siguiente sección para asignar a estas copias extra obtenidas por el algoritmo de expansión y reclutamiento un origen y destino metabólico

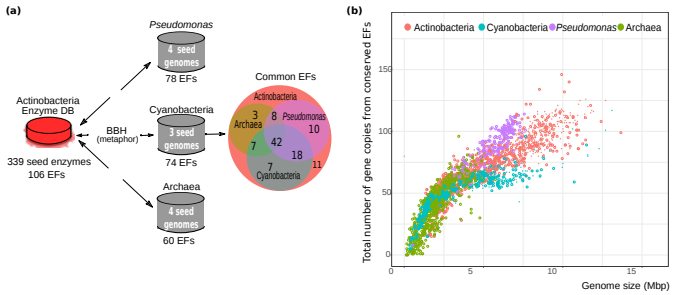

Figure 2:

Algoritmo de reconstrucción filogenética y visualización

Las secuencias de las familias enzimáticas expandidas fueron alineadas con muscle v3.2[@edgar_muscle_2004] y curadas con Gblocks v0.91b[@castresana_selection_2000]. Los parámetros de Gblocks fueron fijados en incluir 5 posiciones como el mínimo tamaño de un bloque y diez posiciones como el máximo de posiciones contiguas no conservadas. Posiciones ausentes de más del 50% de las secuencias fueron filtradas y removidas del alineameinto final. Para recontruis filogenéticamente la historia de las enzimas se utilizó FastTree 2.1[@price_fasttree_2010] que es un método muy rápido pensado para miles de secuencias que utiliza un algoritmo definido como “aproximadamente máxima verosimilitud”.

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[@weber_antismash3_2015] fue utilizado en los secuencias de los genomas de la Genome DB Figure 1 panel (b).

La selección de color de las hojas del árbol fue automatizada utilizando Newick Utilities una serie de programas llamados desde la terminal para manipular árboles en formato Newick[@junier_newick_2010]. La anotación funcional de RAST en su versión clásica[@aziz_rast_2008,@overbeek_seed_2014] fue agregada al SVG. Otros

metadatos como el número de copias por organismo y fueron escritos de manera que los árboles producidos por EvoMining sean compatible con la visualización de Microreact[@argimon_microreact_2016].

Los árboles de EvoMining diferencian entre la función metabólica de cada miembro de una familia génica mediante un código de color Figure 1 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[@medema_minimum_2015]. Estos reclutamientos son marcados en azul. Una vez definidos estos dos grupos, una predicción de EvoMining es definida como aquellas hojas más cerca de una hoja azul que de una hoja roja. Estas predicciones consideradas con más posibilidades de pertenecer al metabolismo especializado que al conservado, son coloreadas en verde Figure 1 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 cyan y son llamados predicciones antiSMASH. Si una secuencia es al mismo tiempo predicción EvoMining y predicción antiSMASH se colorea cyan 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 cyan entonces es coloreada de púrpura. Estas enzimas de intersección entre metabolismo conservado y metabolismo especializado son definidas como enzimas de intersección. Finalmente para todas las copias extra que no fueron marcadas como rojas, azules, verdes, cyanes o púrpuras existe el color girs. 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.

Las tres bases de datos de EvoMining

EvoMining DBs 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 de Genome DB comprendió 230 genomas de Actinobacteria, incluyendo 50 géneros diferentes. Gracias a la explosión de datos genómicos disponibles, en EvoMining 2.0 fue posible expandir el número de genomas de Actinobacteria de Genome DB hasta 1245 genomes, incluyendo 193 géneros diferentes. Cuatro diferentes bases Genome DB fueron utilizadas en este trabajo, Actinobacteria, Cyanobacteria, Pseudomonas y Archaea.

Estas bases están disponibles en el repositorio de datos público Zenodo con identificador DOI 10.5281/zenodo.1219709

.

Así pues, adicionalmente a 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 Cyanobacteria (416 genomas), Pseudomonas (219 genomas) y Archaea (876 genomas). Estos taxa fueron elegidos por su diversidad de exploración respecto a BGCs, por ejemplo Actinobacteria posee 602 MIBiG BGCs, Cyanobacteria cuenta con 60 MIBiG BGCs y Pseudomonas 53 MIBiG BGCs. 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 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 BGCs 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[@overbeek_seed_2014] que a la vez realiza anotaciones funcionales . Estos genomas, previo al análisis de EvoMining fueron minados por antiSMASH[@weber_antismash3_2015] 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 [@cruz-morales_phylogenomic_2016]. 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 Cyanobacteria , Pseudomonas and Archaea. Durante el proceso de selección, para evitar excluir familias debido a huecos debidos a problemas técnicos relacionados a la secuenciación o al ensamble de los genomas, se escogieron genomas semilla que estuvieran contenidos en un sólo contig. Los genomas semillas son los proveedores de las secuencias semilla que conforman la base de datos Enzyme DB. Para Cyanobacteria, los genomas seleccionados son Cyanothece 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 [@cruz-morales_phylogenomic_2016]. La herramienta Metaphor [@van_der_veen_metaphor_2014] fue la técnica implementada para obtener los BBH, se fitraron 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, Cyanobacteria, 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

NP DB Los primeros análisis realizados con EvoMining incluían un base de datos de productos naturales NP DB de 226 BGCs reunidos de la literatura y curados manualmente [@cruz-morales_phylogenomic_2016]. En este trabajo la base NP DB que se utilizó para los análisis es MIBiG v1.3[@medema_minimum_2015]. 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 BGCs y un total de $31,023$ secuencias de proteínas.

Observaciones genómicas de EvoMining

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 Cyanobacteria , 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 BGCs), Cyanobacteria (60 MIBiG BGCs) y Pseudomonas (53 MIBiG BGCs); como también Archaea (0 BGCs en MIBiG versión 1.3), que representa un dominio poco explorado en lo que respecta a los genes que forman parte de metabolismo especializado [@charlesworth_untapped_2015].

Basándonos en estas bases de datos del tipo Genome DBs, como es explicado en la Figure 2 panel (a), un conjunto de familias enzimáticas comunes fue identificado. Notablemente, de las originales 106 familias actinobacteriales menos del 50% estaba conservada en los nuevos taxa incorporados. Cada base de datos, una para cada taxon, contiene sólo 42 FEs (Table S1). Esta observación, que 64 FEs no están conservadas en los cuatro taxa refleja lo específico del metabolismo en cada linaje [@jordan_lineage-specific_2001]

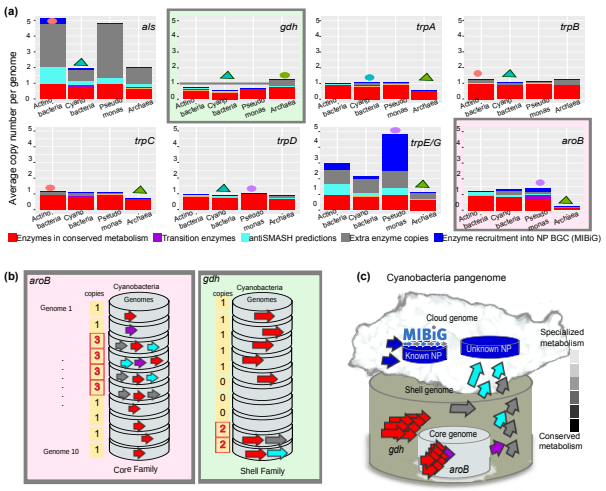

Figure 3: EvoMining perfiles de enzimas conservadas seleccionadas. (a) Patrones de expansión de las ocho familias conservadas cuyas copias adicionales participan en la biosíntesis de scytonemin. El conjunto completo de 42 EF se muestra en la Fig. S11. La codificación de colores es la siguiente: rojo 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.

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 es más grande que el phylum Cyanobacteria y que el dominio Archaea Figure 2 panel(b). Este resultado puede deberse a que no se han descubierto tamaños de genoma Archaea de tamaño comparable a los de Streptomyces o Pseudomonas (> 5Mbp). Cyanobacteria a pesar de tener genomas grandes no tiene tantas expansiones. Esta observación, es posible que sea generalizable a todas las EFs de metabolismo conservado o bien que se deba a un sesgo en la selección de las familias que componen a la EFs.

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 Cyanobacteria 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 NP BGCs sintetizadores de estos metabolitos no han sido caracterizados [@charlesworth_untapped_natural_products_Archaea_2015]. Por ello EvoMining una herramienta de minería genómica que puede ayudar a explorar linajes poco minados con el potencial de descubrir nuevas rutas metabólicas.

La figura muestra que aunque 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. Este incremento parece cambiar su patrón en todos los linajes a partir de 5Mbp Figure 2 (2b). Acorde a estos resultados 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.

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. (Fig. S11). En contraste, Actinobacteria es el máximo en 26.2% de las FEs, mientras que Archaea and Cyanobacteria empatan en ser el linaje con más expansiones sólo en el 9.5% de los casos (Table S1). Aunque existen familias como la acetylornitino aminotransferasa o la acetolactato sintasa (ALS) que están expandidas en todos los linajes (coordenadas A1 y E1, Fig. 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 Cyanobacteria .

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 XX. Es decir, 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 3
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. 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 3 panel (b). Los ocho casos mostrados en la figura son todos parte de un cluster biosintético de Cyanobacteria, descrito en las secciones posteriores de este trabajo.

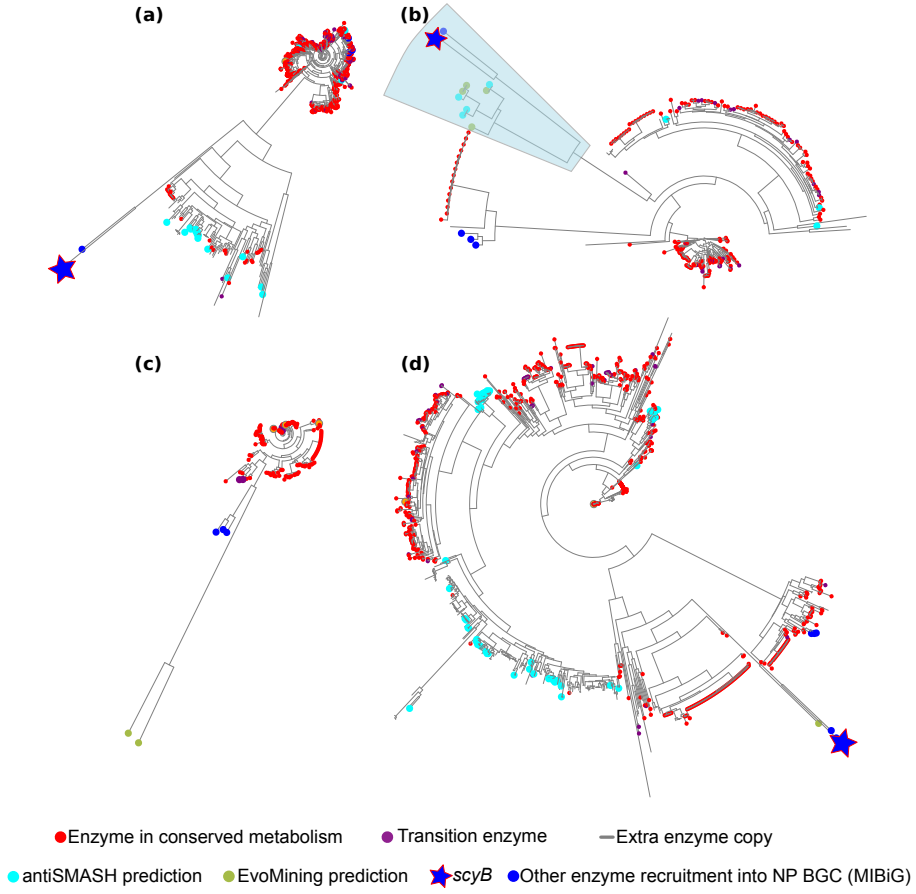
(a)
(b)
(c)
(d)
Enzyme in conserved metabolism
Transition enzyme
Extra enzyme copy
antiSMASH prediction
EvoMining prediction
scyB
Other enzyme recruitment into NP BGC (MIBiG)

Figure 4: n

Con estas observaciones sospechamos que GDH es miembro del shell genome [@koonin_genomics_2008] en los taxa Actinobacteria , Cyanobacteria 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 n un taxon como en el caso de un domino o un phylum en contraposición con taxones coservados 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 3 panel (a). Las modas de número de copias también son informativas por ello son mostradas en la figura XXX. Una sóla copia extra puede ser la que sea reclutada en metabolismo especializado.

En la figura XXX 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 cyan, 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 cyan. Figure 3 panel (b) .

Las expansiones encontradas por EvoMining en la familia GDH incluían predicciones de antiSMASH para Actinobacteria, Cyanobacteria y Archaea, no así para Pseudomonas. La secuencia del reclutamiento de GDH por los clusters biosintéticos scytonemin y el policétido pactamycin [@kudo_cloning_2007], es suficientemente parecida como para que EvoMining la detecte como expansión en Actinobacteria, Cyanobacteria 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 4 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 BGCs. A partir de estos resultados Figure 3 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 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 3 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 BGCs novedosos[@navarro-munoz_computational_2018].

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 4 , 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, Cyanobacteria 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 scytonemin[@balskus_investigating_2008,@soule_comparative_2009] un cluster biosintético de Cyanobacteria

GDH y ALS en el cluster scytonemin 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 α-cetoglutarato y amonio. De acuerdo al uso de cofactores GDH puede dividirse en tres

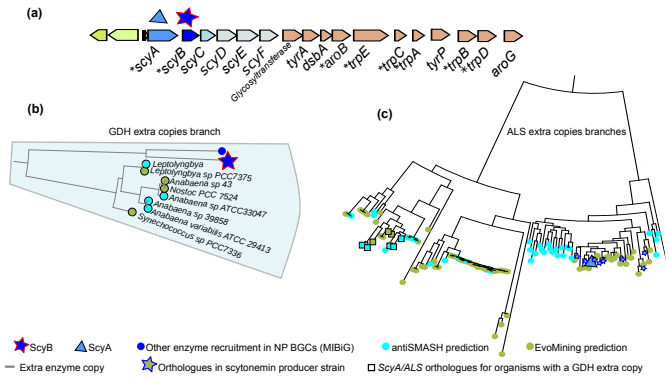

Figure 5:

clases, la primera usa NAD+ y es nombrada como GDH(NAD+). La segunda clase utiliza NADP+ y es conocida como GDH(NADP+). La tercera clase utiliza ambos cofactores NAD+ y NADP+; por lo que se le conoce como GDH(NAD+ y NADP+)[@engel_glutamate_2014].

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+) is utilizada para la oxidación del glutamato mientras que GDH(NADP+) para fijar amonio, algunas enzimas de Archaea funcionan bien con ambos cofactores, es decir tienen promiscuidad de cofactores[@engel_glutamate_2014]. La especificidad por NAD o NADP probablemente emergió en repetidas ocasiones, una evidencia a favor de esta hipótesis es que se ha mostrado que algunas mutaciones pueden revertir la especificidad[@lilley_partial_1991]. 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 4 panel (a) ,y en Cyanobacteria Figure 4 panel(b). Las expansiones están prácticamente ausentes en *Pseudomonas* Figure 4 panel (c). En contraste, un número significativo de expansiones es encontrado en Archaea Figure 4 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)+[@consalvi_glutamate_1991]. 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)[@ferrer_nadp-glutamate_1996], incluyendo muchas predicciones de antiSMASH, y sólo dos marcadas como metabolsimo 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* , *Natrialbaceae* y *Natronococcus*. Los genes codificante se encuentran un contexto de posible síntesis de terpenos. Este contexto incluye enzimas relacionadas al geranyl pyrofosfato, un precursor de todos los terpenos y terpenoides[@tholl_terpene_2006]. 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 4 panel(d).

En contraste con la amplia expansión de GDH relacionada a las adaptaciones metabólicas en Archaea, el árbol de Cyanobacteria tiene copias extra sólo en el 4.5% de sus genomas (Figure 4 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 scytonemin. ScyB es pues parte de la síntesis de scytonemin, un pigmento amarillo producido por muchas Cyanobacterias como protección contra la radiación UV-A solar[@balskus_genetic_2010]. *Nostoc punctiforme* PCC 73102 es el organismo productor de scytonemin 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 scytonemin se observó que *scyB* se conserva cerca del gene *scyA* Figure 5 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 Cyanobacteria. 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 Cyanobacteria , 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 5 (c). Estas predicciones incluyen más de veinte organismos conocidos como productores de scytonemin [@balskus_investigating_2008]

Además, ramas cercanas muestran secuencias de ALS queson p redicciones de antiSMASH, reforzando la sugerencia de que esta seción 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 5 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 expansin 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 GDH EF y homólogos de ScyB no fue suficiente para reconstruir una rama de *scyB* con todas las expansiones sugeridas por la ocurrencia del cluster de scytonemin. 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.

El cluster de scytonemin es un cluster promiscuo.

El primer cluster caracterizado de scytonemin, mostrado en la figura Figure 5 panel (a) y en Figure 6 , comprende 18 genes [@soule_comparative_2009]. Además de genes reguladores, este BGC incluye genes biosintéticos scyABC; otros genes conservados cuya función queda por dilucidar, scyDEF; y 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 el ácido shikímico, parecen haber sido reclutadas para proveer de los precursores L-triptofano y prefrenato necesarios para la síntesis de scytonemin. 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 al metabolismo especializado Figure 6. ALS une dos piruvatos, transformándolos en S-2-acetolactato [@liu_acetohydroxyacid_2016], mientras que ScyB cataliza la unión de indol-3-piruvato con p-hydroxy-fenil-ácido pirúvico. Análogamente GDH convierte L-glutamate en 2-oxoglutarato [@engel_glutamate_2014], mientras que ScyA cataliza una desaminación oxidativa de triptofano. El producto de estas dos enzimas

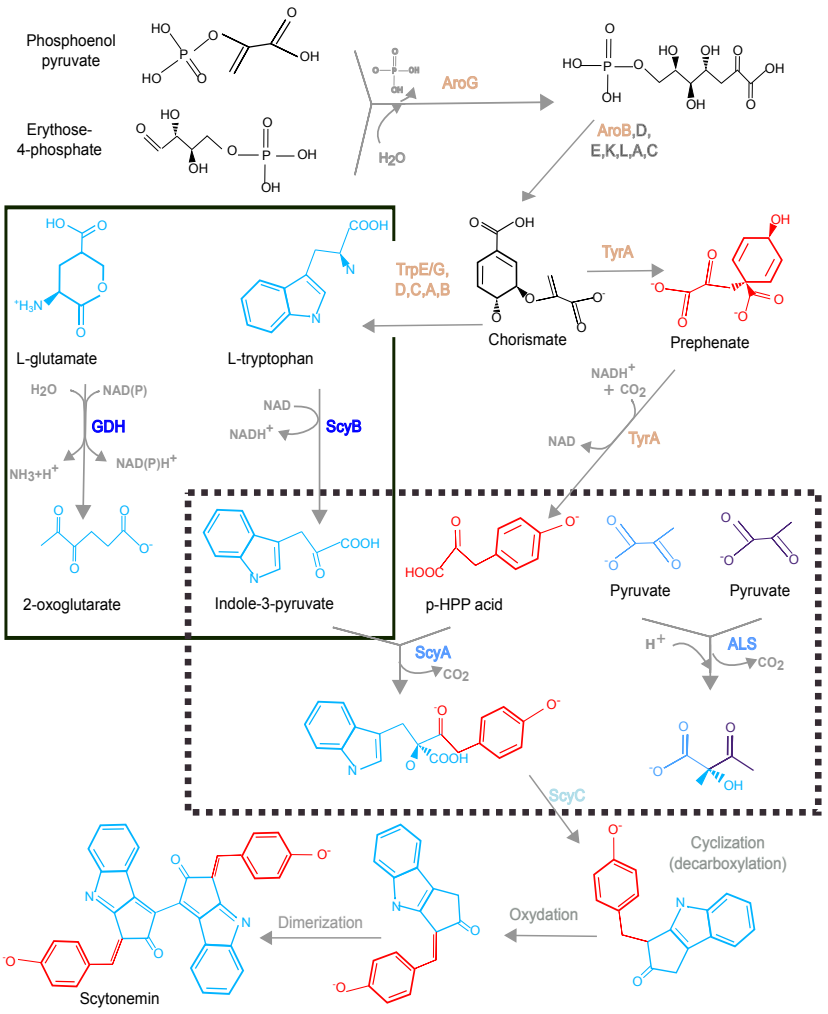
Phosphoenol pyruvate
Erythose-4-phosphate
AroG
H2O
AroB,D, E,K,L,A,C
L-glutamate
L-tryptophan
TrpE/G, D,C,A,B
Chorismate
TyrA
Prephenate
H2O
NAD(P)
GDH
NH3+H+
NAD(P)H+
NAD
NADH+
ScyB
NADH+ + CO2
NAD
TyrA
2-oxoglutarate
Indole-3-pyruvate
p-HPP acid
Pyruvate
Pyruvate
ScyA
CO2
H+
ALS
CO2
ScyC
Cyclization (decarboxylation)
Oxydation
Dimerization
Scytonemin

Figure 6:

actuando secuencialmente e un dipéptido, el cual es ciclizado por ScyC. La ruta metabólica culmina con una serie de oxidaciones y dimerizaciones hasta llegar al producto scytonemin, aún no es claro como estos útimos pasos se llevan a cabo [@balskus_investigating_2008].

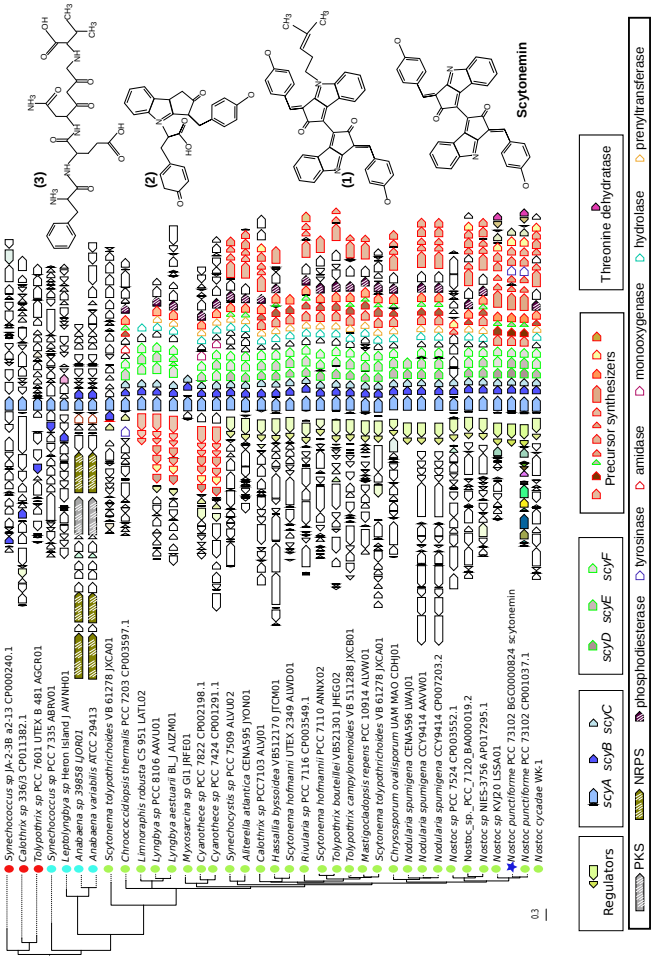

Figure 7:

Además de GDH y ALS el BGC de scytonemin tiene seis familias que son parte de las 42 FEs analizadas aquí, es decir 8 de los genes que participan en la síntesis de scytonemin tienen un origen en metabolismo central y fueron reclutados en el cluster de scytonemin (Figure 3 panel 1, Table S3). (Figure 5 and S3–S9). De estas familias, seis de los siete árboles de EvoMining contienen copias extras identificadas como predicciones de EvoMining predictions, debido a que en su rama de expansión se encuentra el correspondiente gen de scytonemin 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 con enzimas provenientes de las rutas de síntesis de otros pigmentos protectores solares como la shinorina y los aminoácidos de tipo mycosporina Figure 8 [@balskus_genetic_2010] , además de otros productos no relacionados como la welwitindolinona [@hillwig_identification_2014], ambiguina [@li_hapalindole_ambiguine_2015] y la fischerindolina [@li_decoding_2017] (Figs S3–S7). Estos resultados ilustran como EvoMining puede complementar a antiSMASH mediante la identificación de secuencias de NP BGCs no tradicionales.

Para investigar la coocurrencia de ScyA y ScyB, que son necesarias juntas para producir scytonemin, reconstruimos su historia evolutiva conjunta. Para ello se obtuvieron las secuencias de lo genomas donde

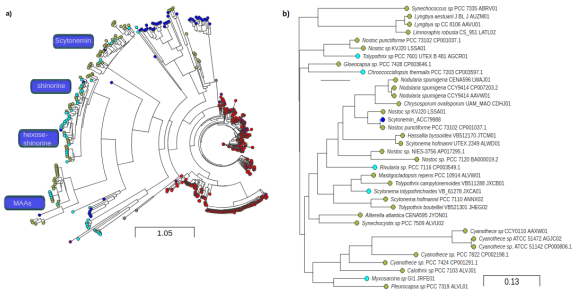

Figure 8:

ambas tenían presencia, se concatenaron sus secuencias y se realizó una filogenia con ellas. Las variantes de la vecindad genómica del BGC de scytonemin fueron visualizadas mediante el uso de corason [@navarro-munoz_computational_2018]. En el siguiente capítulo este software de visualización y organización de vecindades genómicas será explicado con más detalle. Los análisis filogenómicos resultaron en 34 cyanobacterias con diversidad química en el BGC de scytonemin. Es decir en concusión scytonemin 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 scytonemin y correlacionan con episodios de pérdida y ganancia de genes en este locus Figure 7. Al core de genes scyABCD se incorporan genes que realizan ornamentos como hidrolasas, prenil-transferasas , fosfodiesterasas y monooxigenasas , para formar congéneres de scytonemin como los compuestos 1 y 2. La pérdida de los genes *scyDEF* y la aparición de otras enzimas como la tyrosinase 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. Las variantes producidas por la dinámica evolutiva del metabolismo especializado fueron sugeridas mediante el sólo uso de ScyA y ScyB como anzuelos de búsqueda. Estos resultados sugieren el poder predictivo de EvoMining para abrir la exploración de espacios metabólicos típicamente inexplorados por métodos de búsqueda tradicionales de NP BGCs que no consideran la evolución dentro de sus algoritmos de minería genómica.

Consideraciones para el uso de EvoMining

EvoMining fue desarrollado como una herramienta descargable de minería genómica 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 expeciones donde EvoMining pudo predecir enzimas de BGCs 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.

Después de analizar GDH de manera exhaustiva, un EF expandido notablemente en Archaea pero no en otros taxones, proporcionamos un ejemplo de un reclutamiento de una enzima metabólica central por un NP BGC, así como por otras vías metabólicas. Es interesante observar que los EF más expandidos en los análisis de prueba de concepto anteriores de EvoMining [@cruz-morales_phylogenomic_2016] fueron asparagina

sintasa, 2-dehidro-3-deoxifosfoheptanoato aldolasa y 3-fosfosquimato-1-carboxivinil transferasa, que llevan a El descubrimiento de enzimas biosintéticas de arsenolípidos sin precedentes. Cabe destacar que ninguna de estas enzimas formaba parte de los 42 EF analizados en el presente documento, lo que refuerza la idea de que no solo las enzimas conservadas, sino también las enzimas de cáscara con copias adicionales, pueden servir como balizas para el descubrimiento de nuevos BGC de NP. Estas observaciones enfatizan la naturaleza predictiva de EvoMining, que se hizo evidente solo después de que el origen y el destino de las enzimas pudieran rastrearse hasta eventos evolutivos en diferentes niveles, desde la dinámica del genoma que involucra loci grandes, hasta diferentes tasas de mutaciones en el nivel de secuencia proteica.

Los usuarios de EvoMining, por lo tanto, deben definir de antemano el EF más apropiado para un determinado grupo taxonómico. El Enzyme DB seleccionado debe contener un conjunto de EF donde se puedan detectar los patrones de expansión. A su vez, los EF con una distribución restringida a un pequeño porcentaje de genomas no son adecuados para el análisis de EvoMining. También es importante determinar qué EF están compartidos 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 EF curados manualmente que solo incluían enzimas metabólicas centrales, pero, como se demostró aquí, no representaban necesariamente 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 de la cubierta. Esta noción implica la posibilidad de automatizar la integración de Enzyme DB mediante la selección de EF en cualquier linaje genómico dado, evitando la necesidad de definir arbitrariamente qué es el metabolismo central.

Otro punto clave para el mejoramiento de EvoMining fue la disponibilidad de la base de datos MIBiG [@medema_minimum_2015] que permite incrementar consistentemente y sin esfuerzo de curación manual a los BGCS reportados por investigadores de todo el mundo. La versión previa de EvoMining no incluía por ejemplo NP BGC de Cyanobacteria o de Archaea, y fue por esta actualización que la nueva versión de EvoMining pudo identificar correctamente secuencias cercanas a ScyA y ScyB. Sin la presencia de la señal de los BGC de Cyanobacteria en MIBiG estas hojas habrían sido catalogadas como de destino metabólico desconocido.

This maybe the case in Archaea, where some sequences in the GDH tree are labelled as such, possibly related to terpenes as well as to other metabolic fates yet-to-be discovered. The presence of only one Archaea BGC at MIBiG is clearly due to the limited research available of the potential of Archaea to synthesize NPs, as our results suggest that current methods based on previous knowledge from unrelated taxa impose biases that hamper our ability to unlock the metabolic diversity of this domain of life. We anticipate that this situation will be overcome by EvoMining, as it is a less-biased and rule-independent approach.

OTros ejemplos de EvoMining en Actinobateria y Pseudomonas

EvoMining en Rna transferasas

Azoxy

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

TauD en una enzima dedicada al metabolismo de taurina, proviene del operon de *E. coli*. En Pseudomonas también tiene muchas expansiones. Es parte de los clusters rimosamide y detoxin, así como de 15 BGCs 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

Rimosamide y Detoxin tienen en común la enzima TauD,

Un análisis de EvoMining8 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). Dentro de las expansiones de la familia existe un clado,que contiene quince homólogos de *tauD* que pertenecen a clusters biosintéticos experimentalmente caracterizados y depositados en MIBiG, incluyendo los BGCs de detoxin y rimosamides (Table S6). La variedad de BGCs mostrada en este clado abre la posibilidad de encontrar variantes moleculares de estas familias.

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

Table 1: Occurrencias de homólogos de *tauD* en BGCs de metabolismo secundario reportados en MIBiG.

MIBiG BGC	Compound	Class	Producer Organism
653_ADO85576	pentalenolactone	Terpene	*Streptomyces arenae*
678_BAC70706	pentalenolactone	Terpene	*Streptomyces avermitilis* NBRC 14893
163_ACR50790	tetronasin	Polyketide	*Streptomyces longisporoflavus*
961_ABC36162	bactobolin	NRP-Polyketide	*Burkholderia thailandensis* E264
287_AAG05698	2-amino-4-methoxy-trans-3- butenoic acid	NRP	*Pseudomonas aeruginosa* PAO1
846_ctg1_orf9	tabtoxin	Other	*Pseudomonas syringae*
1183_AGC09526	lobophorin	Polyketide	*Streptomyces* sp. FXJ7.023
1156_ADD83004	platencin	Terpene	*Streptomyces platensis*
1140_ACO31277	platensimycin-platencin	Terpene	*Streptomyces platensis*
1140_ACO31282	platensimycin-platencin	Terpene	*Streptomyces platensis*
715_ABW87795	spectinomycin	Saccharide	*Streptomyces spectabilis*
1205_KGO40485	communesin	Polyketide	*Penicillium expansum*
1205_KGO40482	communesin	Polyketide	*Penicillium expansum*
1183_AGC09525	lobophorin	Polyketide	*Streptomyces* sp. FXJ7.023
654_ABB69741	phenalinolactone	Saccharide-Terpene	*Streptomyces* sp. Tu6071
1070_CAN89617	kirromycin	NRP - Polyketide	*Streptomyces collinus* Tu 365

La tabla Table 1 muestra los reclutamientos de *tauD*.