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

Introducción

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 probabliemente ocurrió a traves de promiscuidad enzimática. Los homólogos con la función 'secundaria' suelen ser producto de expansiones previas de genes con la funcion 'primaria' REFERENCIA. En un trabajo previo se demostró para un caso en el que efectivamente las enzima XXX realiza la funcion primaria mientras que XXX cataliza la funcion 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 ezimas 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.

EvoMining es un paradigma que permite ubicar copias extra de familias enzimáticas y organizarlas visualmente acorde a eventos evolutivos para encontrar BGC no tradicionales

Existen varias clases de BGC que son arquetipos de los productos naturales. Entre ellas se encuentran las clases *non ribosomomal 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 sintasas 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 v1.3 (la base de datos de BGC caracterizados experimentalmente) hay 231 BGC [@medema_minimum_2015] (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 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 sintenia 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 [@garciapichel_evidence_1992], un pigmento cianobacteriano que absorve luz UV. Su biosíntesis requiere de ScyB y ScyA, dos enzimas que sostienen la síntesis de este metabolito especializado [@balskus_investigating_2008, @soule_comparative_2009]. 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 [@engel_glutamate_2014] y en la síntesis de aminoácidos de cadena ramificada [@liu_acetohydroxyacid_2016], respectivamente.

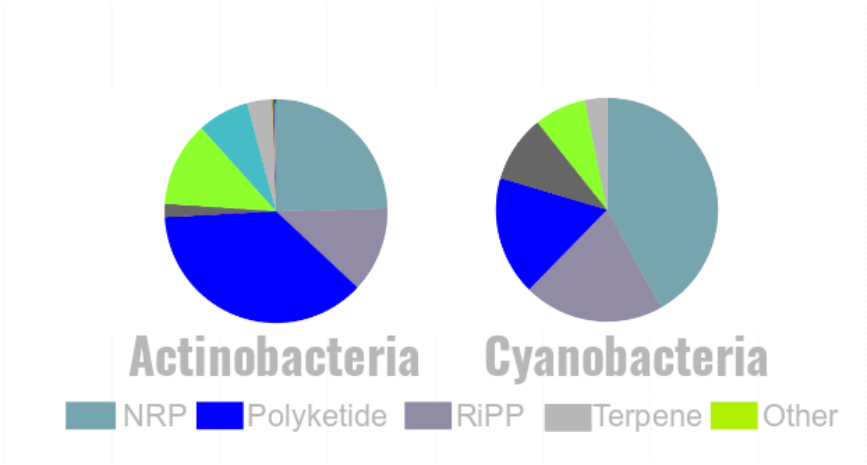

Figure 1: Existen cluster biosintéticos no clasificados (other) entre los reportados en MIBiG

Algoritmos y bases de datos de EvoMining 2.0

EvoMining está compuesto de dos algoritmos: el primero que utiliza a la familia semilla para encuentrar todos miembros de la familia entre todos los genomas blanco para así detectar genomas donde haya habido expansiones e identificar a todos los miembros de la Familia Expandida (FE) 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 algortimos EvoMining necesitan tres bases de datos: i) los genomas blanco, ii) las secuencias de ezimas semilla, y iii) la de productos naturales verificados.

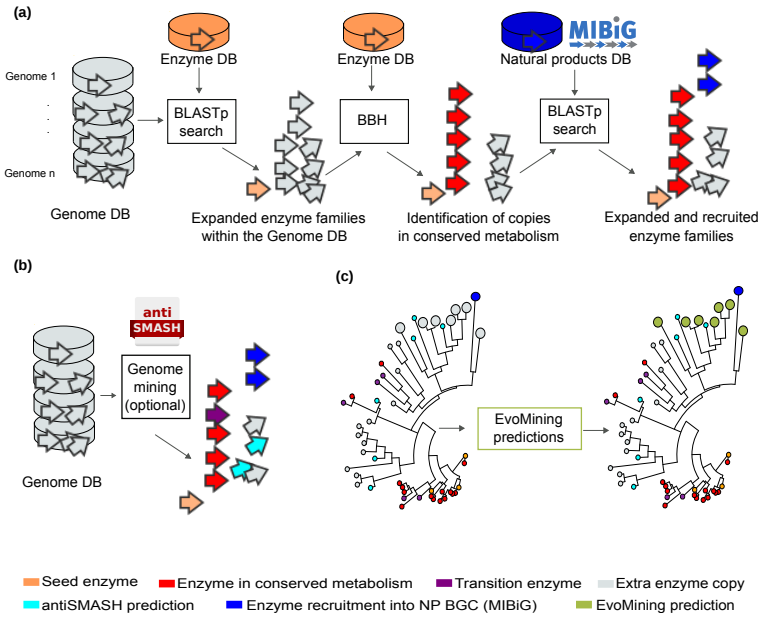

Figure 2: Representación de la tubería bioinformática de los dos algoritmos que componen EvoMining. a) Algoritmo de expansión - clasificación/reclutamiento. Se ingresan la Enzima DB (cilindro anaranjado) y la Genome DB (cilindros grises) para identificar por medio de BlastP a todos los que sean miembrios 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

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 ezimas 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:

a) Secuencia semillla (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.

b) 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.

c) 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). 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 (Figure 6 panel b).

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 subsequentes 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 seucencias 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[@weber_antismash3_2015] fue utilizado en los secuencias de los genomas de la Genome DB Figure 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.

Algoritmo de reconstrucción filogenética y visualización

Una vez que se tienen definidos los genes de una famliia expandida (FE) de enzimas y que ya se conoce cuáles de los homólogos realizan la funcion 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 todavia. Para eso se alínean 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 alínean con muscle v3.2 y son curadas con Gblocks. Las posiciones ausentes en más del 50% de las secuencias se filtran y remueven del alineameinto final. Para recontruir filogenéticamente la historia de las enzimas se utiliza 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". Así se obtiene un árbol en formato newick, mismo que puede ser utilizado con el software de visualización de Microreact[@argimon_microreact_2016] 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 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 sin una categoría definida en el primer algoritmo y qu 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 homologos con la función primaria. Estas predicciones consideradas con

más posibilidades de pertenecer al metabolismo especializado que al conservado, son coloreadas en verde Figure 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 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.

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[@argimon_microreact_2016].

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. 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[@overbeek_seed_2014] 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[@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 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 [@cruz-morales_phylogenomic_2016]. La herramienta Metaphor [@van_der_veen_metaphor_2014] fue 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, 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

-**NP DB** (Base de datos de genes biosintéticos de productos naturales). 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 [@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 BGC y un total de 31,023 secuencias de proteínas.

EvoMining detecta distintas dinámicas evolutivas de las enzimas metabólicas que dependen de la familia enzimática y del linaje taxonomíco.

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 [@charlesworth_untapped_2015].

Basándonos en estas bases de datos de genomas blanco (Genome DB), como es explicado en la Figure 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 [@jordan_lineage-specific_2001].

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 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.

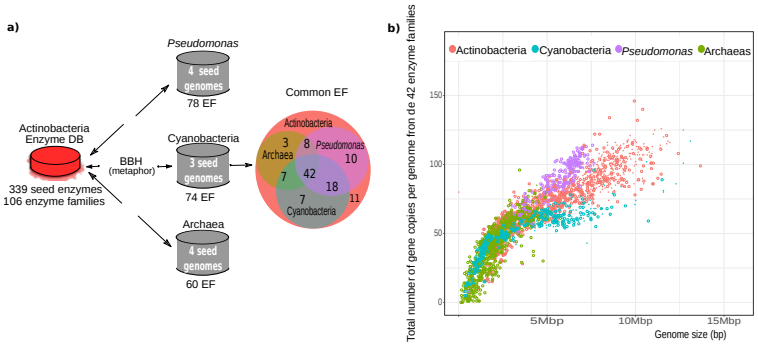

Figure 3: EvoMining Enzyme DB. (a) La base de datos de enzimas de la versión previa de EvoMining se filtró para establecer un conjunto común de 42 familias enzimáticas conservadas para los phyla Actinobacteria y Cyanobacteria, el género *Pseudomonas* y el dominio Archaea. (b) Todos los taxones muestran expansiones de familias enzimáticas que correlacionan con el tamaño del genoma. Las diferencias en las tasas de expansión entre los taxones se observan principalmente después de un tamaño de genoma superior a 5 Mbp. En este umbral, *Pseudomonas* supera las expansiones de Actinobacteria, que a su vez supera a Cyanobacteria en las 42 familias seleccionadas.

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 Figure 4.

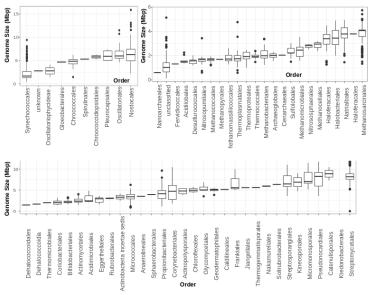

Figure 4: Tamaño de genoma en órdenes de Actinobacteria y Cianobacteria y en clases De Archaea. Actinobacteria y Cianobacteria tienen algunos genomas con tamaños superiores a 6 Mbp, que es el máximo que se encuentra en Archaea. El tamaño promedio en el género *Pseudomonas* es de 5.8 MBbp con un máximo de 7.6Mbp entre los genomas utilizados en este trabajo. (no se muestra en la figura)

Esta observación es congruente con que las archaeocinas, dicetopiperazinas, carotenoides y otros productos nat-

urales de Archaea fueron aislados de especies de Halobacteria, los genes, probablemente en BGC, que sintetizan 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.

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 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.

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 (TABLATABLATABLA 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). En esta figura, para cada familia en los ejes horizontales siempre se muestran en orden cuatro barras: Actinobacteria, Cianobacteria, *Pseudomonas* y Archaea. El código de color es el mismo que el de los árboles de EvoMining, con excepción del verde, pues aun no hay árbol filogenético. Asi pues es como sigue: rojo para el metabolismo conservado, azul para los reclutamientos anotados en MIBiG, cian para predicciones de antiSMASH de pertenencia a un BGC de metabolismo especializado, púrpura para la intersección entre el metabolismo conservado y predicciones antiSMASH y gris para expansiones sin destino metabólico conocido. La letra en la parte inferior y los números a la izquierda son coordenadas para facilitar la identificación de la familia en la Tabla S1. Los triángulos indican el linaje con el mayor número de copias por genoma en promedio, y los círculos representan la menor cantidad de copias. Aunque Archaea tiende a ser los taxones menos expandidos, esta tendencia revierte en las familias A4, C4, G4 (GDH) y B5. GDH y ALS (E1) están encerradas en una cuadro, estas enzimas son el origen de los reclutamientos en el BGC de escytonemina. Muchas otras familias exhiben expansiones sólo en ciertos linajes. Tal es el caso de la fumarato reductasa subunidad de hierro-azufre, cordenada C3 de la figura, 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 6
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 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 6 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) [@koonin_genomics_2008] 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

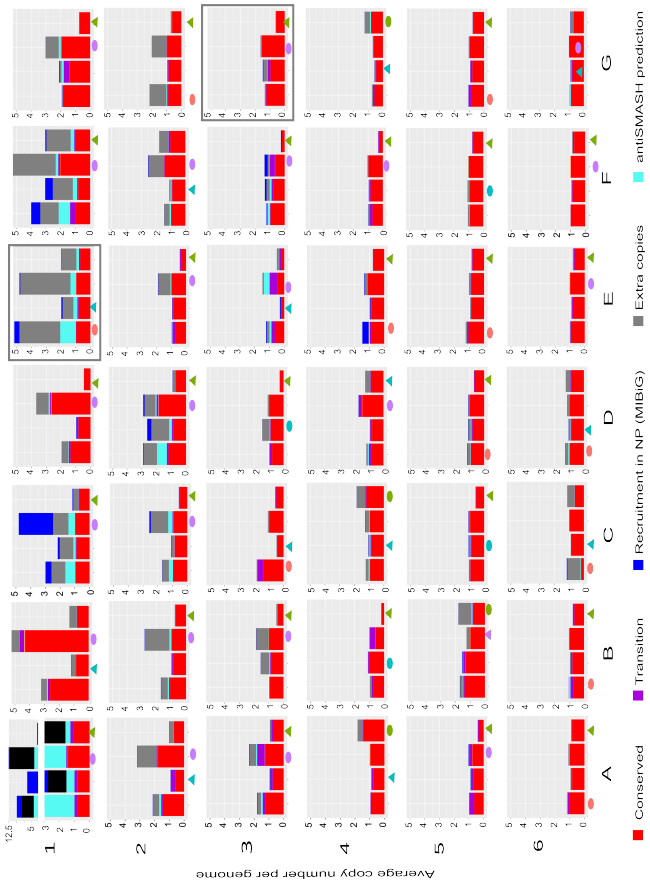

Figure 5: Perfiles de EvoMining de las 42 enzimas conservadas en linajes genómicos seleccionados. Las coordenadas en forma de letras A-G y los números 1-6 se muestran en esta figura para Localice fácilmente la familia y sus propiedades en la Tabla S1.

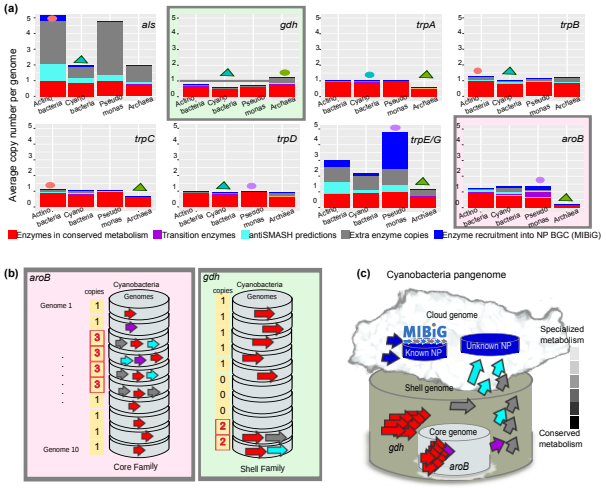

Figure 6: Perfiles de expansiones de EvoMining de enzimas conservadas seleccionadas. (a) Patrones de expansión de las ocho familias conservadas cuyas copias adicionales participan en la biosíntesis de escitonemina. El conjunto completo de 42 EF se muestra en la Fig. S11. La codificación de colores es 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.

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 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 6 panel (a). Las modas de número de copias también son informativas por ello son mostradas en la figura FIGURAFIGURAFIGURA. Una sóla 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 6 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 [@kudo_cloning_2007], 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 7 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 6 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 6 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[@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 **??** , 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 [@balskus_investigating_2008,@soule_comparative_2009] un cluster biosintético de cianobacteria.

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 α-cetoglutarato y amonio. De acuerdo al uso de cofactores, la GDH puede dividirse en tres clases, la primera usa NAD+ y es nombrada como GDH(NAD+). La segunda clase utiliza NADP+ y es

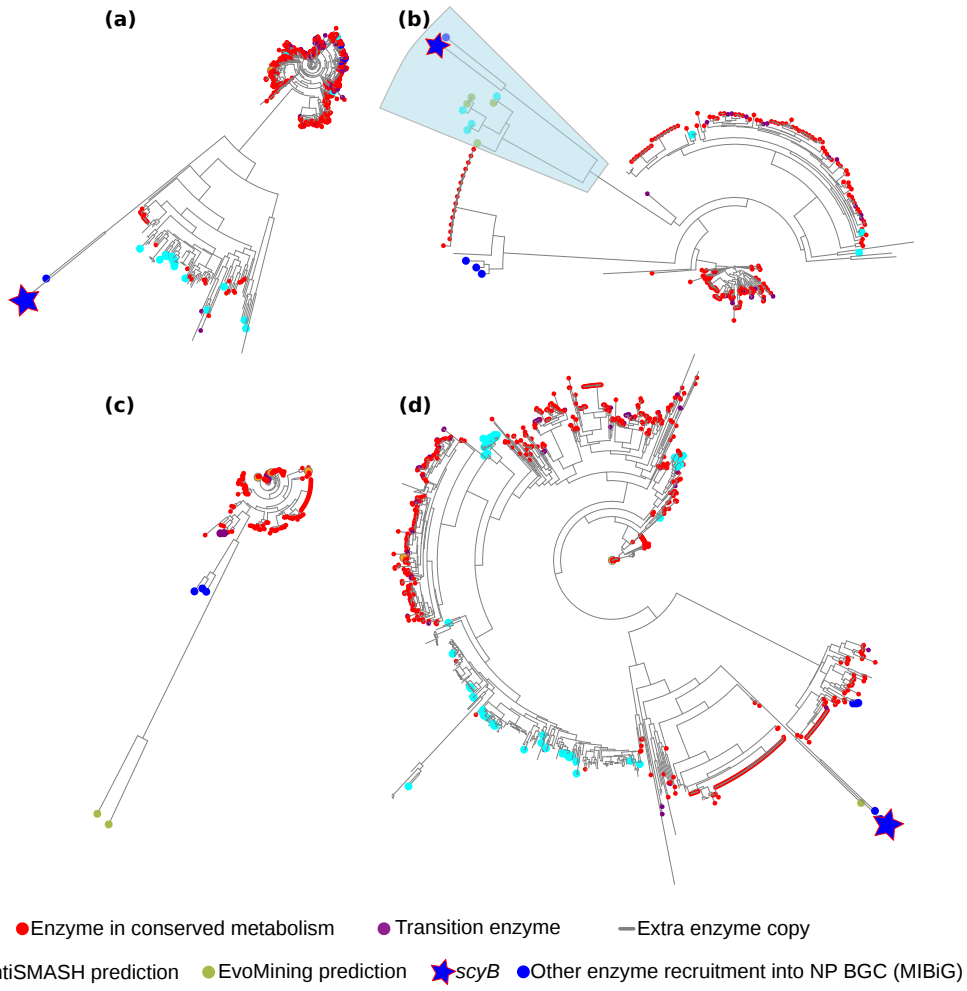

Figure 7: Árboles de EvoMining de glutamato deshidrogenasa en cuatro linajes genómicos. (a) Reconstrucciones filogenéticas específicas del linaje que muestran claras diferencias en los perfiles de expansión en Actinobacteria, Cyanobacteria, *Pseudomonas* y Archaea. Actinobacteria no tiene predicciones de EvoMining, ya que su rama de expansión principal carece de reclutamientos de MIBiG. Sin embargo, es posible que se produzca un metabolito especializado dentro de esas copias de destino desconocido (gris). (b) El árbol de Cyanobacteria posee cuatro predicciones de EvoMining y cuatro de antiSMASH. *scyB* se encuentra junto a esta rama del metabolismo especializado. (c) La mayoría de las copias de *Pseudomonas* están etiquetadas como metabolismo conservado, con solo dos predicciones de EvoMining ubicadas en una rama divergente. *Pseudomonas* tiene un promedio de copias por genoma menor que uno en esta familia lo que se refleja en que casi todas las copias fueron etiquetadas como metabolismo central. (d) Archaea, el taxón más expandido, tiene una rama poblada con expansiones etiquetadas como hits de antiSMASH (cian), pero sin ninguna predicción de EvoMining. Los cuatro linajes tienen reclutamientos de MIBiG, pero *scyB* solo fue reconocido por Actinobacteria, Cyanobacteria y Archaea.

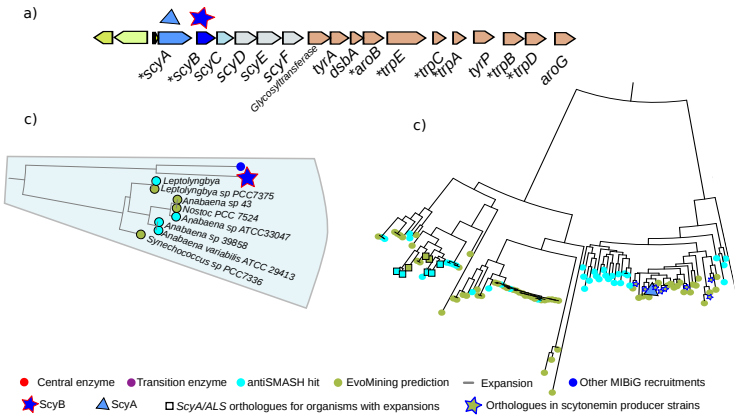

Figure 8: Reclutamientos de GDH y ALS por el cluster biosintético de escitonemina. (a) El BGC de escitonemina de *Nostocpunctiforme* se compone de genes reguladores (verde), genes que participan en la biosíntesis de escitonemina (azul) y de genes dedicados al suministro de precursores (marrón). Se encontró que ocho familias enzimáticas del BGC de escitonemina tienen su origen dentro de las 42 familias enzimáticas conservadas. Estas ocho familias comunes están marcadas con asteriscos. (b) Acercamiento de la rama de expansión de Cyanobacteria GDH cerca de ScyB. Inesperadamente, muchos de los productores conocidos de escitonemina no se encuentran en esta rama. (c) Acercamiento de la rama ScyA, que muestra las expansiones de ALS correcta y exclusivamente marcadas por EvoMining como un destino en el metabolismo especializado. Los productores de escitonemina conocidos están marcados con estrellas. Los cuadrados indican expansiones dedicadas al metabolismo especializado ubicado en la vecindad genómica de las expansiones de GDH que coinciden con la rama ScyB. Los árboles de EvoMining de TrpA, TrpB, TrpC, TrpD, TrpE y AroB de Cyanobacteria están disponibles en Microreact.

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+) es 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 7 panel (a) ,y en cianobacteria Figure 7 panel(b). Las expansiones están prácticamente ausentes en *Pseudomonas* Figure 7 panel (c). En contraste, un número significativo de expansiones es encontrado en Archaea Figure 7 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 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[@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 7 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 7 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[@balskus_genetic_2010]. *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 8 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 8 (c). Estas predicciones incluyen más de veinte organismos conocidos como productores de escitonemina [@balskus_investigating_2008]. Además, ramas cercanas muestran secuencias de ALS que son predicciones 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 8 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.

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 icluso aquellos BGC que en un mismo organismo pueden producir varias moléculas muy similares. El *cluster* de la escitonemina mostrado en la figura Figure 8 panel (a) y en Figure 9 comprende 18 genes [@soule_comparative_2009]. 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 al metabolismo especializado Figure 9. ALS une dos

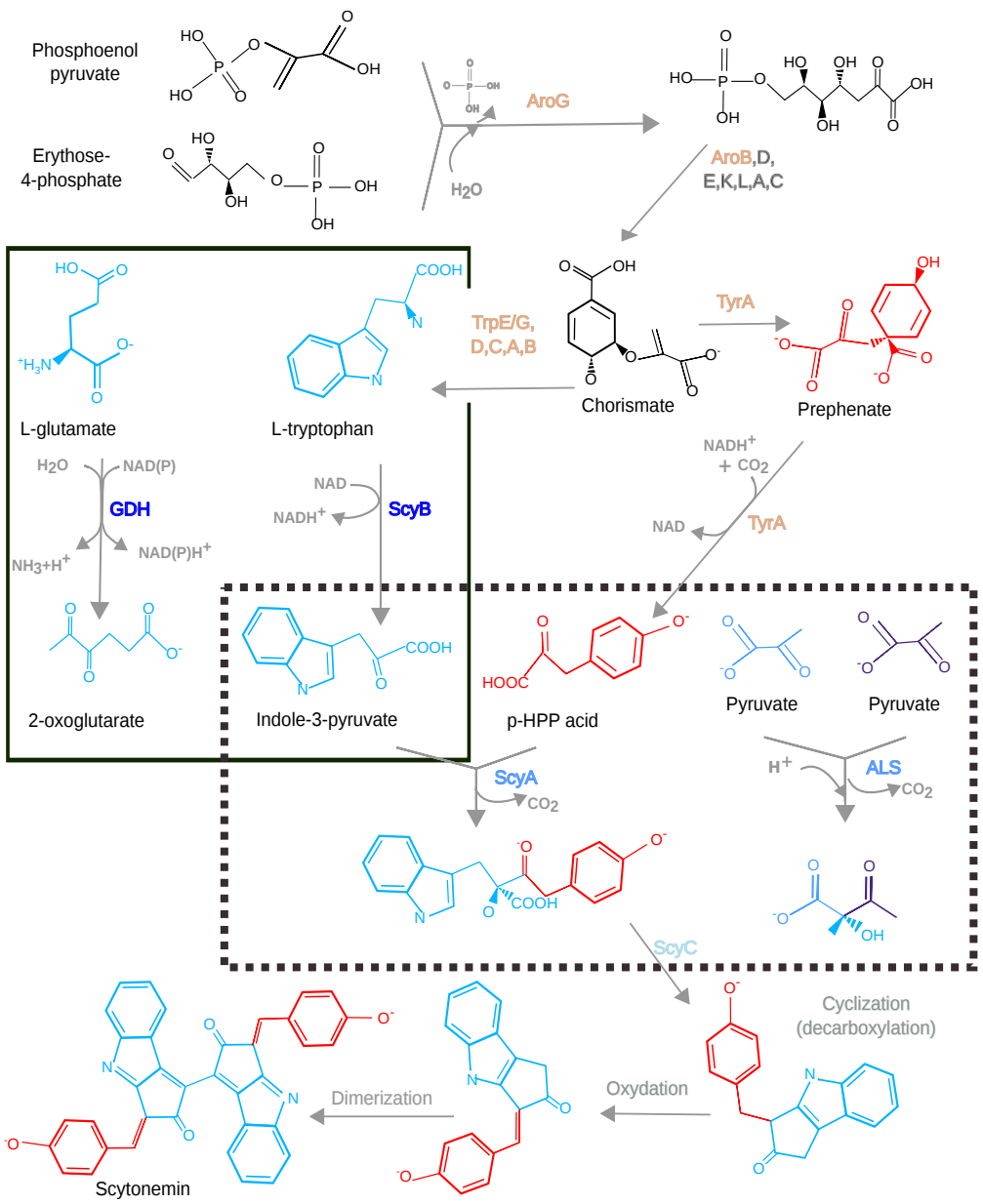

Figure 9: Origen metabólico y destino de GDH / ScyA y ALS / ScyB en la biosíntesis de escitonemina. AroG y AroB participan en la síntesis de corismato, un intermediario que se transforma en los precursores que conducen a sustratos de ScyA, es decir, l-triptófano y prefenato. La reacción catalizada por ScyB que convierte el triptófano en indol-3-piruvato es similar a la conversión de l-glutamato en 2-oxoglutarato, catalizada por GDH (cuadrado con un contorno sólido). ScyA cataliza la descarboxilación de indol-3-piruvato y ácido p-hidroxi-fenil-pirúvico (p-HPP) para formar un dipéptido que sirve como un precursor de escitonemina. Esta reacción es análoga a la descarboxilación de dos piruvatos por la enzima ALS original (rectángulo con un contorno de puntos). ScyC realiza una ciclación seguida de pasos de oxidación y dimerización que concluyen con la ruta de scytonemin. Las enzimas del BGC de escitonemina dedicadas a la síntesis de precursores están coloreadas en marrón y las enzimas biosintéticas en azul.

piruvatos, transformándolos en S-2-acetolactato [@liu_acetohydroxyacid_2016], 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 [@engel_glutamate_2014], 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 útimos pasos[@balskus_investigating_2008].

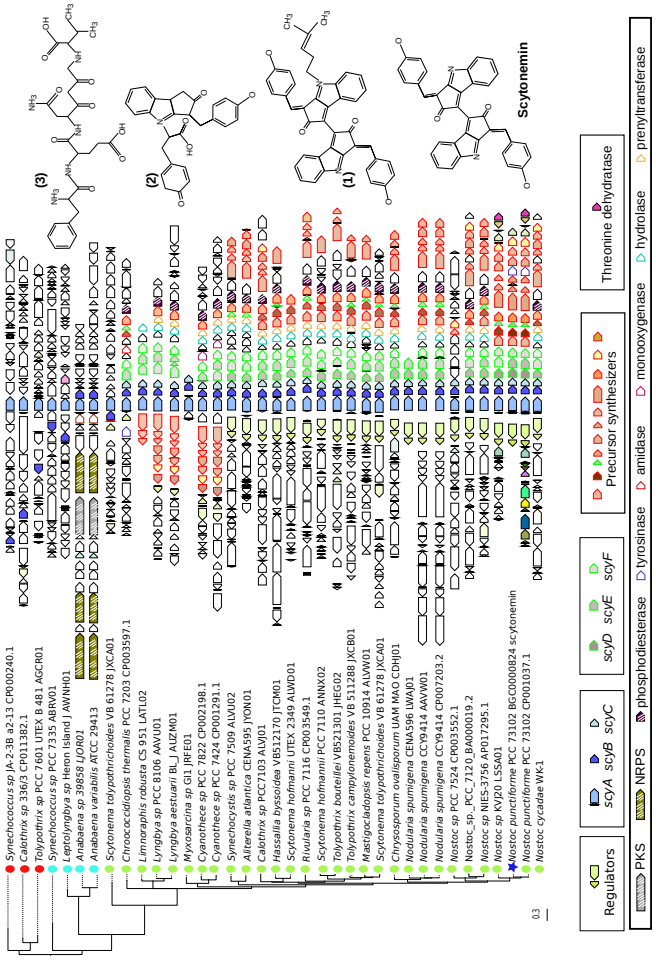

Figure 10: El análisis filogenómico de scyA y sycB mostró una diversidad química en torno a la scytonemin BGC. Las proximidades genómicas que contienen tanto scyA como scyB en cianobacterias se muestran junto a una reconstrucción filogenética utilizando las secuencias de proteínas de estos dos genes.

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 6, Figure 8 . 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 11 [@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] . Estos resultados ilustran como EvoMining puede complementar a antiSMASH mediante la identificación de secuencias de BGC no tradicionales.

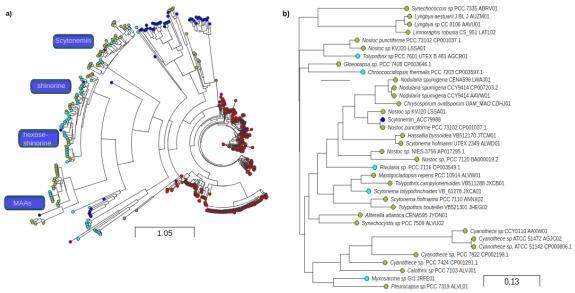

Figure 11: Árbol de EvoMining de AroB. El número promedio de copias de esta familia es 1.3, la moda es uno con 29.5 porciento del total de los genomas superando este umbral. Estas características se reflejan en tres ramas llenas de copias adicionales marcadas ya sea como predicciones de antiSMASH o como predicciones de EvoMining. La primera rama de abajo hacia arriba tiene los compuestos MAA como reclutamientos. La segunda rama fue reclutada para la síntesis de shinorine y la tercera rama para la de escitonemina. Estos tres productos naturales MAAs, shinorine y escitonemina son protectores solares. En contraste a las familias del operón de triptofano, AroB no tiene Welwitindolinone, Ambiguine o fischerindoline como reclutamientos

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 [@navarro-munoz_computational_2018]. 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 concusión 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 10. En esta figura la clasificación de EvoMining para la familia ALS se muestra en un círculo acorde con los colores de EvoMining. 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 validada mediante trabajo experimental. 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.

EvoMining en RNA transferasas

Azoxy

EvoMining en 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 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.

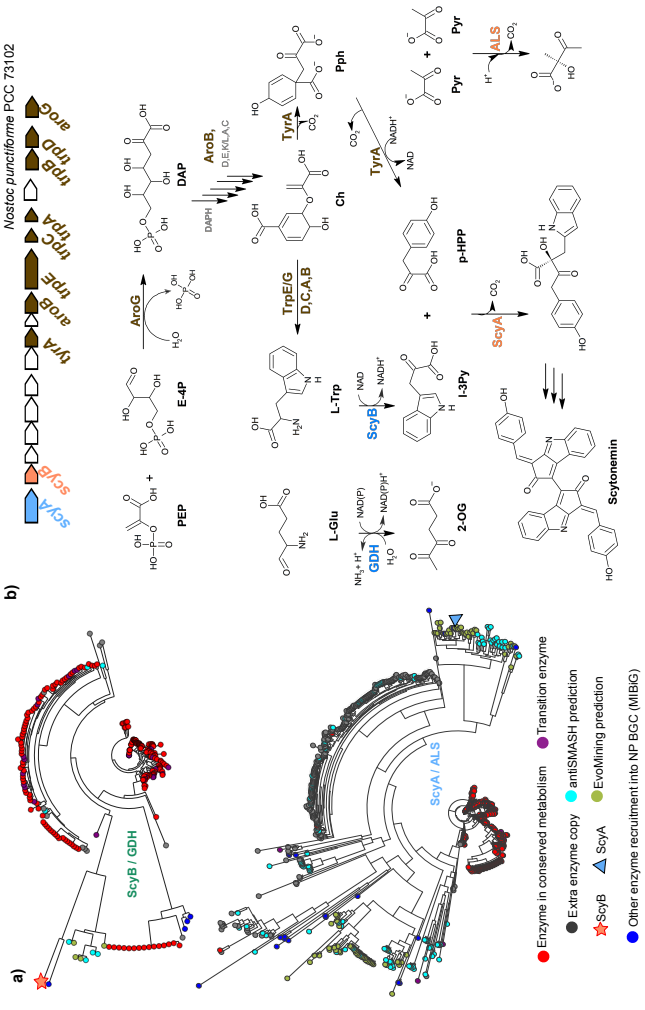

Figure 12: scytonemin

AQUI ARBOL EVOMINING TAUD y GRAFICA CON tauD

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

Table 1: Homólogos de *tauD* en BGC 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*.

Consideraciones finales sobre el uso de EvoMining

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 expeciones 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 [@cruz-morales_phylogenomic_2016] fueron asparagina sintasa, 2-dehidro-3-deoxifosfoheptanoato aldolasa y 3-fosfosquimato-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 [@medema_minimum_2015] 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 ael 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 ezimas, 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 exporar los patrones de neofuncionalización de forma exahustiva 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.