CENTRO DE INVESTIGACIÓN Y ESTUDIOS AVANZADOS DEL INSTITUTO POLITÉCNICO NACIONAL

UNIDAD IRAPUATO

Estudio de la promiscuidad enzimática a nivel enzima, familia y ruta metabólica

Tesis que presenta

Nelly Sélem Mojica

Para obtener el grado de

Doctor en Ciencias

En la especialidad de

Biología Integrativa

Director de Tesis: Francisco Barona Gómez

Irapuato, Guanajuato agosto, 2019

Agradecimientos

Este trabajo de Tesis de doctorado se desarrolló en el Laboratorio de Evolución de la Diversidad Metabólica a cargo del Dr. Francisco Barona Gómez, en el Laboratorio Nacional de Genómica para la Biodiversidad (Langebio) del CINVESTAV-IPN Irapuato. Se agradece el apoyo del CONACyT a través del programa de becas para estudios de posgrado (beca No. 204482) y del SICES del Estado de Guanajuato.

Quiero agradecer a mis compañeros de equipo César Aguilar, Pablo Cruz, Karina Verdel, Ernesto Verduzco, Lorena Hernández, Jana Hiltner, Cuauhtémoc Licona, Alan Yáñez, Daniel Díaz, Paulina Mejía y Ana Juárez por compartir sus conocimientos y su tiempo conmigo. Por la revisión de mi trabajo quedo en deuda con Abraham Avelar, César Aguilar, Pablo Cruz, Lianet Noda y Adriana Espinosa. Por su contribución a mi formación y su interés en emprender proyectos doy gracias a Manuel Villalobos, mi compañero en BetterLab. Agradezco también a Strain Biotech y DNAbits empresas que me ayudaron a desarrollar la faceta de emprendedora de desarrollo de software. A Daniel Hernández mi estudiante de verano le agradezco el honor de haberme escogido como su primera mentora. A Ernesto de CONABIO le extiendo un reconocimiento por su excelente soporte técnico. A los auxiliares y amigos Christian Eduardo Martínez e Hilda Ramos les agradezco su participación en esta investigación tanto con apoyo técnico como con sugerencias innovadoras y aportaciones directas a este trabajo.

Gracias a mi asesor el Dr. Francisco Barona Gómez por incorporarme en su laboratorio y por guiarme de la mejor manera en estos años. A los doctores que fueron miembros de mi comité tutorial Jean Philippe Vielle, Johan van Horebeek, Luis Delaye, Alexander de Luna, Mauricio Carrillo y Chris Henry les agradezco sus aportaciones multidisciplinarias y sus críticas oportunas a mi trabajo. Su contribución a mi formación es invaluable, sin ellos habría sido imposible este camino desconocido de primera estudiante de doctorado de Biología en México.

Agradezco a Ross A. Overbeek de Argonne National laboratory por contribuir en gran manera a mi visión de la genómica computacional y emocionarme al introducirme a Woese y el increíble mundo de las Archaeas. A mis amigos y colaboradores en la Universidad de Wageningen: Marnix Medema y Jorge Navarro por su guía en el estudio de los productos naturales. A Pilar por hacer feliz mi estancia en Holanda. A Michael Mulloney de Northwestern University que con su trabajo experimental retroalimenta los hallazgos bioinformáticos.

Gracias especiales a mi mejor amigo Abraham Avelar por su paciencia, sus enseñanzas, revisión, edición del trabajo, esfuerzo y aportaciones a la vida.

Gracias a la Cocina Tepeyac por su apoyo durante mi doctorado y su compromiso con el deporte mexicano. Gracias también a mis amigas Ximena Morgado y Erika Cruz por su apoyo en las experiencias vividas.

Finalmente, gracias a mi familia, mis padres, mis abuelos, mis tíos por todo su apoyo en el largo y difícil camino que hemos recorrido juntos.

Resumen

Las enzimas son moléculas que catalizan reacciones químicas convirtiendo sustratos en productos. A las enzimas que pueden realizar más de una función química, i.e. catalizar más de una reacción, se les conoce como enzimas promiscuas. Las nuevas funciones resultan de mutaciones en la secuencia que codifica la enzima. Estas mutaciones conllevan a la síntesis de nuevos productos que son sujetos de selección. A su vez estos productos pueden ser incorporados como sustratos de otras enzimas, generando así innovaciones en rutas biosintéticas. Entendiendo la promiscuidad como un mecanismo evolutivo que genera diversidad metabólica propongo generalizar el concepto de promiscuidad enzimática a promiscuidad de una ruta metabólica. En esta tesis abordo el fenómeno de promiscuidad distinguiendo tres niveles: enzima, familia enzimática y ruta metabólica.

Debido a la abundancia de metabolitos en los organismos es prácticamente imposible probar que una enzima no es promiscua. Para esta conclusión se necesitaría probar que la enzima no cataliza reacciones en ninguno de estos sustratos. En cambio, para mostrar promiscuidad se necesitan conocer al menos dos sustratos sobre los cuales la enzima catalice una reacción. Esta observación me llevó a que si bien en este momento no podemos identificar si una enzima es o no promiscua, sí es posible observar marcas dejadas por la adquisición o pérdida de una función, es decir cambios en promiscuidad. En mi trabajo identifiqué que en bacterias los cambios en promiscuidad pueden estar asociados a cambios en la secuencia, en el número de copias de los genes codificantes, en reclutamientos a funcionales, en la vecindad genómica y en la flexibilidad de la molécula. Así pues, para sugerir promiscuidad desarrollé una serie de herramientas que me permitieran sugerir promiscuidad mediante el uso de genómica comparativa en linajes procariontes.

Para identificar cambios en el número de copias de familias enzimáticas, es necesario primero identificar ¿cuáles son estas familias que están cambiando? es decir entender cuáles son las familias enzimáticas conservadas dado un linaje taxonómico. Si las expansiones provienen de rutas de metabolismo central, cómo podemos identificar los genes que codifican este metabolismo. Para abordar estas preguntas, en el primer capítulo diseñé y desarrollé Orthocore, lo que me permite determinar cuáles son las más familias más conservadas dado un linaje taxonómico. En estas familias, resuelvo también cómo identificar

los ortólogos más cercanos y entonces los posiblemente más relacionados a la función primaria. También se muestra que no todas las familias del *core genome* pertenecen a rutas del llamado “metabolismo central”, es decir las rutas de glicólisis, síntesis de aminoácidos, etc. Si no que, también estarán en el *core genome* familias marcadoras, es decir aquellas que están distribuidas en todo el linaje pero que no se encuentran presentes en otros linajes cercanos. Algunas de estas familias marcadoras pertenecen a rutas biosintéticas de metabolismo especializado, como en el caso de *clv* en *Clavibacter michiganensis.* Una aplicación final de este capítulo es utilizar las familias en el *core genome* para reconstruir filogenias difíciles.

Una vez identificadas las familias conservadas quedaban por explorar los cambios en el número de copias, así como las marcas de reclutamientos enzimáticos en nuevas rutas metabólicas. El objetivo de EvoMining, un paradigma de minería genómica concebido previamente a mi trabajo es la búsqueda de nuevas rutas metabólicas mediante la identificación de enzimas divergentes con marcas de reclutamiento. Al identificar la relación entre EvoMining y la promiscuidad continué el desarrollo de EvoMining hasta convertirla en una herramienta interactiva que permite explorar expansiones y reclutamientos sobre distintos linajes taxonómicos. Los principales resultados del capítulo dos son: existen patrones de expansión diferentes para cada familia enzimática según el linaje taxonómico, no todas las familias tienen expansiones, puede haber reclutamientos sin expansiones así como también expansiones sin reclutamientos conocidos y finalmente, no sólo las familias del *core genome* sino también familias de enzimas pertenecientes al *shell genome* i.e. presentes en más del 50% de los organismos, tienen frecuentemente copias extra que contribuyen a la formación de familias del metabolismo especializado.

Las familias de enzimas con reclutamiento señaladas por EvoMining pertenecen a nuevas copias biosintéticas codificadas en clústeres de genes ubicados en la vecindad genómica inmediata de la copia extra. Observé que estas vecindades suelen tener o bien enzimas clásicas de metabolismo especializado (por ejemplo, NRPS y PKS) o bien otras copias extra de familias enzimáticas del *shell genome*. Además, el contenido génico es variable, aunque suele haber un core de genes en estas vecindades también existe variedad en la presencia y ausencia de familias accesorias, lo que sugiere una variedad de metabolitos, es decir una

exploración del espacio químico. Por ello en el capítulo tres propuse generalizar la noción de familia enzimática a familia de vecindades genómicas (clústeres biosintéticos). Para ello desarrollé CORASON una herramienta que considera una familia de vecindades a aquellas que comparten una enzima *query* y al menos alguna otra de la vecindad de referencia. Además, CORASON recupera Orthocore para mostrar la historia filogenética del clúster. La familia de clústeres rimosamida - detoxina es sugerida como una ruta biosintética promiscua debido a la gran cantidad de variantes génicas encontradas. Efectivamente, nuevas variantes moleculares de rimosamida - detoxina fueron encontradas por nuestros colaboradores y comprobadas mediante espectrometría de masas.

En el último capítulo se investiga cuáles de los cambios genómicos están presentes en la familia HisA, cuya ortóloga PriA en Actinobacteria es promiscua. En esta familia no se conocían duplicaciones previamente a mi trabajo. De hecho, en los géneros donde se ha caracterizado PriA como promiscua no se encuentran duplicaciones, pero EvoMining sí localiza en Actinobacteria tanto seis copias extra como un reclutamiento con cambios de vecindad genómica, lo que señala una posible nueva promiscuidad para PriA sugiriendo además el nuevo sustrato. Además, HisF se revela como resultado de una duplicación de HisA en Archaea, Cianobacteria y *Pseudomonas.* Aunque en Actinobacteria no se ha encontrado que HisA sea promiscua para la función HisF es posible que en otros linajes donde la divergencia de secuencia es menor sí puedan encontrarse miembros de la familia HisA capaces de catalizar la reacción HisF. En otros aspectos, simulaciones de dinámica molecular me sugirieron afinidad por nuevos sustratos, misma que no se vio reflejada en actividad catalítica en ensayos enzimáticos. Finalmente, comencé un diseño experimental para medir la interacción de los sustratos nativos de PriA en un mismo ensayo.

Así pues, la respuesta a ¿es esta enzima promiscua?, puede ser mejor comprendida a través de las preguntas ¿esta familia enzimática tiene miembros promiscuos en este linaje taxonómico? ¿De los ortólogos de la familia cuáles son promiscuos para estos determinados sustratos? y finalmente ¿son promiscuas las rutas metabólicas en las que participa la enzima? En este trabajo muestro como la genómica comparativa nos ayudar a guiar información relevante para las respuestas a estas preguntas.

Abstract

Enzymes are molecules that catalyze chemical reactions by converting substrates into products. Enzymes that perform more than one chemical function, i.e. catalyze more than one reaction, are known as promiscuous enzymes. New functions arise as a result from mutations in the sequences encoding enzymes. These mutations lead to the synthesis of new products that are subject to selection. In turn, these products can be incorporated as substrates of other enzymes, thus generating innovations in biosynthetic routes. After understand promiscuity as an evolutionary mechanism that generates metabolic diversity, I propose to generalize the concept of enzymatic promiscuity to the metabolic pathway promiscuity. In this thesis I discuss the phenomenon of promiscuity distinguishing three levels: enzyme, enzyme family and metabolic pathway.

Due to the metabolite's abundance in every organism, it is practically impossible to prove that an enzyme is not promiscuous. Prove that an enzyme is not promiscuous is equivalent to prove that the enzyme does not catalyze reactions on any of these substrates. Instead, to show promiscuity, it needed to have prior knowledge about at least two substrates on which the enzyme catalyzes a reaction. This observation led me to the fact that although at this time we cannot identify whether an enzyme is promiscuous or not, it is nevertheless possible to observe marks left by the acquisition or loss of a function, that is, it is possible to observe changes in promiscuity. In my work I identified that in bacteria the changes in promiscuity may be associated with changes in the sequence, in the number of copies of the coding genes, in recruitments to new pathways, in the genomic neighborhood and in the flexibility of the molecule. The identification of two functions in the same phylogenetic tree, has also been a mark of promiscuity. So, to understand promiscuity I developed a series of tools that would allow me to suggest promiscuity using comparative genomics in prokaryotic lineages.

To identify changes in the copy number of enzymatic families, it is first necessary to identify which are these families that are changing? That also means to understand which enzymatic families are conserved given a taxonomic lineage. If the expansions come from central

metabolic pathways, how can we identify the genes that encode for this metabolism. To address these questions, in the first chapter I designed and developed Orthocore, a tool that allows me to determine the most conserved families in a taxonomic lineage. In these families, I also resolve how to identify the closest orthologs and then those copies possibly most related to the primary function. It is also shown that not all families in the core genome belong to the so-called "routes of central metabolism", such as glycolysis and amino acid synthesis, marker families are also in the core genome. Marker families are those families that are distributed throughout the selected lineage but that are not present in other nearby lineages. Some of these marker families belong to specialized metabolism biosynthetic pathways, as is the case of *clv* in *Clavibacter michiganensis*. A final application of this chapter is the use families in the core genome to reconstruct difficult phylogenies.

Once identified the conserved families, it remained to explore the following promiscuity footprints: changes in the number of copies, two functions annotated in the same tree, and enzymatic recruitments into new metabolic routes. The goal of EvoMining, a genome mining paradigm conceived prior to my work is the search for new metabolic pathways by finding divergent enzymes with recruitment marks. After identifying the relationship between EvoMining and promiscuity, I continued EvoMining development until it became an interactive tool that allows us to explore expansions and recruitments on different taxonomic lineages. The main results of chapter two are: each enzyme family show different expansion patterns for several taxonomic lineages, not all families have expansions, there are recruitments without expansions as well as expansions without known recruitments and finally, not only core families genome but also shell families i.e. families present in more than 50% of organisms, often have extra copies that contribute to the formation of families of specialized metabolism.

Recruitments in enzyme families identified by EvoMining usually are encoded in biosynthetic gene clusters located in the immediate genomic neighborhood of the extra copy. I noticed that these neighborhoods usually contain either classic enzymes of specialized metabolism (e.g. NRPS, or PKS) or other extra copies of enzymatic families from the shell genome. In addition, the gene content is variable, although there is usually a core of genes in these neighborhoods there is also variety in the presence and absence of accessory families, which

suggests a variety of metabolites, that is, an exploration of the chemical space. Therefore, in chapter three I proposed to generalize the notion of an enzymatic family to a family of genomic neighborhoods (biosynthetic clusters). To this end, I developed CORASON a tool that considers a family of neighborhoods as those that share a query enzyme and at least some other gene of the neighborhood of reference. CORASON also recovers Orthocore to show the phylogenetic history of the cluster. The family of rimosamide-detoxin clusters' is suggested as a promiscuous biosynthetic pathway due to the large number of gene variants found. Indeed, new molecular variants of rimosamide - detoxin were confirmed by our collaborators and tested by mass spectrometry.

The last chapter investigates which of the genomic changes are present in the HisA family, whose orthologous PriA in Actinobacteria is promiscuous. In this family there were no known duplications prior to my work. In fact, in those genera where PriA has been characterized as promiscuous, no duplications were found, but EvoMining does find in Actinobacteria: six extra copies and one recruitment, both with changes in its corresponding genomic neighborhood. These changes may indicate a possible new promiscuity for PriA, also suggesting the new substrate. In addition, another member of the histidine operon HisF is pointed by EvoMining as a result of a duplication of HisA in Archaea, Cianobacteria and Pseudomonas. Although in Actinobacteria it has not been found that HisA is promiscuous for HisF function, it is possible that in other lineages with lesser sequence divergence, members of the HisA family can be found capable of catalyzing HisF reaction. In other aspects, molecular dynamics simulations suggested affinity for new substrates, which was not reflected in catalytic activity in enzymatic assays. Finally, I began an experimental design to measure the interaction of native PriA substrates in the same trial.

The answer to is this enzyme promiscuous? could be better understood by considering the following questions, does this enzymatic family have promiscuous members in this taxonomic lineage? Of the family orthologs which are promiscuous for certain known substrates? And finally, are the metabolic pathways in which the enzyme participates promiscuous? This work uses changes identified by comparative genomics that helps to guide insights to answer those questions.

Artículos derivados de este trabajo

1. Pablo Cruz-Morales, Johannes Florian Kopp, Christian Martinez-Guerrero, Luis Alfonso Yáñez-Guerra, Nelly Selem-Mojica, Hilda Ramos-Aboites, Jörg Feldmann, Francisco Barona-Gomez. **Phylogenomic analysis of natural products biosynthetic gene clústeres allows discovery of arseno-organic metabolites in model streptomycetes.** Genome biology and evolution 8 (6), 1906-1916 **(2016)** DOI: 10.1093/gbe/evw125

2. Pablo Cruz-Morales, Hilda E Ramos-Aboites, Cuauhtémoc Licona-Cassani, Nelly Selem-Mójica, Paulina M Mejía-Ponce, Valeria Souza-Saldívar, Francisco Barona-Gómez. **Actinobacteria phylogenomics, selective isolation from an iron oligotrophic environment and siderophore functional characterization, unveil new desferrioxamine traits.** FEMS microbiology ecology 93 (9) **(2017)** DOI: 10.1093/femsec/fix086

3. Jana K Schniete, Pablo Cruz-Morales, Nelly Selem-Mojica, Lorena T Fernández-Martínez, Iain S Hunter, Francisco Barona-Gómez, Paul A Hoskisson. **Expanding primary metabolism helps generate the metabolic robustness to facilitate antibiotic biosynthesis in Streptomyces** MBio 9 (1), e02283-17 **(2018)** DOI: 10.1128/mBio.02283-17

4. Karina Gutiérrez-García, Edder D Bustos-Díaz, José Antonio Corona-Gómez, Hilda E Ramos-Aboites, Nelly Sélem-Mojica, Pablo Cruz-Morales, Miguel A Pérez-Farrera, Francisco Barona-Gómez, Angélica Cibrián-Jaramillo. **Cycad Coralloid Roots Contain Bacterial Communities Including Cianobacteria and Caulobacter spp. That Encode Niche-Specific Biosynthetic Gene Clusters.** Genome biology and evolution 11 (1), 319-334 **(2018)** DOI: 10.1093/gbe/evy266

5. Enrique Jesús Delgado-Suárez, Nelly Selem-Mojica, Rocío Ortiz-López, Wondwossen A Gebreyes, Marc W Allard, Francisco Barona-Gómez, María Salud Rubio-Lozano. **Whole genome sequencing reveals widespread distribution of typhoidal toxin genes and VirB/D4 plasmids in bovine-associated nontyphoidal Salmonella.** Scientific reports 8 (1), 9864 **(2018)** DOI: 10.1038/s41598-018-28169-4

6. Nelly Selem-Mojica, Cesar Aguilar, Karina Gutiérrez-García, Christian Martínez-Guerrero, Francisco Barona-Gomez. **EvoMining reveals the origin and fate of natural products biosynthetic enzymes.** Microb Genom. Apr 4. doi: 10.1099/mgen.0.00026. **(2019)** DOI: 10.1099/mgen.0.000260

7. Jorge Navarro-Muñoz, Nelly Selem-Mojica, Michael Mullowney, Satria Kautsar, James Tryon, Elizabeth Parkinson, Emmanuel De Los Santos, Marley Yeong, Pablo Cruz-Morales, Sahar Abubucker, Arne Roeters, Wouter Lokhorst, Antonio Fernandez-Guerra, Luciana Teresa Dias Cappelini, Regan Thomson, William Metcalf, Neil Kelleher, Francisco Barona-Gomez, Marnix H Medema. **A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data**. BioRxiv, 445270 **(2019)**

Resúmenes de congresos donde se presentó este trabajo

1. Sélem-Mojica Nelly, Cruz-Morales Pablo, Mejía-Ponce Paulina M, Aguilar César, Ramos-Aboites Hilda, Barona-Gómez Francisco. **CORe Analysis of Syntenic Orthologs to Prioritize Natural Products Biosynthetic Gene Glusters (CORASON).** ISBA 2018

2. Sélem-Mojica Nelly, Cruz-Morales Pablo, Aguilar César, Ramos-Aboites Hilda, Jorge Navarro-Muñoz, Nelly Selem-Mojica, Michael Mullowney, Regan Thomson, William Metcalf, Neil Kelleher, Marnix Medema, Barona-Gómez Francisco. **Genome Mining methods to identify molecular and genetic variations.** Oaxaca Challenges and Synergies in the Analysis of Large-Scale Population-Based Biomedical Data from November 26 to December 01, 2017.

3. Jorge Navarro-Muñoz, Nelly Selem-Mojica, Michael Mullowney, Satria Kautsar, James Tryon, Elizabeth Parkinson, Emmanuel De Los Santos, Marley Yeong, Pablo Cruz-Morales, Sahar Abubucker, Arne Roeters, Wouter Lokhorst, Antonio Fernandez-Guerra, Luciana Teresa Dias Cappelini, Regan Thomson, William Metcalf, Neil Kelleher, Francisco Barona-Gomez, Marnix H Medema. **A computational framework for rapid exploration and prioritization of biosynthetic diversity from large-scale genomic data**. abstract ID 2899 8th Congress of European Microbiologists (FEMS2019)

Software desarrollado listado según el capítulo donde fue utilizado

Capítulo	GitHub	DockerHub
Capítulo 1	Distribución de myRAST Desarrollo de Orthocore Tutorial de Orthocore Desarrollo de clavigenomics Tutorial de clavigenomics	nselem/myrast nselem/orthocores nselem/clavigenomic
Capítulo 2	Sistematización de EvoMining Tutorial de EvoMining	nselem/evomining
Capítulo 3	Desarrollo de CORASON Tutorial de CORASON Tutorial de CORASON - BiG SCAPE	nselem/corason
Capítulo 4	Distribución de EVcouplings Análisis recursivo de rutas	nselem/evcouplings nselem/ev_dependencies nselem/ev_data
Tesis	Otros datos, R markdown y scripts de R para figuras	

Enlaces de datos genómicos utilizados en este trabajo

1. Nelly Selem-Mojica, Cesar Aguilar, Karina Gutiérrez-García, Christian Martínez-Guerrero, Francisco Barona-Gomez. (2018). EvoMining genomic and enzyme databases for Actinobacteria, Cianobacteria, *Pseudomonas* and Archaea. Zenodo. http://doi.org/10.5281/zenodo.1219709

2. Jorge Navarro-Muñoz, Nelly Selem-Mojica, Michael Mullowney, Satria Kautsar, James Tryon, Elizabeth Parkinson, Emmanuel De Los Santos, Marley Yeong, Pablo Cruz-Morales, Sahar Abubucker, Arne Roeters, Wouter Lokhorst, Antonio Fernandez-Guerra, Luciana Teresa Dias Cappelini, Regan Thomson, William Metcalf, Neil Kelleher, Francisco Barona-Gomez, Marnix H Medema. (2018). Genomic data for "A computational framework to explore large-scale biosynthetic diversity". http://doi.org/10.5281/zenodo.1532752

Índice general

Introducción

Las enzimas catalizan reacciones químicas transformando sustratos en productos. Durante el siglo XX, las enzimas fueron percibidas como catalizadores altamente específicos, sin embargo, esta percepción cambió con el descubrimiento de que pueden participar en varias reacciones [Jensen, 1976]. Esta capacidad de catalizar varias funciones químicas se conoce como promiscuidad enzimática. Por ejemplo, se ha calculado que *Escherichia coli* contiene al menos 404 enzimas promiscuas [Nam, 2012]. Otro estudio ha mostrado que en mutantes sencillas de 104 genes esenciales de *E. coli* en el 20% de la auxotrofía puede ser rescatada debido a la actividad promiscua de otra enzima [Patrick, 2007]. La relevancia de la promiscuidad radica tanto en su papel como mecanismo de evolución de la función enzimática [Jensen, 1976; Pandya, 2014; Khersonsky 2010], así como en la necesidad de su detección para la corrección de modelos de flujo metabólico y la determinación de efectos secundarios en drogas farmacológicas. A pesar de su frecuencia e importancia aún se está en el proceso de entender las causas y las características observables de la promiscuidad enzimática.

Para estudiar la promiscuidad es necesario contar con una definición, algunos autores emplean el término promiscuidad para describir actividades enzimáticas distintas a la función principal conocida [Khersonsky, 2010]. Otros lo ven como una actividad secundaria fortuita [Copley, 2003] que pudo aparecer de forma accidental o inducida artificialmente [Hult, 2007]. Otros más, cuando una enzima puede operar sobre un amplio rango de sustratos, prefieren llamarla multiespecífica [Khersonsky, 2010]. A la acción de realizar distintas funciones catalíticas, ya sea al catalizar varias reacciones químicas o bien una misma reacción en sustratos diferentes se le conoce como promiscuidad enzimática [Obrien, 1999]. Existen varios tipos de promiscuidad enzimática.

-Por sustrato cuando la reacción es la misma, pero se lleva a cabo en distintos sustratos. Como ejemplo tenemos a la familia PriA, una isomerasa de Actinobacteria que actúa abriendo anillos de 5 carbonos en dos sustratos ProFAR y PRA [Barona-Gómez, 2003]. Otro ejemplo es la familia de las betalactamasas [Risso, 2014].

-Catalítica cuando la enzima utiliza diferentes mecanismos de reacción y/o residuos catalíticos, e.g. la quimotripsina puede catalizar reacciones de amidasa y fosfotriesterasa en un mismo sitio activo. [Obrien, 1999]

-Por condiciones del entorno, cuando la enzima cambia su conformación dependiendo de las condiciones químicas y físicas presentes como pH, temperatura, solventes orgánicos y salinidad e.g. algunas lipasas pueden actuar como sintetizadoras de ésteres en lugar de hidrolasas en presencia de solventes orgánicos [Hult, 2007; Kumari 2007].

La promiscuidad por sustrato es importante en términos evolutivos, por ejemplo, la *enzyme commission number* (EC) separa las enzimas en clases, donde a cada enzima se asignan 4 dígitos, los tres primeros corresponden a la reacción y el último al sustrato; el mayor número de sustratos (4306 clases) que de reacciones químicas (234 en el tercer nivel) sugiere que la mayor variación evolutiva se da a nivel de sustrato y no de reacción [Li, 2004]. Otra evidencia de la importancia de la multiespecificidad por sustrato está en el descubrimiento de las superfamilias, enzimas mecanística y estructuralmente relacionadas que divergen en su afinidad por sustrato [Glasner, 2006; Baier 2016].

Si bien existen familias de enzimas con alta especificidad por sustrato, otras familias como el citocromo P450 [Bloom, 2007; Nath, 2008] y las β-lactamasas [Zou, 2015] son promiscuas. Es posible que la visión previa de alta especificidad se deba a que las primeras rutas metabólicas estudiadas pertenecen al metabolismo central, donde la especificidad puede haber sido favorecida por presiones de selección [Firn, 2009]. En contraste en el metabolismo especializado o secundario, es decir aquel que no es esencial para la gran mayoría de procariontes, sino que funciona para producir metabolitos especializados que facilitan la adaptación del organismo a ciertos ambientes parece que la promiscuidad está presente [Weng, 2012]. Se favorece especialmente la capacidad de producir varios metabolitos, es decir la promiscuidad de producto, ya sea debida a la promiscuidad de sustrato de una enzima o bien a las diferencias en el contenido génico del clúster de genes al que pertenece dicha enzima de metabolismo especializado. Al ampliar las enzimas estudiadas, esta visión previa de alta especificidad ha cambiado dando paso al conocimiento de más enzimas con funcionalidad [Jia, 2013], sin que por ello se afecte la eficiencia catalítica por la función primaria [Aharoni, 2005]. En 1976 el interés por la promiscuidad comenzó por

su influencia en la evolución de la función enzimática [Jensen, 1976; Pandya, 2014], las aproximaciones variaron desde la aparición de la síntesis funcional [Dean, 2007], cuando la disponibilidad de genomas permitió la combinación de análisis filogenéticos con técnicas de biología molecular, bioquímica y biofísica. En 2003 la biofísica de las proteínas entra en escena al postularse que la diversidad conformacional durante la dinámica molecular debe incidir en la aceptación de distintos sustratos. Recientemente se ha investigado su papel en efectos secundarios en drogas farmacológicas [Nobeli, 2009; Hopkins, 2009; Nath, 2010, Von-Eichborn, 2011; Zhang, 2012; Zou, 2015]. Entre 2005 y 2010 se avanza del estudio de una sola familia enzimática hacia el interés por propiedades globales, por ejemplo, dado un genoma se investiga la distribución de familias promiscuas en subsistemas metabólicos [Nam, 2012]. En estos años, surge el desarrollo de índices que reflejan las características bioquímicas de enzimas promiscuas [Nath, 2010]. En 2010, comienzan los intentos por desarrollar un método computacional de predicción de promiscuidad [Carbonell, 2010]. Desde 2012 a la fecha, a la par que las aproximaciones bioinformáticas se multiplican [Nagao, 2014; Cheng, 2012], se desarrollan investigaciones de aspectos biofísicos [Noda-garcia, 2013; Zou, 2015], bioquímicos [Verdel-Aranda, 2015; Plach, 2016] y evolutivos [Copley, 2015; Espinosa-Cantu, 2015] de enzimas promiscuas reafirmando que todos estos aspectos están relacionados al fenómeno. En las siguientes secciones se describirán trabajos importantes sobre la relación que guarda la promiscuidad con expansiones genómicas y flexibilidad molecular. Además, se hablará sobre análisis bioquímicos y metabólicos para la descripción del fenómeno.

Función biológica de la promiscuidad enzimática

¿Por qué existe la promiscuidad enzimática? Se tiene evidencia de dos papeles biológicos: el primero proporcionar robustez a la red metabólica de un organismo mediante redundancia de reacciones de otras enzimas [Patrick, 2007]; el segundo permitir plasticidad evolutiva, es decir materia prima para la adaptación a variaciones ambientales [Aharoni, 2005; Sanchez-Ruiz, 2012; Martínez-Núñez, 2015] mediante la adquisición de nuevas funciones químicas. Respecto a la robustez, se probó que sobre expresar enzimas promiscuas puede rescatar pérdidas génicas [Patrick, 2007]. De 104 knockout sencillos de genes esenciales para $E. coli$ K-12, 20% de las auxotrofias pudieron ser suprimidas por la sobre expresión de plásmidos que contenían enzimas promiscuas. Otro ejemplo que aporta a la robustez es PriA, enzima

de la ruta de histidina que realiza en la ruta del triptófano la reacción E.C. 5.3.1.24 [Barona-Gómez, 2003]. En cuanto a la plasticidad se propone que para que la promiscuidad pueda dar origen a la aparición de nuevas funciones la actividad promiscua debe proveer una ventaja fisiológica inmediata para poder ser seleccionada positivamente, además una vez que una función promiscua se vuelva relevante se debe poder mejorar mediante pocas mutaciones derivando en el intercambio entre la actividad promiscua y la principal [Khersonsky, 2010].

Aun cuando el producto de la promiscuidad genera metabolitos que no se integran al metabolismo central de la célula, su efecto es positivo ya que estos metabolitos podrían colaborar a la adaptación al entorno participando por ejemplo en una relación de simbiosis o de competencia con otros organismos. Este tipo de metabolitos, por lo general, no son dañinos [Notebaart, 2014; Linster, 2013, Khanal, 2015] y pueden servir como bloques de construcción para vías metabólicas nuevas [Ma, 2013; Adams, 2014, Soskine, 2010]. La respuesta inmediata de adaptación de un organismo podría ser una consecuencia de su grado de promiscuidad.

PriA como ejemplo de la dinámica evolutiva entre la especialización y la promiscuidad de una familia de enzimas.

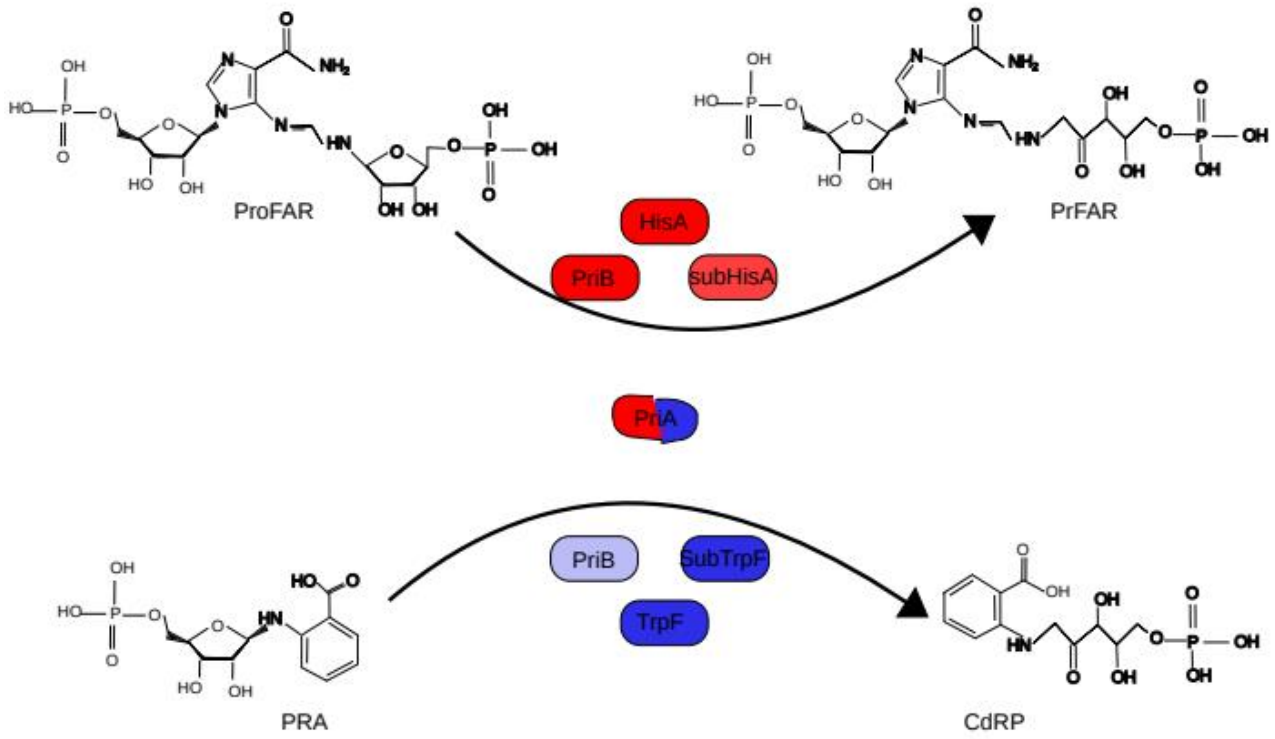

Figura 1 PriA cataliza la isomerización de los sustratos nativos ProFAR y PRA. Convertir ProFAR es un paso de la síntesis de histidina catalizado por HisA en otros linajes bacterianos. Análogamente la isomerización de PRA es catalizada por TrpF en otros linajes. En Actinobacteria PriA se ha dividido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

Los operones *his* y *trp* de histidina [Fondi, 2009] y triptófano [Merino, 2008] respectivamente, participantes del metabolismo de aminoácidos, están ampliamente distribuidos en los organismos bacterianos. En Actinobacteria la familia promiscua PriA participa simultáneamente en ambas rutas biosintéticas (Figura 1), desarrollando el equivalente a la función TrpF en la ruta de histidina y a la función TrpF en la ruta de triptófano. Como HisA isomeriza el sustrato ProFAR y como PRA isomeriza el sustrato PRA. Para su estudio se han generado datos bioquímicos, genómicos y estructurales. En bacterias gram negativas están presentes los operones *his* y *trp* y en lugar de PriA existen dos familias HisA y TrpF. PriA es homóloga de HisA y adquirió la función de TrpF en Actinobacteria coincidentemente

con que en este phylum *trpF* está ausente de la mayoría de sus miembros. PriA se ha diversificado un conjunto de subfamilias en Actinobacteria, por ejemplo, en *Streptomyces*, existe la subfamilia PriB disminuye su actividad de TrpF debido a la presencia de una copia de *trpF* en alguna parte del genoma [Verduzco-Castro, 2016]. En otras Actinobacterias *trpF* se pierde totalmente por ello PriA la familia homóloga de HisA, se vuelve promiscua [Barona-Gómez, 2003] realizando tanto la función química correspondiente a HisA como la de TrpF. Finalmente, en la familia subHisA se pierde la función TrpF debido posiblemente a la ganancia del operón *trp* completo [Noda-García, 2013] y en la familia subtrpF se conserva solo a la función TrpF debido a la pérdida del operón *his* [Juarez-Vázquez, 2017]. Existen al menos 43 familias de Actinobacteria sin explorar respecto a la funcionalidad de PriA. En la tabla 1 se muestran las constantes catalíticas de PriA estimadas en diferentes organismos para estos dos sustratos.

Fuente	Familia	HisA *in vivo*	TrpF *in vivo*	$Kcat^{ProFAR}[M^{-1}s^{-1}]$	$Km^{ProFAR}[\mu M]$	$\frac{Kcat}{Km}ProFAR$	$Kcat^{PRA}[M^{-1}s^{-1}]$	$Km^{PRA}[\mu M]$	$\frac{Kcat}{Km}PRA$	Referencia
Escherichia coli	HisA	-	-	1.6	4.9	3.1	-	-	0	Henn-Sax 2002
Escherichia coli	TrpF	-	-	-	-	0	12.2	34.5	2.82	Sterner 1996
Mycobacterium tuberculosis	PriA	-	-	19	0.23	12	21	3.6	0.17	Due 2011
Mycobacterium smegmatis	PriA	*	*	2.6 ± 0.5	0.85 ± 0.04	0.33	7.9 ± 2.4	3.1 ±0.43	0.39	Verduzco-Castro 2016
Streptomyces globisporus	PriA	*	*	4.2 ± 0.8	0.74 ± 0.03	0.18	11 ± 1.0	3.8 ± 0.2	0.34	Verduzco-Castro 2016
Streptomyces coelicolor	PriA	-	-	3.6 ± 0.7	1.3 ± 0.2	0.4	5.0 ± 0.08	3.4 ± 0.09	0.7	Noda-Garcia 2010
Streptomyces ipomoeae	PriB	*	*	3.8 ± 0.2	0.82 ± 0.02	0.21	60.8 ± 1.1	8.25 ± 0.4	0.14	Verduzco-Castro 2016
Streptomyces sp. Mg1	PriB	*	*	13.2 ± 3.4	0.92 ± 0.19	69	129.6 ± 34	0.29 ± 0.04	0.0022	Verduzco-Castro 2016
Streptomyces sp. C	PriB	*	*	11.4 ± 3.4	2.53 ± 0.74	0.22	149. 9 ± 29	1.4 ± 0.12	9	Verduzco-Castro 2016
Streptomyces sviceus	PriB	*	*	3.9 ± 0.89	0.69 ± 0.04	0.18	24.5 ± 4.0	1.6 ± 0.29	67	VVerduzco-Castro 2016
Corynebacterium diphteriae	subHisA	-	-	4.4 ± 0.5	2.6 ± 0.3	0.59			0	Noda-Garcia 2013
Corynebacterium jeikeium	PriA	-	-	2.3 ± 0.2	0.9 ± 0.08	0.39	5.1 ± 1.0	1.6 ± 0.16	0.31	Noda-Garcia 2013
Corynebacterium striatum	subHisA	-	-	6.9 ± 0.7	2.1 ± 0.5	0.3			0	Noda-Garcia 2013
Corynebacterium diphteriae L48I-F50L-T80S	subHisA			4.5 ± 1.5	0.6 ± 0.08	0.13	133 ± 10	0.05 ± 0.01	0.0004	NNoda-Garcia 2013
Actinomyces urogenitalis DSM 15434	PriB	*	*	2.1 ± 0.5	1.8 ± 0.2	0.9	26.3 ± 6.3	0.37 ± 0.09	14	Verduzco-Castro 2016
Actinomyces odontolyticus ATCC 17982	subTrpF		*	-	-	0			0.02	Juarez-Vazquez 2017
Actinomyces oris K20 BABV01	PriA					0.02			0.01	Juarez-Vazquez 2017
Actinomyces sp. oral taxon 171 str. F0337	PriA					0.01			4	Juarez-Vazquez 2017
Actinomyces sp. oral taxon 848 str. F0332	subTrpF		*	-	-	0			0.0001	Juarez-Vazquez 2017
Actinomyces urogenitalis DSM 15434	PriA					0.01			0.02	Juarez-Vazquez 2017
Bifidobacterium adolescentis L2-32	PriA	*	*			0.2			0.1	Juarez-Vazquez 2017
Bifidobacterium gallicum DSM 20093	PriA	*	*			0.1			0.04	Juarez-Vazquez 2017
Bifidobacterium longum ATCC 15697	PriA	*	*			0.1			0.3	Juarez-Vazquez 2017
Camera CAM1	Metagenoma			1.7 ± 0.1	0.3 ± 0.03	0.2	40 ± 7	3.5 ± 0.04	0.09	Noda-Garcia 2015
CAM1 A81G	Metagenoma			1.7 ± 0.2	0.1 ± 0.01	0.06	32.2 ± 1.7	1.9 ± 0.1	0.06	Noda-Garcia 2015
CAM1 A81S	Metagenoma			4.0 ± 0.9	0.2 ± 0.03	0.04	23.5 ± 6.5	0.5 ± 0.1	0.02	Noda-Garcia 2015
CAM2	Metagenoma			n.d.	n.d.	0	n.d.	n.d.	0	Noda-Garcia 2015
PriA Ancestral	Ancestral			9.4±1.6	0.3±0.009	0.03	4.3±0.4	0.6±0.02	0.13	Verduzco, Noda, sin publicar
PriA SubHisA	Ancestral			3.7±1.01	0.5±0.03	0.1	-	-	0	Verduzco, Noda, sin publicar
SubHisA Ancestral	Ancestral			6.3±0.7	0.15±0.03	0.02	-	-	0	Verduzco, Noda, sin publicar
SubHisA PriA	Ancestral			27.7±3.4	0.05±0.005	2	167.82	0.03±0.002	0.0001	Verduzco, Noda, sin publicar
Streptomyces acidiscabies		0	0	163.6	0.1					Verduzco*
A visco			46	1.37	36	3.4				Juárez*

23

Relación del pangenoma con la promiscuidad enzimática

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

En el Dominio Bacteria se estima que alrededor de 200 familias de secuencias están altamente conservadas. Si bien no están en el $core$, estas secuencias están compartidas por 90% de los genomas [Halachev, 2011]. El *core* depende de los genomas seleccionados, entre menos amplio sea el rango filogenético elegido mayor será el tamaño del $core$. Dada su conservación el $core\ genome$ puede utilizarse para trazar mejores relaciones filogenéticas que las obtenidas con el uso exclusivo de marcadores como la subunidad 16s del RNA ribosomal o la proteína RpoB. El *dispensable* y el *shell genome* son juntos el conjunto complemento del *core genome*, es decir todas aquellas secuencias que están ausentes de uno o más organismos del grupo y por lo tanto no son necesarias para todos, sino sólo posiblemente para el organismo que las posee. Como en estos grupos la presión de selección está relajada respecto al *core-genome* [Firn, 2009] es el conjunto ideal donde la plasticidad genómica tiene facilidades para desarrollarse.

Esta idea puede restringirse a subsistemas metabólicos para identificar genes cuyas enzimas están en proceso de cambio de función química, por ejemplo, en este trabajo se encontró que el gen *trpF* está presente en sólo 49 de 290 genomas analizados del género

Streptomyces. Por lo que *trpF* se encuentra en el *dispensable genome* de este género taxonómico, posiblemente adquiriendo una nueva función [Ma, 2013]. Para evitar problemas técnicos del cálculo del pangenoma existen modelos de medición de variabilidad del genómica entre especies bacterianas [Kislyuk, 2011].

Modelos bioinformáticos de promiscuidad

Con el fin de reducir la inversión en el proceso de experimentación, se han implementado algoritmos computacionales para predecir promiscuidad enzimática [Carbonell, 2010; Cheng, 2012; Nagao, 2014; Noda-García, 2015; Garcia-Seisdedos, 2012]. Estos procedimientos cuentan con un conjunto de aprendizaje, unos descriptores del conjunto, una fase de ajuste de parámetros y finalmente una predicción. En 2010, Carbonell propone un algoritmo de soporte vectorial basado en subsecuencias de distinto tamaño que llama huellas moleculares. En este trabajo aplicado sobre 500,000 proteínas reportadas en la enciclopedia de Kioto de genes y genomas (KEGG) se reporta 85% de éxito en detección de enzimas promiscuas anotadas en KEGG. En 2012, Cheng compara los métodos de random forest y soporte vectorial en 6799 proteínas provenientes de la base de datos *Universal Protein Resource* (UniProt). Las enzimas son descritas con subsecuencias de aminoácidos incorporando además características biofísicas como polaridad. Se utiliza como grupo de control a familias de enzimas donde nunca se ha reportado una enzima promiscua.

Un aspecto no considerado en estos métodos es que hay familias de enzimas con alta identidad de secuencia entre sus miembros, con cambios bruscos en promiscuidad, debidos por ejemplo a la dinámica genómica [Noda-García, 2013; Verdel-Aranda, 2015], o bien que ortólogos de la misma familia enzimática pueden variar en su grado de promiscuidad [Bloom, 2007]. Esta característica dificulta considerar solo la secuencia conlleve a buenos predictores de promiscuidad. Obtener una predicción positiva utilizando los modelos existentes significa que, dada esa secuencia, en su familia se conoce previamente un elemento promiscuo y que además sus subsecuencias de cierto tamaño son suficientemente similares. Estos enfoques no pueden predecir *de novo*, en familias donde la promiscuidad no ha sido previamente detectada experimentalmente, pues no consideran aspectos evolutivos ni mecanísticos de las enzimas.

Otra limitante a los enfoques descritos es que mezclan en su conjunto de entrenamiento fenómenos distintos de promiscuidad. Cheng p. g. incluye enzimas *moonlight* que, si bien poseen funciones adicionales a la catalización, son distantes a las enzimas promiscuas [Copley, 2003]. Además, en ambos casos mezclan en el mismo conjunto enzimas bacterianas y eucariotas, con lo que si existía una huella basada en secuencia entonces ésta puede diluirse por la gran distancia taxonómica entre estos grupos.

Promiscuidad *in vitro* y promiscuidad *in vivo*

La ganancia de promiscuidad no sólo puede entenderse como la capacidad de convertir más sustratos [Carbonell, 2010], sino también como la mejora de la capacidad catalítica respecto a ellos. El I-index [Nath, 2008], está definido como un rango de valores entre 0 y 1 que tiende a 1 entre más parecida sea la actividad de la enzima sobre distintos sustratos, la capacidad catalítica es medida en términos del cociente de *Michaelis - Menten* $\frac{K_{cat}}{K_m}$. El índice ha sido utilizado para predecir la afinidad por sustrato del citocromo P450 [Nath, 2010]. Una limitante del índice I es que se deben conocer los sustratos a los que la enzima es afín; sin embargo, se puede sospechar que una enzima ha ganado promiscuidad aun sin conocer sus potenciales sustratos. Otro punto por señalar es que las variables K_{cat},K_m son mediciones realizadas *in vitro* y no se consideran todos los sustratos presentes *in vivo*. Para solventar esta dificultad e investigar variaciones de sustratos nativos se pueden buscar productos similares a los ya conocidos por medio de análisis bioinformáticos o metabólicos [Nesvizhskii, 2007] como los empleados en la detección de rutas no conservadas en la biosíntesis de productos naturales [Medema, 2015]. En particular para este fin se ha utilizado espectroscopia de masas $\frac{MS}{MS}$, [Nesvizhskii, 2007; Campbell, 2012] combinada con molecular networking para identificar productos similares [Yang, 2013; Kocher, 2007]

El papel de la dinámica molecular en la promiscuidad

La mayoría de las estructuras tridimensionales de proteína se han obtenido por cristalización de rayos X. Aunque mucho se ha hablado de la relación estructura función, al cristalizar se obtienen estados conformacionales homogéneos, que bien pueden no ser la única conformación que adopta la proteína en solución [James, 2003]. En particular en el problema de promiscuidad, se ha observado que la variación funcional no queda obviamente reflejada

en la variación estructural, lo que sugiere un rol significativo para la dinámica molecular [Parisi, 2015; Noda-García, 2013; Zou, 2015]. Se postula que un aspecto de la dinámica molecular relevante para la diversificación de especificidad por sustrato es el número de confórmeros [Javier-Zea, 2013]. Por ejemplo, en la Actinobacteria *Corynebacterium diphtheriae* parece que el contexto genómico correlaciona con pérdida de promiscuidad de PriA ya que al poseer el genoma una copia de *trpF*, la enzima perdió esta función química conservando solo la función EC 5.3.1.16 correspondiente a la ruta de histidina. Esta subfuncionalización se refleja en la pérdida de estados conformacionales cambiando desde un estado en *C. diphtheriae* hasta cuatro presentes en la dinámica de PriA de *M. tuberculosis* [Noda-García, 2013].

Las regiones rígidas de una enzima proporcionan orientación adecuada con respecto a los grupos catalíticos, mientras que las regiones flexibles permiten al sitio activo adaptarse a los sustratos con diferentes formas y tamaños [Copley, 2003]. Esta consideración sugiere que la flexibilidad del sitio activo es otra característica de la dinámica molecular a considerar para obtener información de la capacidad de ligación de una enzima a distintos sustratos [Gatti-Lafranconi, 2013]. Recientemente el índice de flexibilidad dinámica (dfi) se utilizó como una medida cuantitativa basado en la respuesta a perturbaciones de aminoácidos (PRS). Este índice se incrementó en regiones cercanas al sitio activo de beta lactamasas promiscuas respecto al correspondiente dfi de β-lactamasas especialistas existentes [Zou, 2015].

Modelo biológico: la diversidad de Actinobacteria

Al escoger un conjunto acotado para investigar familias de enzimas promiscuas se debe recordar que la funcionalidad es jerárquica por lo que, para mejorar la anotación, es deseable reflejar el proceso evolutivo y restringirse a un grupo de organismos taxonómicamente relacionados [Cruz-Morales, 2016]. Actinobacteria es un phylum que posee promiscuidad tanto en el metabolismo periférico como en el core metabólico. Entre datos públicos (NCBI) y privados están disponibles en 2014 alrededor de 1200 genomas no redundantes de especies de Actinobacteria. Como punto de partida, se han estudiado las relaciones filogenéticas y grupos de ortología [Li, 2003; Waterhouse, 2013], en particular para identificar relaciones entre las familias del phylum Actinobacteria, se han obtenido árboles multilocus de entre 100 y 157 genomas [Gao, 2012, Sen, 2014]. Estos estudios sugieren cómo separar

los genomas disponibles para hacer el cálculo de grupos de ortología. Finalmente, se han realizado estudios de plasticidad genómica en *Streptomyces* considerando 5 y 17 organismos de los 300 genomas disponibles en la actualidad [Zhou, 2012; Kim, 2015] donde reportan respectivamente 2,018 familias en el *core genome* y 32,574 en el pangenoma. Además de Actinobacteria, otros linajes con riqueza metabólica conocida en metabolitos secundarios como Cianobacteria y $Pseudomonas$ también serán considerados en este trabajo. Finalmente, Archaea un dominio desconocido, pero del que comienzan a haber genomas disponibles ~800 genomas públicos en NCBI en 2016. Todos estos linajes pueden ser considerados en la búsqueda de variación de promiscuidad.

Modelo metabólico biosíntesis de aminoácidos.

Al hacer el cálculo vemos que *Streptomyces*, un género del phylum Actinobacteria cuenta en su genoma con un promedio de 8316 secuencias codificantes según la especie. Gran parte de estas secuencias pueden ser agrupadas en subsistemas metabólicos como metabolismo de carbohidratos o de lípidos; de estos subsistemas uno de los más amplios es el metabolismo de aminoácidos con entre 429 y 910 secuencias según el organismo. La síntesis de aminoácidos es un subsistema presente en todas las especies, pero con suficientes variaciones que permiten hacer observaciones evolutivas. En un gran número de Actinobacterias las rutas de histidina y triptófano de 7 y 11 pasos respectivamente convergen en una enzima funcional llamada PriA, que realiza tanto la función de HisA como la de TrpF [Barona-Gómez 2003]. La cantidad de familias en el subsistema de metabolismo de aminoácidos, su variabilidad, su conservación entre distintos grupos taxonómicos y la existencia de estos ejemplos en Actinobacteria posicionan al metabolismo de aminoácidos como un buen punto de partida para la búsqueda de promiscuidad tanto de familias promiscuas como de miembros promiscuos de las mismas.

En las cuatro décadas de estudio de la promiscuidad enzimática, hemos aprendido que es un fenómeno distribuido en distintos subsistemas metabólicos [Nam, 2012] y que su existencia puede deberse tanto al desarrollo de nuevas funciones para fines adaptativos [Jensen,1976; Obrien, 1999; Firn, 2009; Copley, 2015], como al rescate de una función perdida [Patrick, 2007; Barona-Gómez, 2003]. Por ello la dinámica de pérdida y ganancia de genes asociada al contexto genómico en bacterias se relaciona con cambio en la función

enzimática [Zhao, 2013; Zhao, 2014, Copley, 2015]. Precisando, respecto a la ganancia de genes, se postula que la funcionalidad precede la duplicación [Hughes, 1994; Obrien, 1999]. Lo que implica que dada una duplicación muy posiblemente previamente la promiscuidad estuvo presente [Gerlt, 2001; Huang, 2012; Noda-Garcia, 2013; Risso, 2014].

Se han desarrollado técnicas bioquímicas y metabólicas de medición [Nath, 2008], así como algoritmos computacionales de predicción de promiscuidad [Carbonell, 2010; Cheng, 2012, Noda-García, 2015]. Un aspecto que mejorar dentro del modelado es la restricción del conjunto de estudio a un grupo taxonómico tan reducido que exista congruencia en las familias de ortología y a la vez tan amplio que permita observar efectos evolutivos; el phylum Actinobacteria ha probado tener ejemplos de promiscuidad. Si bien la secuencia no ha sido suficiente para la correcta predicción de promiscuidad [Verdel-Aranda, 2015, Copley, 2015], es posible que, dentro de las técnicas computacionales, incorporar consideraciones evolutivas o bien flexibilidad durante la dinámica molecular nos proporcionan sugerencias sobre promiscuidad de los miembros de una familia [James, 2003; Javier-zea, 2013; Noda-Garcia, 2013; Zou, 2015].

Antecedentes

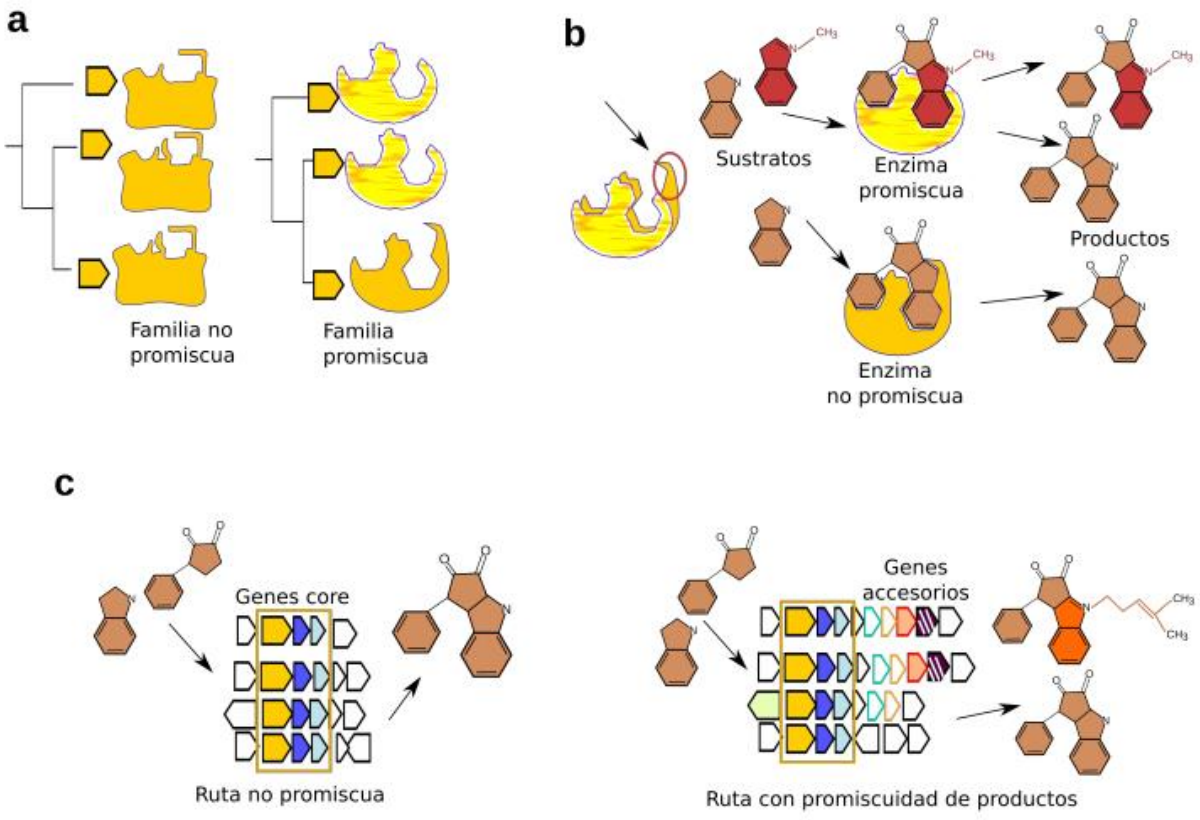

Figura 2: La promiscuidad puede entenderse a distintos niveles metabólicos: enzima, familia enzimática y ruta biosintética.

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

Si se entiende a la promiscuidad como funciones alternativas de alguna unidad molecular, puede observarse promiscuidad a distintos niveles: desde un mismo gen que presenta splicing alternativo, una misma enzima con funciones alternativas, o bien una familia enzimática donde al menos algunos homólogos codifican enzimas promiscuas [Nobeli, 2009]. Existen también rutas que generan productos metabólicos alternativos [Lamble, 2003; Weng, 2012], si se considera como la unidad de estudio a una ruta biosintética podemos generalizar la noción de promiscuidad al concepto de rutas promiscuas. La (Figura 2) muestra tres niveles en los que se puede estudiar la promiscuidad: i) Distinguiendo familias de enzimas promiscuas de familias especialistas, ii) Distinguiendo enzimas específicas de

enzimas especialistas en una familia enzimática promiscua y finalmente iii) encontrando promiscuidad en rutas de metabolismo especializado.

Para clarificar la diferencia entre identificar familias promiscuas y miembro de familias promiscua se propone el siguiente ejemplo: PriA en Actinobacteria y HisA en enterobacteria son ambas familias que isomerizan ProFAR en la ruta de síntesis de histidina, pero solo la familia PriA es promiscua pues puede además isomerizar el sustrato PRA durante la síntesis de triptófano [Barona-Gómez, 2003]. Sin embargo, dentro de Actinobacteria, existen miembros no promiscuos de PriA. Cinéticas *in vitro* han mostrado que en algunas especies del género *Actinomyces* los homólogos de *priA* codifican para enzimas monofuncionales en alguno de los dos sustratos [Juárez-Vázquez, 2017]. Así pues, dentro de una familia promiscua no todos los miembros tienen esta propiedad [Bloom, 2007]. En este trabajo se buscará encontrar marcas de promiscuidad a nivel familia, enzima y ruta biosintética en el metabolismo especializado.

Antecedentes conceptuales

Figura 3: Antecedentes conceptuales de promiscuidad metabólica

Algunos autores han intentado identificar enzimas promiscuas mediante aprendizaje máquina utilizando únicamente la secuencia de aminoácidos. Estos enfoques no han distinguido entre los problemas identificación de familias promiscuas e identificación a nivel de enzima [Carbonell, 2010]. Hasta ahora utilizar únicamente la información de la secuencia de aminoácidos no ha sido suficiente para identificar una familia promiscua sin conocer previamente al menos un miembro promiscuo de ella. Por otra parte, diferenciar la promiscuidad a nivel de enzima una vez que se conoce una familia promiscua se dificulta cuando la identidad de secuencia es alta. Tal es el caso de las PriA que han perdido la promiscuidad en algunos miembros del género *Actinomyces* [Juárez-Vázquez, 2017]. En este caso el establecimiento de las filogenia de los miembros del género permitió entender la distribución de la promiscuidad.

Para mejorar nuestro entendimiento del fenómeno de promiscuidad, además de la comparación de secuencias es necesario integrar otros elementos al análisis, Es difícil medir la promiscuidad en términos absolutos, por ejemplo, no se puede aseverar que una enzima es “no promiscua” sin haber previamente descartado a todos los posibles sustratos del universo químico. Además, incluso enzimas que resultan no promiscuas en análisis *in vitro* resultan que sí son promiscuas al examinarlas *in vivo* [Noda-García, 2012]. Sin embargo, debido a que al adquirir una nueva función existe un umbral donde la función ancestral es conservada, es plausible intentar transformar el problema de “encontrar promiscuidad” al de “encontrar cambios en promiscuidad” al relacionar estos cambios con las huellas que dejan los cambios funcionales [Soskine, 2010; Aharoni, 2005; Bloom 2007]. Entre los elementos relevantes que correlacionan con la adquisición de una función alternativa se encuentran además de divergencia de secuencia, diversidad en vecindad genómica (Zhao, 2014), la pérdida o ganancia de genes [Juárez-Vázquez, 2017], las expansiones génicas, i.e. el crecimiento del pangenoma dentro de un grupo taxonómico [Martínez-Núñez, 2015] y finalmente cambios estructurales o de flexibilidad durante la dinámica molecular [Zou, 2015; Gatti, 2013]. Estos elementos tienen en común que reflejan un cambio en alguna propiedad genómica o biofísica observable en el registro evolutivo, de lo que se deriva que el buscar cambios en la promiscuidad de una enzima, familia o ruta, resulta más factible por ahora que la búsqueda intrínseca de promiscuidad.

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

El establecimiento de un marco de conservación permite distinguir cambios

La función de una enzima es un concepto jerárquico, dependiente de la filogenia de un organismo [Szklarczyk, 2015]. Por ello, para poder encontrar marcas de cambio funcional, por ejemplo, diferencias en número de copias de una familia, primero es importante trabajar en la construcción de un marco filogenético consistente que permita ordenar inclusive organismos de la misma especie. La dificultad de esta tarea consiste en que, si los organismos que se desea ordenar son muy cercanos, marcadores clásicos como el gen de 16s rRNA son también muy parecidos en secuencia y no permiten resolver las relaciones entre ellos. Este caso dificulta la construcción de un árbol filogenético de *Actinomyces* y con ello se imposibilitaba encontrar patrones en la matriz de presencia / ausencia de genes [Juárez-Vázquez, 2017].

Para solventar la falta de resolución de genes individuales en organismos cercanos puede utilizarse el conjunto de todos los genes comunes en un linaje. Este conjunto es conocido como el *core genome*. Se han desarrollado herramientas bioinformáticas para este problema, por ejemplo, phyloPhlan [Segata, 2013] fija 400 genes comunes en bacteria y trata de localizarlos dado un conjunto de genomas sin importar su linaje. Este enfoque enfrenta que el contenido génico de genomas procariontes suele ser muy variable debido a mecanismos como transferencia horizontal y duplicación génica [Land_insights_2015, Koonin, 2015]. Esta variabilidad puede impedir encontrar muchos de estos 400 genes en un cierto linaje. Entre organismos de la misma especie pueden suceder fenómenos como que el *core* siempre se reduzca al aumentar un nuevo genoma y que el conjunto total de familias génicas (pangenoma) siempre aumenta. Aunado a esta observación biológica están también las limitaciones técnicas, hay genes que no aparecen en un genoma porque este fue mal secuenciado o ensamblado y por lo tanto estos genes disminuyen el tamaño del core. Para distinguir cambios en el contenido génico en un linaje de organismos es útil establecer primero un orden entre los genomas del linaje a analizar. Para ello, un camino que generaliza

la propuesta de phyloPhlan es localizar los genes del *core* exclusivos de cada grupo de genomas. Al momento existe publicada *metaphor*, una herramienta de selección de ortólogos [Van-der-Veen, 2014] pero al comenzar este trabajo no existía un método disponible para ello. Por ello el capítulo uno de esta tesis habla del desarrollo de Orthocore, que obtiene los genes del *core* y además los concatena y entrega un árbol filogenético.

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

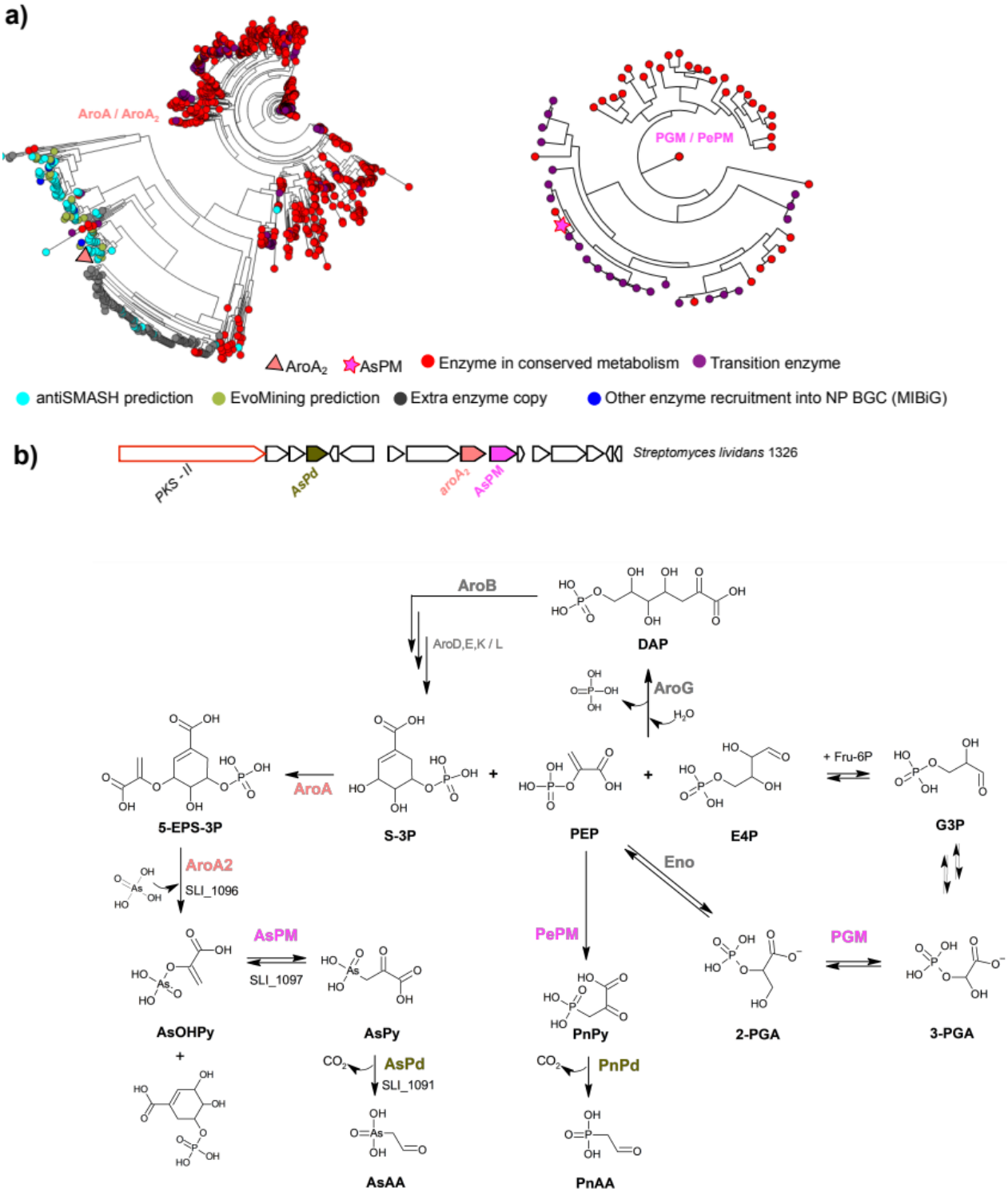

Figura 4. El paradigma de EvoMining permitió localizar familias de enzimas de metabolismo central con copias extra que habían sido reclutadas para una nueva función en metabolismo especializado. Un ejemplo de ello es AroA, que cataliza una reacción similar, pero sobre un sustrato diferente. La copia primaria cataliza una reacción sobre un metabolito con azufre y la secundaria sobre un metabolito con arsénico

Una parte del metabolismo especializado está compuesta por familias enzimáticas que evolucionaron de rutas de metabolismo central [Caetano-Anolles, 2009]. En las familias expandidas, ya sea por duplicación o por transferencia horizontal, las expansiones pueden retener la función química de las rutas centrales [Schniete, 2018; Verdel-Aranda, 2015], así como también la función alternativa suele estar presente aún a bajos niveles antes de la divergencia o duplicación [Soskine, 2010; Aharoni, 2005, Bloom, 2001]. Por tanto, las familias con expansiones en un linaje son candidatas a ser familias promiscuas en él. Se ha notado que en el linaje en que una familia enzimática es promiscua hay una zona de cambio en promiscuidad [Noda-García, 2015], Las expansiones de rutas centrales que participan en la síntesis de productos naturales son candidatos para presentar cambios en promiscuidad tanto a nivel familia como a nivel enzima. Por ello la observación de la retención de función ancestral al aparecer una función alternativa proporciona una zona favorable para la búsqueda de promiscuidad a nivel de enzima. En la familia existirá un gradiente de promiscuidad donde las más cercanas a la zona de duplicación o divergencia tienen más indicaciones de tener un cambio en promiscuidad que las más conservadas y cercanas al metabolismo central. Estas copias extra son candidatas para pertenecer a rutas de metabolismo especializado.

Como prueba de concepto esta idea de minar genomas incorporando información evolutiva permitió la identificación de la biosíntesis de arseno lípidos [Cruz-Morales, 2016]. En este trabajo se desarrolló la idea de EvoMining, un método de minería genómica que permite localizar familias enzimáticas con copias extra en su linaje genómico que además presentaran marcas de reclutamiento a metabolismo especializado. La aplicación de las ideas de EvoMining llevaron a identificar una expansión en el número de copias de la enzima AroA en Actinobacteria. Se demostró experimentalmente que la copia secundaria era parte de un clúster de síntesis de arsenolípidos, . Es interesante remarcar que las funciones tanto primaria como secundaria son similares, lo que cambia es la especificidad de sustrato en cada copia. En la copia primaria el sustrato tiene un átomo de fósforo, en la secundaria uno de arsénico. Otros genes del clúster biosintético también parecen ser expansiones de familias del metabolismo central. Este descubrimiento se realizó con la primera versión de EvoMining que contó con 200 genomas de Actinobacteria [Cruz-Morales, 2016], una base de datos de secuencias de enzimas de productos naturales y otra base de datos de

secuencias de enzimas de rutas centrales curada a mano. En esta primera versión quedaba pendiente poder utilizar EvoMining como plataforma de exploración de linajes personalizables con genomas provistos por el usuario, así como sistematizar la información del conjunto de enzimas consideradas como partícipes de BGC de productos naturales.

La búsqueda de productos naturales cuenta entre sus premisas que estos se producen en vecindades genomas llamadas clústeres y que además clústeres cercanos (ya sea en contenido génico o en la secuencia de sus componentes), exploran variaciones metabólicas, es decir sus enzimas catalizan reacciones sobre sustratos parecidos, aunque no idénticos [Cruz-Morales, 2016; Medema, 2015]. Los clústeres de genes biosintéticos (BGC) de productos naturales que han sido caracterizados experimentalmente están depositados gracias el esfuerzo de toda la comunidad en la base de datos *Minimal information about a biosynthetic gene clúster* MIBiG. Actualmente MIBiG cuenta con 1400 BGC.

EvoMining ha demostrado incorporar efectivamente premisas evolutivas para localizar enzimas pertenecientes a estos BGC aun cuando pertenezcan a clases no conocidas. Análogamente otra herramienta de minería genómica ha utilizado con éxito la búsqueda de copias extra para encaminar su búsqueda específicamente a BGC de resistencia a antibióticos [Alanjary, 2017]. Desarrollar EvoMining en combinación con algoritmos de búsqueda de cambios en la vecindad genómica la harán una plataforma ideal para abordar el problema de las familias, proporcionando una solución la dificultad de no tener conocimiento previo de un miembro promiscuo en la familia investigada. Respecto al problema de los miembros, se propone explorar variaciones en vecindad genómica, flujo génico y dinámica molecular, como candidatos a reflejar la variación en promiscuidad. Uno de los modelos modelo biológicos a explorar será el phylum Actinobacteria, un grupo de bacterias reconocido por su diversidad metabólica donde se ha probado la existencia de promiscuidad enzimática. Existen otros modelos interesantes como $Pseudomonas$ y Cianobacteria, linajes donde según MIBiG hay abundancia de BGC. En contraste Archaea no tiene más que un BGC reportado, esto puede deberse a un fenómeno biológico o bien a falta de exploración en dicho dominio. Aunque se ha avanzado tanto en el conocimiento de la promiscuidad en Archaea [Martinez-Núñez, 2017] como en la detección de productos naturales, sin necesariamente caracterizar los genes biosintéticos [Charlesworth, 2015]

EvoMining podría ayudar a entender los patrones de expansión reclutamiento y sugerir familias con promiscuidad enzimática en este dominio.

EvoMining es una plataforma bioinformática de análisis de expansión y reclutamiento de familias que puede ser utilizada para la identificación de enzimas que participan en clústeres de síntesis de productos naturales. Si se combinara EvoMining con la premisa de que vecindades distintas son marcadoras de funciones químicas distintas, al encontrar una familia expandida con vecindades genomas diferentes se podría solventar la deficiencia de otros métodos bioinformáticos consistente en que para identificar familias promiscuas se debe conocer previamente un miembro promiscuo de la misma. Así pues, al combinar EvoMining con herramientas de vecindad genómica tanto de comparación como de visualización estaremos mejorando su funcionalidad en la identificación de familias promiscuas.

La genómica comparativa como herramienta en la priorización de clústeres promiscuos

La promiscuidad nos interesa por su producción de variantes. Si bien en rutas centrales rescata la función, en metabolismo secundario crea nuevas variantes moleculares que permiten adaptación, de hecho, pangenomas grandes correlacionan con aparición de nuevas funciones enzimáticas. Considero que el concepto de promiscuidad puede ser extendido a un nuevo nivel considerando a los clústeres de genes biosintéticos como una nueva unidad. Así pues, existiría la promiscuidad de una enzima, promiscuidad de una familia y promiscuidad de un clúster.

Expansión y contextos genómicos como herramienta de anotación funcional

Al evaluar la herramienta de análisis de promiscuidad PROMISE [Carbonell, 2010] en un set de datos de la familia HisA/PriA [Noda-García, 2015] obtuve que en su mejor desempeño es (huella molecular de tamaño 6) clasifica correctamente casi todas las no promiscuas, (HisA) pero no sucede lo mismo con la familia PriA donde tiene éxito en 16 de 45 casos. Al aplicar el mismo tamaño de huella a 9 miembros promiscuos de la familia IlvC no consigue predecir correctamente ninguno de ellos reflejando tal vez que en su conjunto de entrenamiento no

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

La diversidad enzimática existente es el resultado de un proceso de expansión, mutación y selección que se ha desarrollado durante el transcurso de la historia evolutiva [Khersonsky, 2010; Pearson, 2012]. Existe evidencia de que cierto grado de promiscuidad o divergencia funcional precede a la duplicación génica [Hughes, 1994]. Por este motivo detectar expansiones ya sea duplicaciones o transferencias horizontales [Treangen, 2011], puede ser un buen punto de partida para determinar divergencia funcional y promiscuidad. No todas las expansiones denotan cambio de función enzimática, algunas pueden ser meros accidentes, sin embargo, dado que la función de una enzima suele estar relacionada con sus vecinos [Overbeek, 1999; Zhao, 2014; Zhao, 2013, Verdel-Aranda, 2015], una expansión en una vecindad genoma diferente de la tradicional será un referente de adquisición de una nueva función y entonces un indicador de existencia previa de promiscuidad.

Para sistematizar el estudio de contextos y vecindades genómicas se desarrolló Search Tool for the Retrieval of Interacting Genes/Proteins STRING [Snel, 2000], que cuenta con una anotación de ortología jerárquica y consistente, realizada en 2000 organismos en cuyo marco interacciones de proteínas con implicaciones funcionales son predichas tanto de novo por información genómica de coocurrencia como por minería de datos en artículos publicados. STRING es una base de datos, y como tal no permite agregar nuevos genomas para su análisis. Sus 2000 organismos incluyen especies tanto bacterianas como eucariotas. Al existir tanta diversidad, los genomas disponibles para un género o clase específicos son escasos, p. g. de los más de 300 genomas disponibles de *Streptomyces* solo 24 están incluidos.

Para resolver la baja cobertura de STRING hacia ciertos grupos taxonómicos se pueden desarrollar scripts de vecindad genómica utilizando RAST (Rapid Annotation using Subsystem Technology); un servicio interactivo de anotación automática de genomas de bacterias y arqueas [Aziz, 2008; Overbeek, 2014] donde la función de cada gen se asigna de acuerdo con conocimiento previo de subsecuencias de organismos cercanos filogenéticamente, cuando es posible se incluye en un subsistema metabólico. Estamos en una era de explosión de datos genómicos, próximamente se espera contar con millones de

genomas bacterianos incluso provenientes de bacterias no cultivables, por ello los algoritmos deben ser constantemente optimizados a los nuevos volúmenes de datos [Medema, 2015]. Ante esta expectativa sería muy útil desarrollar algoritmos de análisis genómico que sean de código libre o al menos interactivos para que cada laboratorio pueda personalizarlos para sus propios genomas.

Finalmente, no solo la vecindad genómica inmediata puede ser utilizada como distintivo en la búsqueda de promiscuidad, diferencias en el contexto genómico en genes relacionados con una enzima promiscua, sin importar su ubicación dentro del genoma también pueden ser relevantes para la pérdida o ganancia de función química [Noda-García, 2013; Juárez-Vázquez, 2017].

Contexto y vecindades genómicas

En 2012 fueron analizados 102 genomas de 29 familias de Actinobacteria [Noda, 2012]. sugiriendo que al menos en *Corynebacteria* el contexto y la vecindad genómica incidían en la subfuncionalización de PriA en subHisA [Noda-García, 2013]. Respecto a IlvC, otra familia involucrada en la síntesis de aminoácidos fue estudiada y caracterizada bioquímicamente en 1 $Corynebacterium$ y 8 $Streptomyces$ [Verdel-Aranda, 2015]. Para ampliar estos resultados, utilizando la anotación de RAST y una generalización de la definición de vecindad de STRING, se diseñó un algoritmo para identificar vecindades similares así como uno de visualización de contexto, ambos disponibles como software libre en GitHub nselem/perlas .

El algoritmo de clasificación de vecindades permite agruparlas en clústeres y calificar estos clústeres según su conservación dado un grupo de bacterias. La definición de vecindad y similitud de vecindad está descrita posteriormente en los métodos. El algoritmo fue aplicado a la familia IlvC en 290 *Streptomyces* resultando 9 clústeres Datos entre los más poblados el primero cuenta con 279 elementos, otro con 9 elemento y dos más con 7 miembros, resultados experimentales son congruentes con que existe divergencia funcional entre miembros de clústeres distintos [Verdel-Aranda, 2015]

Estudio de la familia PriA

Caracterización *in vivo*

No todas las familias promiscuas provienen de expansiones, tal es el caso de PriA en Actinobacteria, donde no tiene expansiones y hasta el momento no se le conoce participación en rutas de metabolismo especializado. La promiscuidad de PriA parece debida al rescate de la función TrpF. Algunas enzimas PriA no han mostrado promiscuidad *in vitro,* pero si *in vivo* ya que sobreviven en un medio sin triptófano, es decir *in vivo* complementan la función TrpF.

Caracterización bioquímica *in vitro.*

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

Modelado de dinámica molecular

La dinámica es un método que permite hacer simulaciones de partículas que sirve para obtener información de propiedades macroscópicas de un conjunto de átomos [Petrenko, 2001; Kukol, 2008]. Es útil en el marco de mi proyecto porque permite la exploración del espacio conformacional, y se ha visto que este está relacionado con la actividad de la enzima [Sikosek, 2014], además dado un confórmero permite verificar su estabilidad. Resuelve la ecuación de movimiento de Newton con base a una configuración inicial, las fuerzas interatómicas como los enlaces covalentes, las fuerzas de Van der Waals y la carga de las partículas [Campbell, 2012]. Entonces para generar una simulación de dinámica molecular, debe contarse con una estructura como punto de partida, ya sea esta cristalográfica o modelada de novo o por homología. El laboratorio de bioinformática y biofísica

computacional ha desarrollado un protocolo de generación de modelos homólogos estructurales y dinámicas moleculares; con este pipeline se han generado dos estructuras de Camera [Noda-García, 2015], 30 estructuras y dinámicas de miembros de Actinobacteriaceae y Bifidobacteriaceae [Juárez-Vázquez, 2017] y finalmente una estructura de subHisA de *Corynebacterium diphteriae*. En el género *Streptomyces*, interesante debido a su variación en contexto genómico y en mediciones *in vitro* aún no se modelan dinámicas moleculares ni se habían realizado estructuras por homología previo a este trabajo.

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

Objetivos

Objetivo General

Estudiar el fenómeno de promiscuidad tanto desarrollando estrategias para identificar familias promiscuas, miembros promiscuos en dichas familias o clústeres biosintéticos promiscuos dentro de un grupo taxonómico, como comparando variaciones de promiscuidad *in vitro* con variaciones en contexto genómico y en propiedades relacionadas a la estructura tridimensional en miembros de una familia.

Objetivos particulares

1. Entender cuáles son los genes más conservados en el *core genome* en un linaje taxonómico y aplicar este conocimiento.
2. Identificar de familias enzimáticas con cambios en vecindades genómicas como características informativas provenientes de datos filogenómicos.
3. Entender la evolución de clústeres biosintéticos y vecindades genómicas diversas, para poder organizarlas jerárquicamente tal y como se pueden organizar las secuencias de enzimas. Con este conocimiento proponer y un BGC promiscuo y verificar la hipótesis de rutas promiscuas
4. Estudiar la relación entre historias filogenómicas y procesos biofísicos con la promiscuidad *in vitro*, a través de mediciones de ciertas características de la familia PriA.

Estrategias

Obtener información genómica de diversos linajes genómicos.

Colectar genomas de Actinobacteria, Cianobacteria, *Pseudomonas* y Archaea de NCBI y de colecciones privadas.

Anotar consistentemente las secuencias codificantes de estos genomas.

Utilizar un anotador automatizado y desarrollar los scripts necesarios para automatizar la anotación de los genomas.

Establecer las relaciones filogenéticas de los genomas colectados.

Mediante el uso del *core genome* construir un árbol filogenómico que permita establecer un marco sobre el cual hablar de cambio.

Establecer las relaciones filogenéticas de los genomas colectados.

Desarrollar Orthocore. Proponer una métrica para identificar los ortólogos más conservados de una familia génica.

La promiscuidad en familias enzimáticas.

Desarrollar EvoMining hasta el punto de convertirla en una distribución funcional de código libre disponible para otros investigadores. Agregar características que faciliten la investigación como la visualización de la vecindad genómica de los miembros de una familia y otros metadatos como el número de copias por organismo.

Identificar cambios en la vecindad genómica en familias selectas de enzimas de metabolismo central.

Identificar familias de metabolismo central en los linajes genómicos seleccionados.

Identificar mediante EvoMining las familias de metabolismo central con expansiones y reclutamientos que pueden presentar promiscuidad enzimática.

Desarrollar las herramientas bioinformáticas necesarias para entender la conservación y evolución de BGC u otras vecindades genómicas a las que pertenecen las expansiones de las familias centrales previamente seleccionadas.

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

Dados los sustratos conocidos de PriA investigar las posibles correlaciones entre mediciones de constantes catalíticas, contexto genómico, vecindad genómica, estructura tridimensional y cambio en el número de copias en diversos linajes.

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

-Ampliar el contenido de EvoMining para poder analizar nuevas bases de datos genómicas, agregando por ejemplo los nuevos genomas colectados de Actinobacteria, así como los genomas de Cianobacteria, $Pseudomonas$ y Archaea.

-Sistematizar la base de datos de metabolismo central.

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

Seleccionar miembros homólogos de la familia de enzimas para modelado molecular.

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

Medir cinéticas enzimáticas, contexto genómico y vecindad genómica

Estudiar la existencia de distintas vecindades genómicas. Determinar la cinética enzimática de PriA respecto a sustratos atípicos sugeridos por los estudios estructurales. Obtener

mediante una colaboración 37 modelos estructurales por homología y modelar docking molecular.

Metodología

A continuación, describiré la metodología para cada una de las estrategias expuestas previamente. Todos los scripts desarrollados fueron escritos en Perl y están disponibles en GitHub https://github.com/nselem/perlas y empacados como un contenedor de Docker.

La promiscuidad en familias enzimáticas.

Actinobacteria genómica

Para obtener información genómica del phylum Actinobacteria mediante la colección de genomas de NCBI se revisaron todas las familias de Actinobacteria de la base de NCBI "*genome"* y se seleccionaron los genomas con mínimo 5 genes por contig. Se crearon scripts para utilizar la interfaz e-utils de NCBI y descargar estos genomas desde la terminal a partir de una lista de identificadores. El mismo procedimiento se aplicó a Cianobacteria, *Pseudomonas* y Archaea.

Anotación

Para anotar consistentemente las secuencias codificantes de estos genomas se utilizó el anotador automatizado RAST y se desarrollaron los scripts necesarios para anotar los genomas desde la terminal, conectado así NCBI y RAST. Se empacaron estos scripts en una distribución Docker de myRAST.

Filogenia de la base de datos genómica

Establecer las relaciones filogenéticas de los genomas colectados mediante el uso del *core genome* para construir un árbol filogenómico, y para encontrar genes marcadores de un linaje taxonómico. Para obtener el *core genome* y en base a ello reclasificar los genomas se diseñó el algoritmo Orthocore que generaliza el concepto de Best Bidirectional Hits (blast all vs all).

Orthocore. Se realiza un blast all vs all de genomas deseados. Para cada secuencia, centrado en cada genoma se realiza una lista (estrella) de sus mejores hits bidireccionales.

Si las listas de todos los genomas coinciden es un BBH múltiple y se agrega la lista al core genome. Una vez con el *core genome* completo se puede reconstruir la filogenia. Este método fue exitoso en la detección de una familia marcadora de *Clavibacter michiganensis*.

Organización y presentación de familias extendidas mediante el desarrollo una plataforma bioinformática.

Se desarrolló EvoMining como una plataforma interactiva de consulta, se complementa con programas de visualización de árboles filogenéticos y contextos genómicos.

Algoritmo de reconstrucción filogenética y visualización

Las secuencias de las familias enzimáticas expandidas fueron alineadas con MUSCLE v3.2 [Edgar, 2004] y curadas con Gblocks v0.91b [Castresana, 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. 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, 2010]. La anotación funcional de RAST en su versión clásica [Aziz, 2008; Overbeek, 2014] fue agregada al SVG.

Algoritmo de visualización

Para facilitar el análisis visual de una vecindad genómica y a la vez generar imágenes de alta calidad exportables para su uso en publicaciones, se desarrolló código cuya salida es el formato *Scalable Vector Graphics* (SVG). Este formato es básicamente un archivo de texto XML que contiene instrucciones para que navegadores como Chrome o Firefox realicen un dibujo. Al ser vectores, las imágenes generadas en SVG no pierden resolución al ser escaladas y justamente por ser escalables permiten explorar con detalle grandes cantidades de datos organizados por ejemplo en árboles filogenéticos. Los scripts extraen para cada gen información necesaria como coordenadas, dirección, función química. La anotación de función proviene de la anotación de RAST, la visualización de vecindades genómicas fue desarrollada *de novo*. Los primeros análisis de EvoMining, antes de la existencia de esta plataforma fueron desarrollados en el lenguaje Perl; este lenguaje cuenta con un módulo para facilitar la elaboración de SVG (perlmaven/SVG 2015) por lo que se decidió utilizarlo y

continuar el desarrollo en Perl para no agregar nuevos requerimientos y facilitar su portabilidad.

Se amplificará EvoMining de los 200 genomas con que contaba su versión inicial a los 1200 colectados en Actinobacteria, además, se incluirán otros linajes genómicos. Se transformará la curación manual de su base de datos de rutas centrales a la anotación por subsistemas de RAST. Finalmente se presentará la variación en vecindades genómicos como una herramienta adicional que ayude en la búsqueda de promiscuidad en familias de enzimas pertenecientes al metabolismo central.

Desarrollo de bio contenedores

Para desarrollar herramientas bioinformáticas relacionadas al estudio de la promiscuidad se adoptó el enfoque de los contenedores bioinformáticos, todo el código fue depositado y documentado en GitHub y distribuido a través de un contenedor Docker siguiendo las especificaciones generales de desarrollo de contenedores [Schulz, 2016; Gruening, 2018].

Identificar cambios en la vecindad genómica en familias selectas de enzimas de metabolismo central.

Se definirá como familia de vecindades genómica aquellas que compartan una enzima query y al menos un gen con el clúster de referencia. Los miembros de esta familia serán identificados y organizados filogenéticamente en una visualización.

Dinámica molecular

Para generar dinámicas moleculares en primer lugar se recolectarán las estructuras tridimensionales de miembros de PriA de Actinobacteria. Después se procederá a modelar por homología las estructuras tridimensionales faltantes utilizando el pipeline del laboratorio de bioinformática y biofísica computacional. Este pipeline utiliza el software Rosetta para el modelado para las estructuras y GROMACS Groningen Machine for Chemical Simulation, [Van-der-Spoel, 2005] para el modelado de la dinámica molecular. Esta parte del trabajo se realizará en colaboración con el laboratorio de bioinformática y biofísica computacional.

Consideraciones

Falsos negativos respecto a promiscuidad están muy extendidos en la literatura y en las bases de datos, en parte porque la mayoría de las funciones son asignadas por similitud de secuencia y dado un falso negativo el error se propaga en secuencias similares. Por otro lado, es muy difícil demostrar un verdadero negativo a menos que se prueben todas las posibilidades de sustrato para la enzima. Sin embargo, el espacio de sustratos puede acotarse gracias a técnicas como el docking que está íntimamente relacionado con la dinámica molecular [Campbell, 2012; Kukol, 2008]. Limitar el espacio de sustratos puede retroalimentarse con el estudio de la promiscuidad *in vivo* y viceversa.

Con los métodos propuestos en este trabajo sólo se podrá detectar pérdida o ganancia de promiscuidad entre enzimas de organismos respecto a otros miembros dentro un grupo taxonómico, no así el estado de promiscuidad intrínseco a la enzima. Si dada una enzima no se detectan variaciones en contexto, vecindad genética o flexibilidad dentro de un grupo taxonómico cercano, entonces no podemos decir en principio nada acerca de la promiscuidad de la variante, posiblemente es promiscua, pero al mantenerse constante en todos los parámetros descritos, con estos métodos no se puede sugerir promiscuidad. Es posible que al mirar en un grupo taxonómico más amplio se detecte una neofuncionalización de la familia, aunque también es posible que exista una variable z como la flexibilidad de sustrato [Nobeli, 2009; Odokonyero, 2014] que no se esté considerando y que explique o sea el mejor indicador para esta familia de promiscuidad enzimática.

Se debe considerar que si existe una correlación vecindad genómica-promiscuidad, esta no indica causa efecto, más bien, es plausible que la vecindad sea un amplificación de diferencias en secuencia, a un número igual de variaciones en secuencia la existencia de un cambio de vecindad indica un proceso más largo y más cambios, es una amplificación de las marcas dejadas por transformaciones funcionales.

Si bien no se resuelve el problema de anotar promiscuidad automáticamente, este trabajo pretende aprovechar que los contextos genómicos ayudan a la identificación de familias promiscuas para mejorar una plataforma de productos naturales, pretende también una confirmación de que los cambios en la dinámica molecular ayudan a identificar los miembros

más promiscuos hacia actividades recién adquiridas, así como también ser pionero en la investigación de promiscuidad *in vivo*.

Capítulo 1

Orthocore: una herramienta computacional para entender el pangenoma de un linaje genómico.

La organización filogenética de un linaje genómico permite la observación de dinámicas de pérdida y ganancia de familias génicas en organismos cercanos. Si los organismos están desordenados taxonómicamente es difícil apreciar la dinámica de aparición-desaparición por genoma de ortólogos que son miembros de una familia génica. Así pues, ordenar los genomas de un linaje de acuerdo con su historia evolutiva facilita apreciar cambios en el número de copias de una familia. Esta consideración es relevante en el marco de esta tesis ya que cambios en la ocurrencia de promiscuidad pueden estar relacionados a copias extra de organismos cercanos [Verdel-Aranda, 2015]. Como se explicó previamente en la introducción, el pangenoma de un linaje taxonómico está dividido en el *core*, *shell* y *dispensable genome*. Orthocore es un algoritmo que se desarrolló para automatizar la obtención del *core conservado* de un linaje genómico. Con este *core* realiza una reconstrucción filogenética de los organismos del linaje (Figura 1.1). En esta sección se explica el funcionamiento de Orthocore, así como casos en donde fue utilizado, específicamente en casos de estudio de los órdenes *Actinomycetales* [Juarez-Vazquez, 2017], *Micrococcales* [Rodriguez, 2016] y *Nostocales* [Gutierrez-Garcia, 2019] y en el género *Salmonella* [Delgado-Suarez, 2018]. Finalmente se describen otros métodos implementados para entender el pangenoma de un linaje genómico que han sido utilizados en aplicaciones relacionadas al patógeno del tomate *Clavibacter michiganensis* [Rodriguez, 2016].

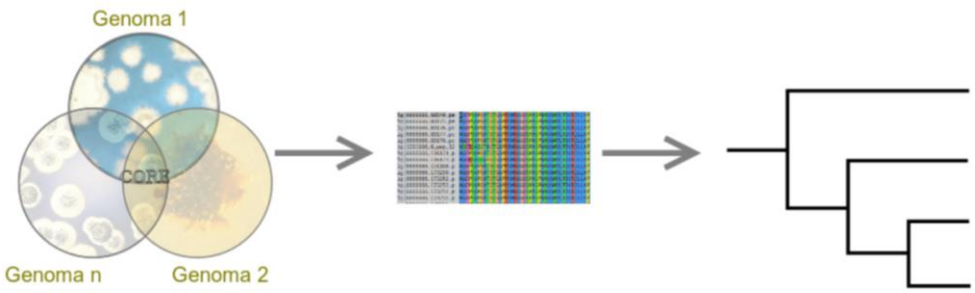

Figura 1.1 Orthocore calcula las familias génicas comunes de un linaje taxonómico. Después de un proceso de filtrado, alineamiento y curación, Orthocore concatena las secuencias de estas familias y entrega una reconstrucción filogenética.

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

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

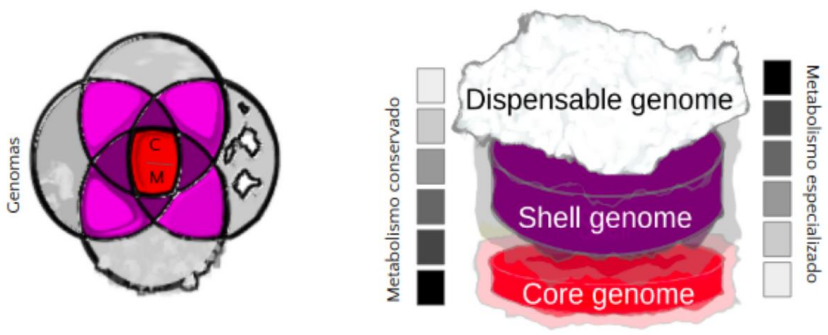

Figura 1.2 El pangenoma de un conjunto de genomas de un linaje puede ser clasificado en varios grupos. En este ejemplo, en el lado izquierdo de la figura se observa en gris el genoma dispensable compuesto por familias génicas presentes sólo en un genoma. En dos tonos de morado observamos el *shell genome*,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

El *core genome* de un linaje se puede usar para encontrar familias marcadoras (Figure 1.3). Estos genes marcadores permiten realizar pruebas que diagnostican la presencia de organismos de ese linaje. A las familias que están presentes en el *core genome* de un linaje A, pero que están completamente ausentes de un linaje B se les llama marcadoras. Por ejemplo, genes conservados en la especie *Streptomyces coelicolor,* pero ausentes en *Streptomyces rimosus* son genes marcadores de *Streptomyces coelicolor* respecto de *Streptomyces rimosus*. Estos mismos marcadores tal vez no sean marcadores respecto de *Streptomyces lividans*, a pesar de la cercanía taxonómica entre estos organismos. La presencia de genes marcadores en el *core* depende de ambos linajes, y ya que es de interés científico y tecnológico encontrarlos en para identificar diferentes taxones es importante contar con algoritmos que permitan encontrarlos de forma automatizada.

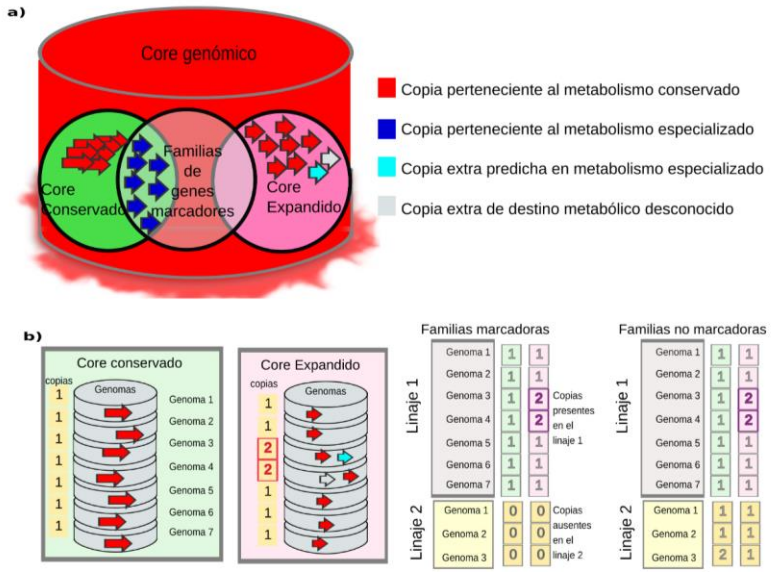

Figura 1.3 El *core genome* puede contener familias con funciones en distintos grupos metabólicos, así como diversidad en el número de copias. Arriba se muestra que en el *core* pueden coexistir familias tanto de copia única como con expansiones. Las familias con funciones en el metabolismo conservado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

El número de familias en el pangenoma ya sea en el *core*, *shell* o *dispensable genome* no sólo depende de la divergencia o proximidad taxonómica de los organismos del linaje seleccionado, también depende de lo variable que sea el contenido génico en los genomas del linaje. A esta característica se le conoce como apertura. Hay especies, por ejemplo, algunos patógenos, cuyo pangenoma se encuentra sumamente cerrado en el sentido de que no importa cuántos genomas se

agreguen, el número de familias parece converger y ser asintótico rápidamente a una cota superior. En cambio, especies o géneros que viven en una gran diversidad de hábitats suelen tener un pangenoma abierto. Esto significa que cada vez que se agrega un nuevo genoma aparecen otras familias que no estaban en los genomas anteriores. En los linajes con pangenoma abierto el número de familias nuevas al agregar un genoma seguirá una tendencia creciente y no asintótica. Además de la apertura, existen otros intentos de cuantificar la diversidad genética de un linaje. Está por ejemplo la fluidez, definida como el promedio de familias únicas entre familias totales por pares de genomas. El pangenoma bacteriano total, es decir el total de familias génicas en el dominio Bacteria es considerado abierto.

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

1.2. El core conservado permite la reconstrucción de filogenias complicadas

Orthocore es el desarrollo bioinformático que realicé para calcular las familias génicas más conservadas del *core genome*. Dos genes son homólogos si poseen un ancestro común, entre los principales grupos de homólogos están ortólogos y parálogos. Los ortólogos provienen de eventos de especiación de un ancestro común mientras que los parálogos evolucionan por eventos de

duplicación. Orthocore obtiene un subconjunto del *core genome*: el *core conservado*, es decir, familias de ortólogos presentes en todos los genomas del grupo y que además son libres de parálogos de difícil identificación. El *core conservado* facilita la organización en árboles filogenéticos de organismos de un linaje genómico.

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

Entre los factores importantes para establecer las relaciones filogenéticas que diferencian a las Archaeas de las Bacteria están los siguientes: 1) la presencia conservada de la subunidad de 16S en los dos dominios y 2) la suficiente divergencia entre estas secuencias en los organismos de dichos dominios. Ahora bien, establecer relaciones filogenéticas entre Archaea y Bacteria es en cierto sentido más sencillo que establecerlas entre organismos pertenecientes al mismo género o inclusive a la misma especie. En ocasiones, como en el caso del género *Streptomyces*, la secuencia de 16S rRNA por sí sola no posee la suficiente variación para resolver la filogenia [Labeda, 2017]. En *Streptomyces* la variación entre estas secuencias suele ser menor al 1%. Para resolver el problema de escasa variación en secuencias de 16S rRNA se pueden concatenar las secuencias de otros ortólogos, siempre que estos aparezcan en todos los organismos que se estén estudiando, es decir, siempre que sean parte del *core* genómico.

1.3. El algoritmo de Orthocore

Orthocore encuentra los genes del *core* conservado en todos los genomas que se le proveen. Para poder ejecutarlo, el usuario tiene que proveer como entrada un conjunto de secuencias genómicas

que ya debieron haber sido procesados por RAST, es decir ya deben tener anotaciones funcionales de los genes presentes en dichos genomas. El algoritmo de Orthocore encuentra los genes que están conservados como ortólogos en todas las secuencias y proporciona como salida una matriz con todos los genes y secuencias del *core,* es decir, todos los genes que resultaron mejores hits multidireccionales (Figura 1.4). Una vez que tiene todas las secuencias realiza una reconstrucción filogenética con las secuencias de aminoácidos y entrega la segunda salida, que es un árbol en formato árbol en formato Newick. Con estas salidas se puede visualizar la filogenia de los organismos en cuestión, identificar cuáles genes son parte del *core*. Una parte central de este algoritmo consiste en determinar cuáles genes realmente son ortólogos confiables para reconstruir la filogenia.

1.3.1. Los mejores hits multidireccionales definen los genes del *core* conservado

Los ortólogos suelen identificarse por similitud de secuencia, pero si se realiza la identificación manualmente también se suelen capturar parálogos que pueden confundir la elucidación de eventos de especiación. Orthocore automatizó la búsqueda de ortólogos y el filtrado de parálogos en genomas procariontes mediante la generalización de la definición del mejor hit bidireccional (BBH por sus siglas en inglés). Dos secuencias son BBH si cada una es el mejor hit de un algoritmo de distancia (BLAST usualmente) en el genoma de origen de la otra. Una primera generalización para obtener el set de ortólogos de una familia del *core* es definir un genoma de referencia y tomar los BBH respecto a ese genoma. En la práctica, esta definición da como resultado distintos resultados según el genoma de referencia, haciendo que algunos parálogos no sean filtrados.

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

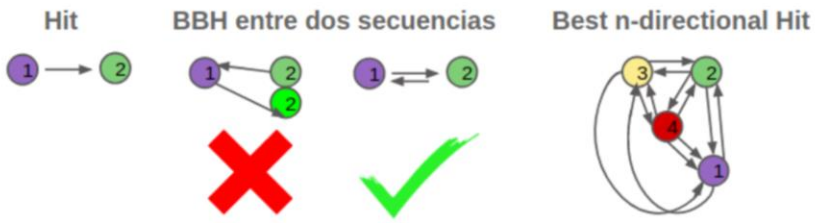

Figura 1.4 Orthocore utiliza los mejores hits n-direccionales para obtener grupos de ortólogos. Un hit es el mejor resultado de una secuencia en otro genoma. Un *Bidirectional Best Hit* (BBH) es el mejor hit bidireccional. La secuencia 2 es el mejor hit de la secuencia 1 en el genoma 2 y recíprocamente,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

1.3.2. Ejecución de Orthocore

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

- Obtiene el *core conservado*: los mejores hits multidireccionales (Blastp).
- Alinea cada familia del *core conservado* (MUSCLE).
- Cura automáticamente cada familia del *core conservado* (Gblocks).
- Concatena las familias del *core conservado* formando un alineamiento multilocus de aminoácidos.
- Provee una reconstrucción filogenética que se genera a partir del alineamiento (FastTree).
- Provee los genes del *core conservado* y su anotación funcional según RAST.

Existen otros algoritmos como OrthoMCL [Li, 2003], y FastOrtho [Wattam, 2014] que dividen pangenomas en clústeres de familias de genes, get_homologues [Contreras-Moreira, 2013] y Metaphor que obtienen el *core* y filtran buscando verdaderas relaciones de homología, y finalmente BPGA [Chaudhari, 2016] que hace reconstrucciones filogenéticas tanto según el *core* como según el pangenoma. Sin embargo, Orthocore resolvió en su momento la necesidad específica de proporcionar un alineamiento concatenada de genes del *core conservado* lista para utilizarse en un árbol multilocus. Adicionalmente, como Orthocore fue diseñado para trabajar con la anotación de la

plataforma RAST, también se obtiene la anotación funcional tanto de familias del *core* como del complemento.

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

1.4. Aplicaciones de Orthocore, identificación del *core conservado* y de familias de genes marcadores.

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

1.4.1. Orthocore de *Actinomycetales* para mejorar el árbol de especies

A pesar de que se sabía que miembros de la familia PriA habían sufrido cambios de promiscuidad y de especificidad entre la función HisA de la vía de síntesis de histidina y la función TrpF de la biosíntesis de triptófano [Noda-Garcia, 2013], se desconocía cómo había ido cambiando la función de esta familia de genes durante los procesos de especiación en *Actinomycetales*. Para eso se necesitaba entender filogenéticamente al orden *Actinomycetales*, el camino que seguimos fue hacer una filogenia con las secuencias de su *core conservado*. Orthocore fue diseñado para resolver este problema. Con el resultado de Orthocore se realizó un árbol de especies donde se observaron patrones de pérdida y ganancia de genes en la vecindad genómica del gen que codifica para PriA. Se encontró que hay clados de *Actinomyces* donde los genes correspondientes a la síntesis de histidina no estaban en la vecindad genómica de PriA, y mediante la realización de cinéticas enzimáticas se comprobó que la actividad de catalizar la reacción correspondiente a HisA estaba

perdida en estos organismos. A estas enzimas se les llamó subTrpF ya que sólo poseían la capacidad de catalizar la reacción correspondiente a la familia TrpF. Del mismo modo existían clados que perdieron los genes de síntesis de triptófano en la vecindad de PriA y estas enzimas se subfuncionalizaron a la familia subHisA. De estos datos se observa que en estos organismos la especiación coincidió con el cambio de promiscuidad en la familia PriA, acorde a la pérdida y ganancia de genes vecinos. Este caso es una muestra promiscuidad puede coocurrir con variaciones en el contexto genómico, pudiendo estos cambios ser una marca para sugerir cambio funcional en una familia. De esta forma, se pudo concluir que los perfiles de promiscuidad de PriA en el orden *Actinomycetales* se relacionan con la especiación [Juarez-Vazquez, 2017].

1.4.2. Uso de Orthocore para entender la evolución de la patogénesis de *Salmonella* en México.

Los genomas conocidos de *Salmonella* en México están tan conservados en sus genes marcadores tradicionales que se dificulta hacer una filogenia. Esta situación es común en cepas de la misma especie. En este caso era importante relacionar las especiaciones de cepas aparentemente no patogénicas y la evolución de las islas de patogenicidad que se sabía estaban presentes en la mayoría de los aislados.
Por ello, como primer paso se desarrolló una distribución de myRAST en un contenedor de Docker que pudiera usarse en cualquier servidor. Esta tubería prepara los datos para Orthocore al realizar la anotación automática de genomas ensamblados en RAST. El protocolo de myRAST está disponible en GitHub y además puede encontrarse en los apéndices de esta tesis. Después, con los genomas ya anotados, Orthocore fue usado para reconstruir la filogenia de estos aislados de *Salmonella*. Además, se buscó cómo fue la evolución de las islas de patogenicidad de estas cepas mediante un análisis de CORASON [Navarro-Munoz, 2018], en el que visualizamos las islas de patogenicidad organizadas filogenéticamente. En este caso se observó una alta conservación de toxinas tifoidales en islas de patogenicidad de *Salmonella*. Éstas fueron identificadas en 76% de las cepas analizadas. Este análisis se publicó como parte de un trabajo donde además se mostró que las islas de patogenicidad no pueden ser marcadores de la enfermedad ya que se encontró ganado sin síntomas de la enfermedad pero que sí contenía bacterias con estas islas [Delgado-Suarez, 2018].

1.4.3. Las bacterias *Nostoc* provenientes del metagenoma de cícadas están en el mismo grupo filogenético

Cianobacteria es un phylum de bacterias que se han adaptado a diversos ambientes. Aunque muchas de ellas son marinas algunas Cianobacterias viven como simbiontes de plantas. En particular las cícadas han desarrollado un tipo especial de raíz donde se sabe que vive como simbionte el género *Nostoc*. La presencia de *Nostoc* en la raíz coraloide de las cícadas es fácilmente distinguible por la formación de un anillo verde conocido como anillo Cianobacterial. En la Figura 1.5 se muestra la filogenia de 76 Cianobacterias de 7 órdenes distintos construida con 198 proteínas del *core conservado* obtenidas por Orthocore. En esta reconstrucción se puede observar que los *Nostoc* asociados a plantas tienden a agruparse en la filogenia [Gutierrez-Garcia, 2019] lo que sugiere que entre las Cianobacterias de nuestro set de datos pudiera ser que las 3 *Nostoc* provenientes de raíz coraloide provengan de un evento de colonización. Queda por mostrar si esta filogenia corresponde a la de las cícadas simbiontes, lo que sugeriría una estrecha coevolución entre ambos organismos. Identificación de genes marcadores de la bacteria del 'cáncer del tomate' *Clavibacter michiganensis*

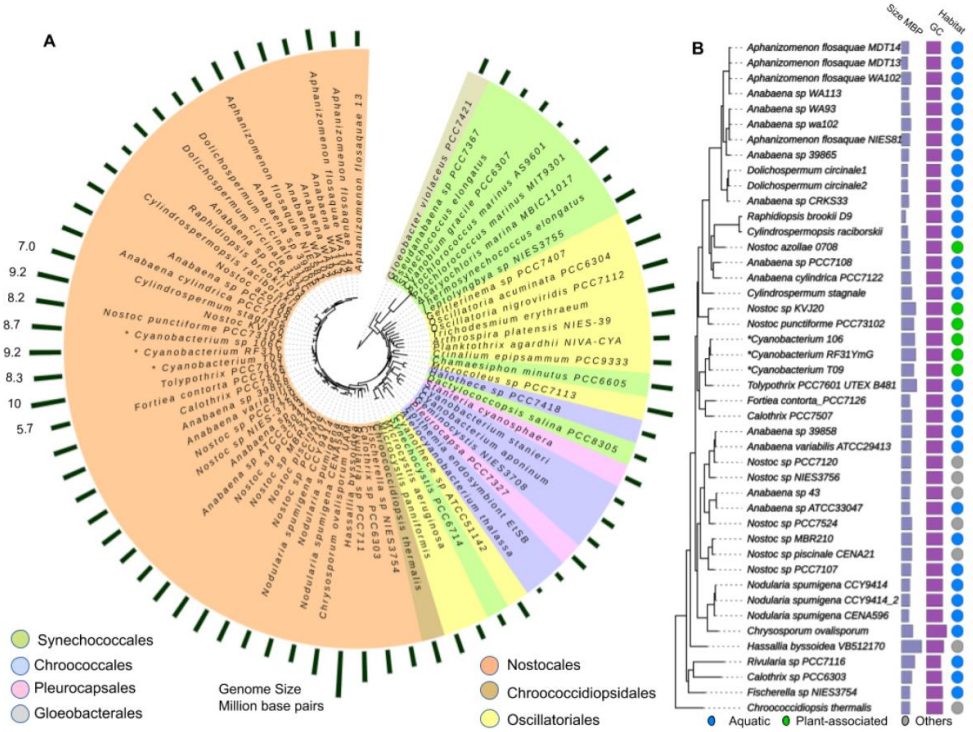

Figura 1.5 Reconstrucción de 76 taxa provenientes de 7 órdenes de Cianobacteria. La matriz final incluyó 45,475 aminoácidos curados de 198 familias de proteínas pertenecientes al *core* conservado. A la derecha se muestra un acercamiento sobre el orden Nostocales. En este orden se incluyen algunas bacterias simbiontes de cícadas. Metadatos como el tamaño de genoma, contenido de GC y hábitat de

origen muestran una posible tendencia de incremento de tamaño en los genomas provenientes del microbioma de plantas.

1.4.4. Identificación de genes marcadores de la bacteria del 'cáncer del tomate' *Clavibacter michiganensis*.

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

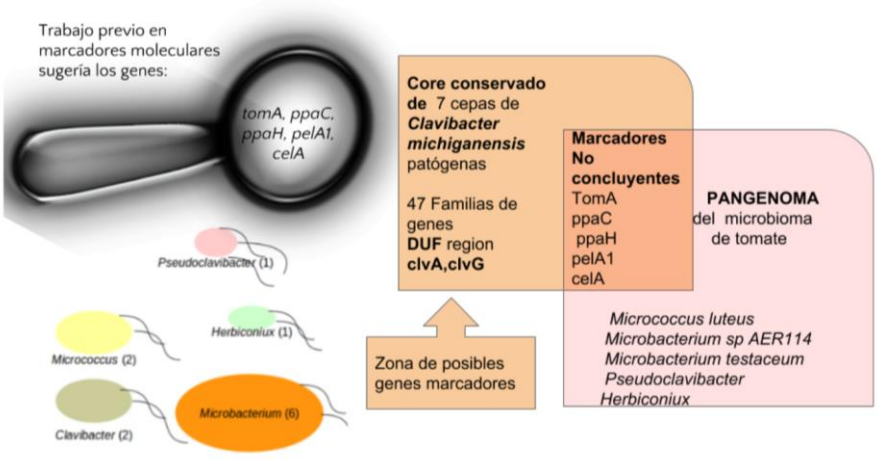

Figura 1.6 Los marcadores moleculares previos a este trabajo no permitían diferenciar correctamente a *Clavibacter michiganensis* respecto al microbioma del tomate en invernaderos mexicanos. Arriba a la izquierda se muestran los antiguos marcadores *tomA*, *ppaC*, *ppaH*, *pel*A1, *celA*. Abajo organismos pertenecientes al mcrobioma del tomate. Orthocore obtuvo el *core* conservado de siete cepas de *Cmm* patógenas, y este *core* se utilizó para definir nuevos marcadores. Aunque *tomA* pertenecía al *core* conservado de *Cmm*, estaba también incluido en el pangenoma del microbioma del tomate. Después de calcular la intersección *core* conservado de *Cmm* y pangenoma del microbioma se obtuvieron entre las familias marcadoras genes *clv* parte del *cluster* biosintético de clavidicina (michiganina).

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

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

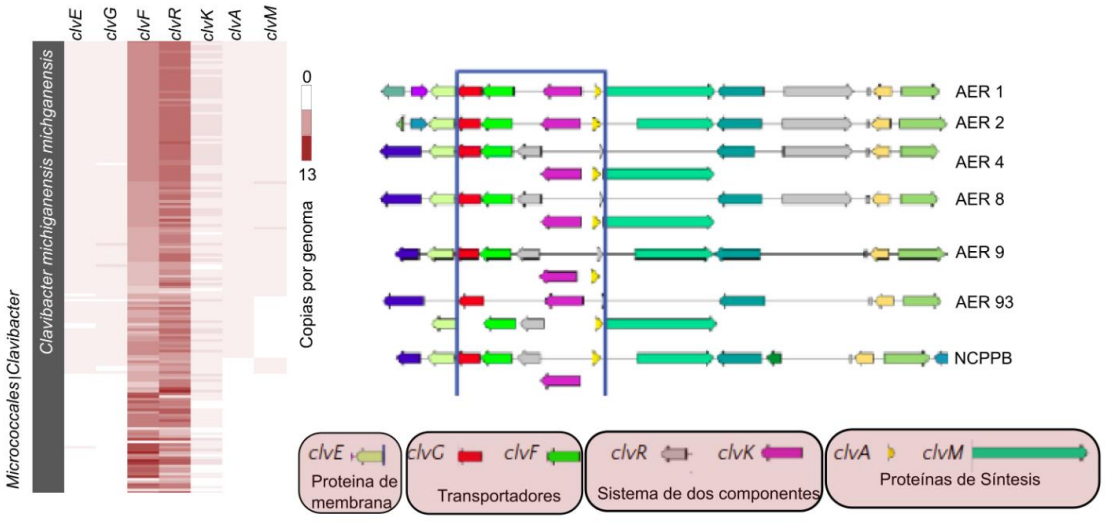

Figura 1.7 El Cluster *clv* es un marcador de la especie *Clavibacter michiganesis* en el orden *Micrococcales*. Los genes *clvEGAM* que son una proteína de membrana, un transportador y dos proteínas de síntesis del BGC *clv* están muy conservados en la especie *Clavibacter michiganensis*. *clvFRK*,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

De hecho, al considerar más genomas dentro del microbioma del tomate, la familia *clvF* no solo no está en el *core conservado* de *Microbacterium* y *Micrococcus*, sino que no está presente en ningún otro genoma distinto a *Clavibacter*. Con distintos niveles de conservación de secuencia los genes *clv* son un buen marcador para distinguir *Cmm* de otras especies, por esta razón estos genes aún se encuentran en uso como genes marcadores de *Cmm*. Previamente se reportó que las familias *clvA, clvF* y *clvG* son exclusivas de *Cmm*, de acuerdo con evidencia experimental [Yasuhara-Bell, 2014] pero sin análisis genómicos que confirmaran que esos genes no estaban presentes en otras especies sin ser expresados. Este descubrimiento ha permitido bajar los costos de identificación de *Clavibacter*, ya que ahora en lugar de enviar a secuenciar el genoma es suficiente identificar por PCR *clvF* en conjunto con otros marcadores.

Con Orthocore además de obtener los genes marcadores podemos obtener también la matriz del *core conservado* para realizar la reconstrucción filogenética de especies cercanas de *Clavibacter* (Figura 1.8). Ya que los productores querían conocer de dónde provienen las bacterias que infectan al tomate con el fin de evitarlas, se determinó la organización taxonómica de cepas de *Cmm* y de otras bacterias del microbioma de tomate. Para esto se han secuenciado y mantenido como datos privados unos doscientos genomas provenientes del microbioma del tomate y se han analizado con Orthocore para obtener matrices multilocus que pueden diferenciar entre cepas de *Clavibacter* de la misma o de diferente especie.

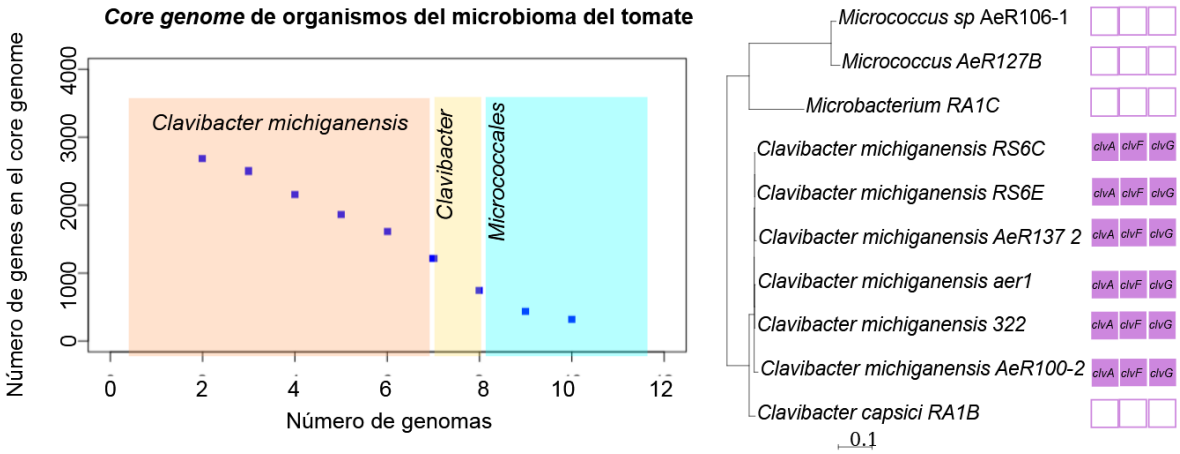

Figura 1.8 El *core genome* de *micrococcales* ayuda a identificar genes marcadores. El *core genome* de siete organismos del microbioma del tomate muestra como decrece este conjunto al agregar un *Clavibacter* que no es de la especie *michiganensis*, y como cae aún más al incluir en el análisis otros *micrococcales*. A la derecha se muestran los genes marcadores, que están en el core de los genomas de *C. michiganensis*, y simultáneamente no pertenecen a ningún organismo fuera de esta especie.

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

1.4.4.1. Clavisual: Identificación de genes marcadores a un cierto porcentaje de grupos seleccionados

La idea de que Orthocore puede ser usado para obtener los genes marcadores de un grupo taxonómico frente a otro fue generalizada en el software Clavisual. Ya se ha explicado previamente que el *core* puede salir vacío por diversas razones, entre ellas baja calidad de los genomas, o que éstos provengan de organismos muy divergentes, verdaderas razones biológicas como dinámica génica o un *core* no convergente. Así pues, es posible que si sólo se utiliza el *core* no se obtengan marcadores. Pero el *core* puede relajarse de varias maneras una de ellas es el *pseudocore*, donde en lugar de multidireccionales hits se toman BBH a un genoma de referencia. Otra forma es establecer un porcentaje de presencia /ausencia de interés. El *pseudocore* consiste en utilizar información previa específica del *core* de un linaje para obtener los genes de y la metodología está depositada en GitHub integrada en el repositorio clavigenomics. El blast fue optimizado cambiando hacer un blast todos contra todos por archivos genómicos individuales genomai_vs_genomaj.blast que luego son concatenados según se necesiten.

Los porcentajes de genomas son diferentes porque al no bastar los mejores hits bidireccionales conservados, todo el pangenoma es decir todos los genes contenidos en los genomas del grupo de interés necesitan ser clasificados por familias, para de ahí obtener las familias que tienen presencia en un porcentaje %p y ausencia en un porcentaje a% del grupo externo. Estos perfiles fueron

desarrollados para Clavisual (Figura 1.9) utilizando FastOrtho para clasificar las familias y de ahí obtener los grupos. Con ellos se consiguieron marcadores para *Kurtobacterium*.

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

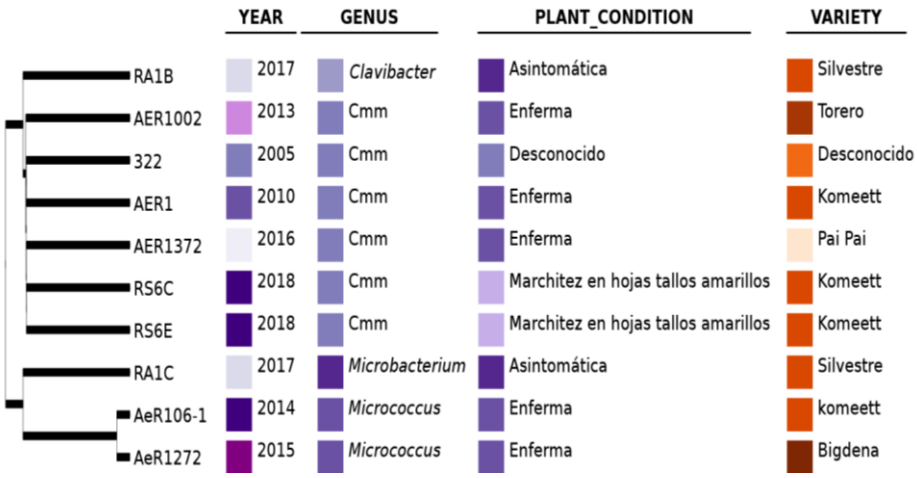

Figura 1.9 Clavisual organiza filogenéticamente cepas de *Clavibacter* y generaliza Orthocore. Clavisual no se restringe a la búsqueda del *core* conservado, permite la identificación de familias conservadas sólo en un cierto porcentaje. Además, puede filtrar estas familias permitiendo sólo las que estén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

1.4.4.2. El pangenoma de *Clavibacter michiganensis* es abierto

Después de desarrollar métodos de identificación de genes marcadores y generalizarlo a obtener grupos con patrones de presencia/ausencia definidos por el usuario, quedaba por responder la pregunta cómo es el pangenoma de *Cmm*. Algunos autores consideran que el pangenoma de patógenos es reducido porque sus genomas suelen sufrir proceso de reducción de tamaño debido a la pérdida de genes. Como *Cmm* es un patógeno de planta quedaba por investigar cómo es su

pangenoma. ¿Es posible saturar el contenido génico de *Cmm* con sólo secuenciar más genomas? Aunque actualmente existen ya herramientas web para el análisis de pangenoma, en su momento se utilizó el software *Bacterial PanGenome Analysis Tool* que se corre desde la terminal. Para facilitar su instalación se desarrolló un contenedor Docker (ver abajo, descripciones técnicas). Como ejemplo de su funcionamiento, se analizó el pangenoma de los mismos diez genomas del tomate utilizados en la visualización de Clavisual (Figura 1.10). Tomando otros siete genomas del género *Clavibacter*, utilizando OrthoVenn tenemos la misma observación, el número de familias de genes agregadas al adicionar genomas, es después de siete genomas casi tan grande como su *core* (Figura 1.11).

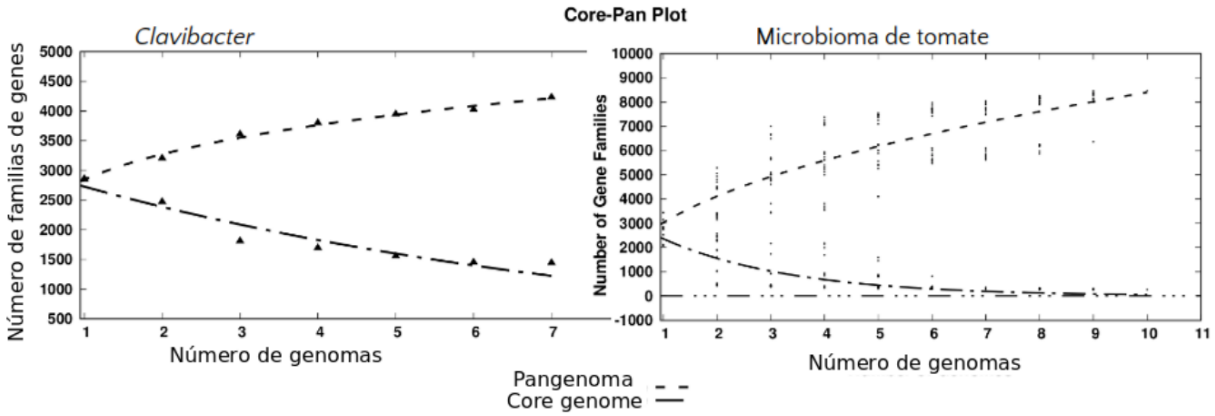

Figura 1.10 El pangenoma de *Clavibacter* es abierto según el análisis de BPGA. En este ejemplo el pangenoma de *Clavibacter* se mantiene creciente mientras que el *core* no alcanza el cero. En contraste, al ampliar el rango taxonómico el *core genome* sí se acerca a cero según BPGA.

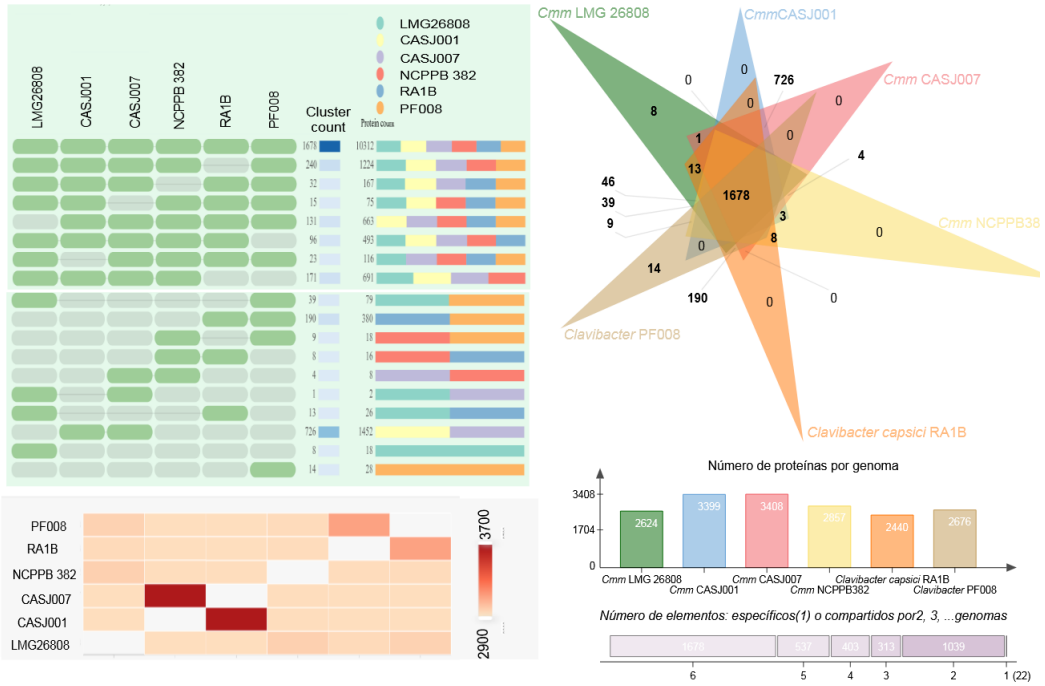

Figura 1.11 Diagrama de Venn del pangenoma de genomas selectos del género *Clavibacter*. VennChart permite visualizar la distribución de las familias génicas mediante un diagrama de Venn siempre que sean pocos genomas. En la parte 1superior izquierda se muestra el número de familias compartidas entre distintos grupos de genomas. Por ejemplo, existen 1678 familias génicas compartidas entre los seis genomas. Entre los grupos de cinco genomas, el más abundante es el que no contiene a RA1B con 243 familias, así pues, RA1B es de los genomas más divergentes de este grupo.

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

Orthocore es pues una herramienta para detectar el *core conservado*, es decir los genes del *core genome* que o bien no tienen copias extra o éstas son tan diferentes de la copia primaria que no pueden ser confundidas con ella. En este capítulo aprendimos que los genes marcadores son parte del *core conservado* de un grupo taxonómico, pero no necesariamente se dedican a lo que se entiende como metabolismo central en procariontes. Es decir, el *core conservado* no sólo está compuesto por rutas centrales como la síntesis de aminoácidos o la glicólisis. Como ejemplo de ello se mostraron familias pertenecientes al clúster de clavidicina *clvABCDEF* que pertenecen al *core conservado* de la especie *Clavibacter michiganensis* pero que salen del *core* cuando se considera todo el linaje de *Micrococcales*, i.e. no son parte del metabolismo central de ese orden. Estas

consideraciones nos explican que no hay una relación uno a uno entre metabolismo central y genes del *core*. Según el linaje que se considere existen genes del *core* que no son esenciales, y existen también genes de metabolismo conservado en Bacteria que no están en el *core* de un linaje, como por ejemplo *trpF* en Actinobacteria.

1.6. Consideraciones finales.

Así pues, en este capítulo aprendimos que con Orthocore podemos detectar los genes del *core genome* que no tienen copias extra. En oposición, en el siguiente capítulo veremos qué pasa con los otros genes del *core*, los que sí tienen copias extra. Estas familias de genes como se mencionó en la introducción son interesantes porque pueden presentar promiscuidad enzimática debido a la retención de la función ancestral. Sin embargo, las familias expandidas no necesariamente tienen que provenir del *core*. El siguiente capítulo veremos que el *shell genome* también puede presentar expansiones y que varias de ellas han sido reclutadas al metabolismo especializado. Esta búsqueda y clasificación de familias que han sido expandidas mediante copias extra fue sistematizada en el desarrollo de la herramienta EvoMining que es de lo que trata el siguiente capítulo.

Capítulo 2

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

2.1. Introducción

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

La promiscuidad enzimática puede buscarse en familias envueltas en procesos de divergencia funcional [Jensen, 1976; Zou, 2015; Copley, 2015; Huang, 2012]. Uno de dichos procesos es la expansión de familias pertenecientes a rutas metabólicas conservadas y su posterior reclutamiento hacia nuevo metabolismo [Caetano-Anolles, 2009; Soskine, 2010; Cruz-Morales, 2016]. Dentro de cada familia los niveles de promiscuidad pueden variar en cada ortólogo [Khanal, 2015, Verdel-Aranda, 2015, Verduzco-Castro, 2016]. 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 neo-funcionalización probablemente ocurrió a través de promiscuidad enzimática. Los homólogos con la función 'secundaria' suelen ser producto de expansiones previas de genes con la función 'primaria' [Huang, 2012; Jensen, 1976; Khersonsky, 2010]. Muchas mutaciones son neutrales a la función primaria [Bloom, 2007], por lo que esta puede ser retenida temporalmente cuando se está ganando una nueva función. EvoMining es un algoritmo que sigue esta estrategia para sugerir

cambios en la promiscuidad de una familia de enzimas dentro de un linaje definido por el usuario. Al cambiar la familia de enzimas en la que se buscan neo-funcionalizaciones 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 *clústeres* de genes distribuidos en un pequeño porcentaje de los organismos de un linaje taxonómico. Estos *clústeres*, conocidos como BGC (*Biosynthetic Gene Cluster*), contienen copias extras de genes de familias que pertenecen al metabolismo conservado. En este trabajo aprovechamos que en la actualidad es posible predecir nuevos BGC mediante estrategias bioinformáticas gracias a la gran cantidad de secuencias disponibles públicamente, así como la facilidad para secuenciar nuevos genomas. La similitud de secuencia de los genes que pertenecen a los BGC, así como su sintenia en diversos organismos de un linaje hacen que genómica comparativa sea de utilidad para intentar localizarlos.

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

2.1.2. 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 postraduccionalmente (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, 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 Figura 2.1 donde se aprecia que un porcentaje de los BGC reportados tanto en Actinobacteria como en Cianobacteria pertenece al grupo “otros”. La ausencia de enzimas biosintéticas conocidas hace que estos BGC sean “atípicos”, difíciles de identificar. Los BGC no tradicionales suelen pasar desapercibidos porque no hay conocimiento previo de ellos que permita reconocerlos.

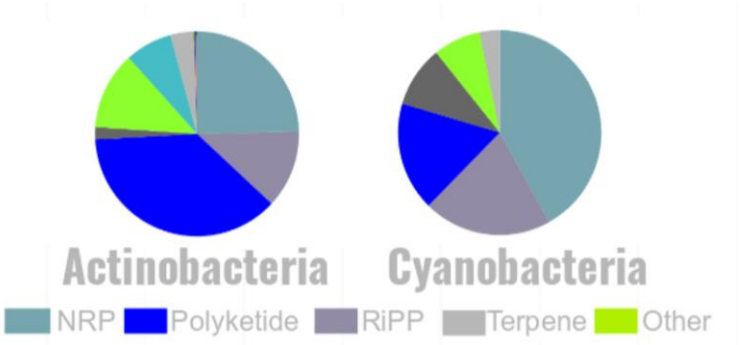

Figura 2.1 Existen *clusters* biosintéticos no clasificados (*other*) entre los reportados en MIBiG.

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 ya no hace la reacción sobre el metabolito con fósforo sino con un análogo que tiene arsénico en su lugar (Figura 4). Luego de identificar esta enzima divergente de su homóloga de metabolismo central pudimos identificar que era parte de una región que además mantiene la sintenia en un clado de los *Streptomyces*. Este descubrimiento 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 [Cruz, 2013]. Otro ejemplo de este escenario es el BGC de la escitonemina [Garciapichel, 1992], un pigmento Cianobacteriano que absorbe luz UV. Su biosíntesis requiere de ScyB y ScyA, dos enzimas que sostienen la síntesis de este metabolito especializado [Balskus, 2008, Soule, 2009].

Curiosamente, ScyB y ScyA son homólogos distantes de las enzimas glutamato deshidrogenasa (GDH) y acetolactato sintasa (ALS), que participan en la desaminación oxidativa reversible del glutamato a α −cetoglutarato y amoníaco [Engel, 2014] y en la síntesis de aminoácidos de cadena ramificada [Liu, 2016], respectivamente. Una descripción amplia del origen y destino de las enzimas del BGC de escitonemina será provista en este capítulo.

2.2. Algoritmos y bases de datos de EvoMining 2.0

EvoMining está compuesto de dos algoritmos: el primero que utiliza a la familia semilla para encontrar 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 BGC de acuerdo a reportes previos. El segundo algoritmo permite la visualización de todas las copias de una familia expandida en un árbol clasificadas según sus posibles destinos metabólicos. Para ello, los algoritmos EvoMining necesitan tres bases de datos: i) los genomas blancos, ii) las secuencias de enzimas semilla, y iii) la de productos naturales verificados (Figura 2.2).

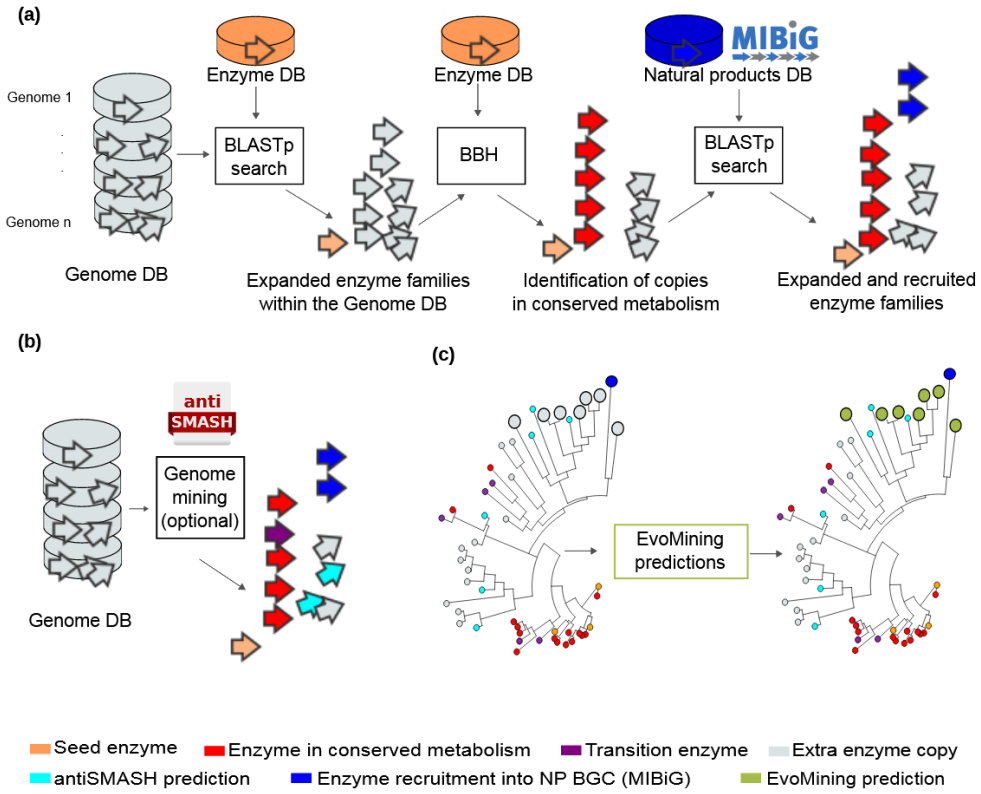

Figura 2.2 Representación de la tubería bioinformática de los dos algoritmos que componen EvoMining. a) Algoritmo de expansión - clasificación/reclutamiento. Se ingresan la Enzima DB (cilindro anaranjado) y la Genome DB (cilindros grises) para identificar por medio de BlastP a todos los que sean miembros de la Familia de las enzimas en Enzima DB, luego se identifican los ortólogos de la base de datos de enzimas (Flechas rojas), finalmente se buscan homólogos de la familia similares a genes reclutados a BGC (Flechas azules), b) EvoMining toma en cuenta las predicciones de antiSMASH. Cuando un gen es encontrado como miembro de un BGC predicho por antiSMASH se colorea de cian, o bien de morado si había sido detectado también como parte del metabolismo central (rojo). c) El Algoritmo de visualización muestra un árbol donde cada homólogo de la familia expandida tiene un código de color que lo clasifica como se describe en la leyenda.

2.2.1. Algoritmo de expansión y clasificación de grupos de homólogos.

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

a) Secuencia semilla (Enzyme DB). Es una secuencia o un conjunto de secuencias de la misma familia que serán interpretadas como las enzimas con una función primaria. Por esta razón es importante que el usuario haga una curación minuciosa en la que se sugiere que utilice genes con evidencia experimental que sí sean de una familia enzimática única y conservada. La implementación actual de EvoMining 2.0 permite que se introduzcan en el archivo de entrada varias familias bien identificadas simultáneamente, sin embargo, el algoritmo procesa una familia a la vez.
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 más dos desviaciones estándar. Siguiendo esta definición EvoMining colorea en un diagrama de calor 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.

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

El último paso de este algoritmo consiste en encontrar miembros de la familia que se encuentren reclutados a BGC reportados anteriormente. En este paso definimos un reclutamiento como una copia extra de una familia de metabolismo conservado que ahora participa en un *clúster* 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 FEs como *query* 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 con las predicciones de antiSMASH. AntiSMASH busca secuencias con dominios conocidos de BGC tradicionales en la base de datos de genomas blanco (Genome DB) y además busca entre los genes circundantes la presencia de otras enzimas que ya hayan sido reclutadas a BGC conocidos. Si alguna enzima de la FE es detectada por antiSMASH nos será también útil para identificarla como el producto de una expansión que está en proceso o se ha neofuncionalizado. La minería genómica tradicional realizada por antiSMASH no es parte de la tubería de EvoMining, pero las predicciones de antiSMASH pueden ser calculadas previamente por el usuario y utilizadas por EvoMining. En este trabajo antiSMASH 3.0 [Weber, 2015] fue utilizado en las secuencias de los genomas de la Genome DB (Figura 2.2, panel b).

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

En conclusión, el algoritmo de expansión - clasificación/reclutamiento trabaja para identificar tres clases de copias en las familias enzimáticas expandidas y tiene como resultado tres salidas. La

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

2.2.2. Algoritmo de reconstrucción filogenética y visualización

Una vez que se tienen definidos los genes de una familia expandida (FE) de enzimas y que ya se conoce cuáles de los homólogos realizan la función primaria, así como los que han sido reclutados a BGC para realizar una función divergente, en este algoritmo se visualiza esa información junto con la inferencia de los genes que no han sido clasificados todavía. Para eso se alinearon 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 con la inferencia de origen-destino.

Las secuencias de la FE detectadas por el algoritmo de búsqueda - clasificación/reclutamiento se alinean con MUSCLE v3.2 y son curadas con Gblocks. Las posiciones ausentes en más del 50% de las secuencias se filtran y remueven del alineamiento final. Para reconstruir filogenéticamente la historia de las enzimas se utiliza FastTree 2.1 [Price, 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, 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 (Figura 2.2, panel b y 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, 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 que están más cerca de una hoja azul que de una hoja roja. Es decir, son los homólogos que están en ramas del árbol que contienen genes descritos como reclutados por BGC y que no están en ramas de los homólogos con la función primaria. Estas predicciones consideradas con más posibilidades de pertenecer al metabolismo especializado que al conservado, son coloreadas en verde en la Figura 2.2, panel c.

Además de los tres destinos metabólicos descritos marcados respectivamente en rojo, azul y verde, se puede opcionalmente agregar información predicha por antiSMASH. Cuando el usuario provee resultados de antiSMASH sobre qué genes pertenecen a un BGC que contiene una enzima típica de metabolismo especializado, estos se colorean en cían y son llamados predicciones antiSMASH. Si una secuencia es al mismo tiempo predicción EvoMining y predicción antiSMASH se colorea cían 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 cían 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, cian o púrpuras se les asigna el color gris. Así pues, gris son aquellas hojas del árbol de las que no se tiene un clave sobre su destino metabólico, estas secuencias son llamadas de destino metabólico desconocido.

En este trabajo se diseñó e implementó el código para hacer el visualizador de EvoMining que permite expandir visualmente 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, 2016].

2.2.3. Actualizaciones de las bases de datos de EvoMining

Tres bases de datos son requeridas como variables de inicio de EvoMining, la primera es el conjunto de secuencias de genomas de un linaje, esta base fue llamada Genome DB. La segunda es un grupo de secuencias de enzimas de metabolismo conservado llamada Enzyme DB. La base de datos se

secuencias de genes que pertenecen a un clúster 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.

-**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, 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, 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 con reconstrucciones metabólicas de los organismos *Streptomyces coelicolor*, *Mycobacterium tuberculosis* y *Corynebacterium glutamicum* [Cruz-Morales, 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, 2016]. La herramienta Metaphor [Van-der-veen, 2014] fue implementada para obtener los BBH, se filtraron aquellas secuencias con menos del 30% de identidad en un alineamiento del 80% de la secuencia de las dos proteínas. Como resultado, las 106 familias de Actinobacteria quedaron reducidas a 42 FEs, compartidas por los genomas semilla de Actinobacteria, Cianobacteria, *Pseudomonas* y Archaea. Las bases de datos Enzyme DBs de todos los linajes están disponibles en Zenodo con número de identificación DOI 10.5281/zenodo.1219709

-**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, 2016]. En este trabajo la base NP DB que se utilizó para los análisis es MIBiG v1.3 [Medema, 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.

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

2.3.1. Los perfiles de expansión de las proteínas dependen del linaje.

Para entender la evolución de las enzimas y rutas metabólicas se seleccionaron cuatro linajes de diversas características los phyla Actinobacteria y Cianobacteria, el género *Pseudomonas* y el dominio Archaea. Todos los resultados en las figuras son presentados en este orden. Estos taxa fueron seleccionados para tener un espectro de análisis que abarcara tanto microorganismos ampliamente reconocidos como productores de NP, es decir Actinobacteria (602 MIBiG BGC), Cianobacteria (60 MIBiG BGC) y *Pseudomonas* (53 MIBiG BGC); como también Archaea (0 BGC en MIBiG versión 1.3), que representa un dominio poco explorado en lo que respecta a los genes que forman parte de metabolismo especializado [Charlesworth 2015].

Basándonos en estas bases de datos de genomas blanco (Genome DB), como es explicado en la Figura 2.3, 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 taxón, contiene sólo 42 FEs (Tabla A.1). 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, 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 (Figura 2.3). 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 Archaea se debe a que no se han descubierto organismos tamaños de genoma Archaea de tamaño comparable a los de *Streptomyces* o

Pseudomonas (> 5Mbp). Actinobacteria y Cianobacteria, aunque tienen genomas > 5Mbp, el incremento en tamaño podría ser porque tienen expansiones de genes que no son del metabolismo más conservado, porque obtienen más genes por transferencia horizontal u otros mecanismos por los que se incremente el tamaño del genoma. Cianobacteria a pesar de tener genomas grandes no tiene tantas expansiones en estas FEs Esta observación, es posible que sea generalizable a todas las FEs de metabolismo conservado o bien que se deba a un sesgo en la selección de las familias que componen a las FEs.

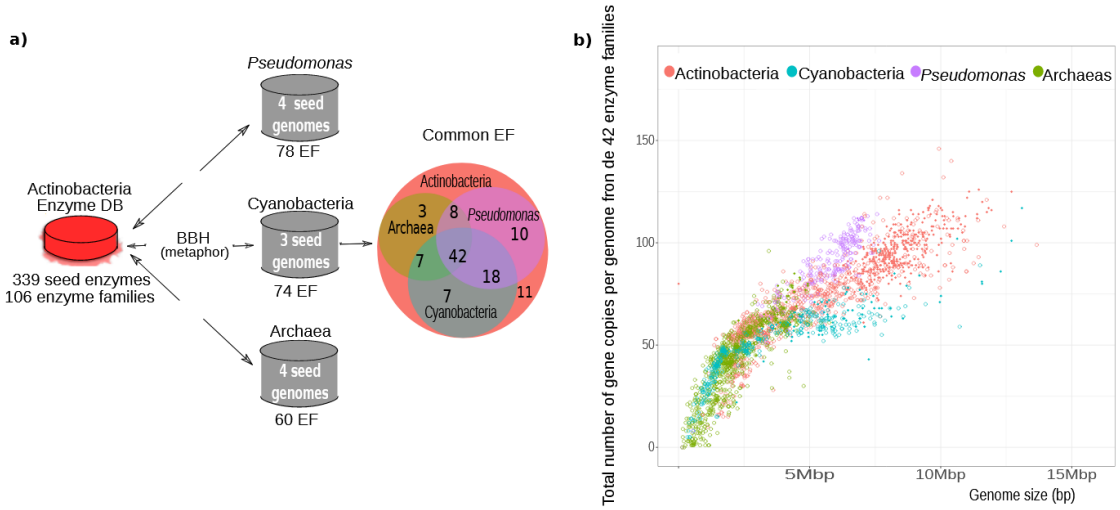

Figura 2.3 EvoMining Enzyme DB. (a) La base de datos de enzimas de la versión previa de EvoMining se filtró para establecer un conjunto común de 42 familias enzimáticas conservadas para los phyla Actinobacteria y Cianobacteria, el género *Pseudomonas* y el dominio Archaea. (b) Todos los taxones muestran 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 Cianobacteria 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 (Figura 2.4).

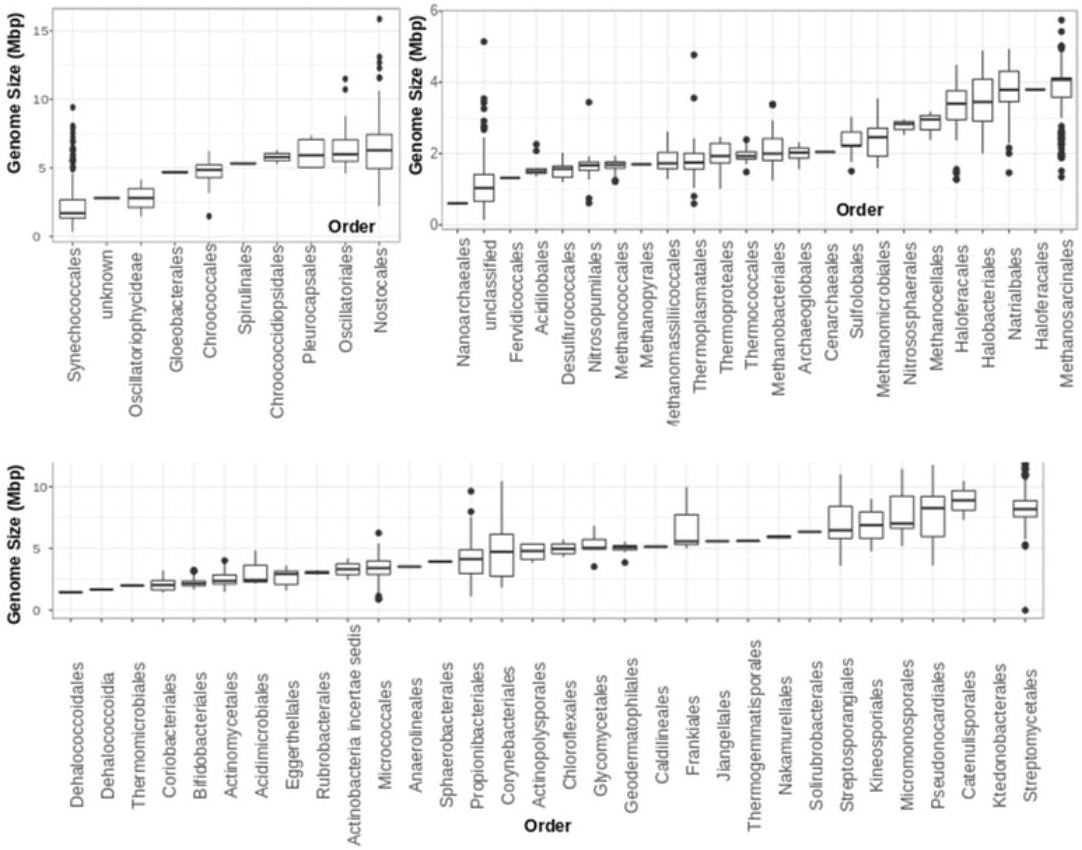

Figura 2.4 Tamaño de genoma en órdenes de Actinobacteria y Cianobacteria yen clases De Archaea. Actinobacteria y Cianobacteria tienen algunos genomas con tamaños superiores a 6 Mbp, que es el máximo que se encuentra en Archaea. El tamaño promedio en el género *Pseudomonas* es de 5.8 Mbp con un máximo de7.6Mbp 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 naturales de Archaea fueron aislados de especies de Halobacteria, los genes, probablemente en BGC, que sintetizan de estos metabolitos no han sido caracterizados [Charlesworth, 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 5 Mbp (Figura 2.3). Por ello se concluye que para ensamblar una base de datos genómica para EvoMining se debe considerar

que las expansiones dependen tanto de los distintos linajes taxonómicos como de la diversidad del tamaño de genoma.

2.3.2. El *shell genome* posee expansiones en sus familias enzimáticas

Pseudomonas posee en promedio más copias por genoma que los otros taxa como se puede ver en la Figura 2.3. De las 42 familias analizadas el 54.8% tiene su máximo número promedio de copias por genoma en este linaje. 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. Aunque existen ejemplo como las familias acetil ornitino aminotransferasa y la acetolactato sintasa (ALS) que están expandidas en todos los linajes (Figura 2.5, coordenadas A1 y E1, Tabla A.1). 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 aún no hay árbol filogenético. Así pues, es como sigue: rojo para el metabolismo conservado, azul para los reclutamientos anotados en MIBiG, cían 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 A.1. 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 un cuadro, estas enzimas son el origen de los reclutamientos en el BGC de escitonemina. Muchas otras familias exhiben expansiones sólo en ciertos linajes. Tal es el caso de la familia fumarato reductasa subunidad de hierro-azufre, coordenada C3 de la figura, muy expandida en Actinobacteria, pero con menos de una copia por genoma en promedio en Cianobacteria.

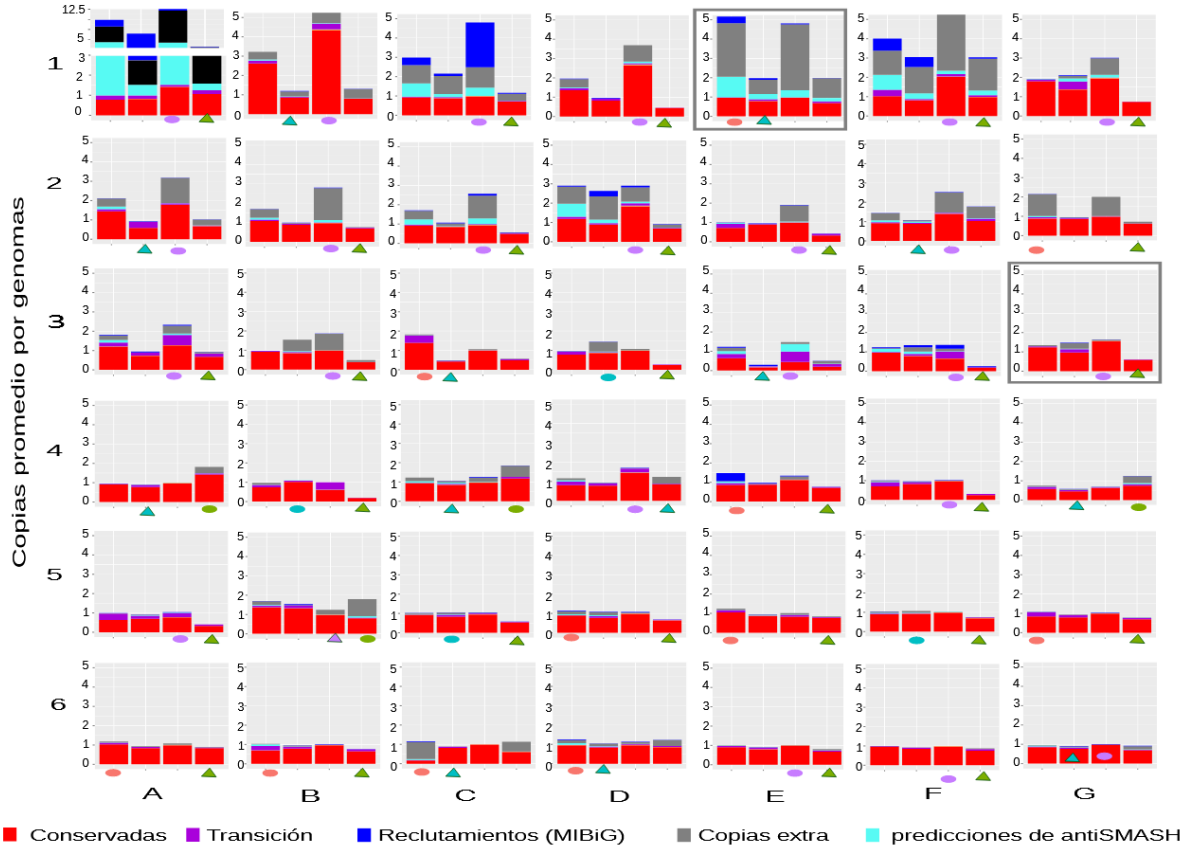

Figura 2.5 Perfiles de EvoMining de las 42 enzimas conservadas en linajes genómicos seleccionados. Las coordenadas en forma de letras A-G y los números 1-6 se muestran en esta figura para Localice fácilmente la familia y sus propiedades en la Tabla A.1.

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. 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 (Figura 2.6, 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 Figura 2.3. Del total de las 42 familias, GDH es una de las cuatro FEs en las que el mayor número de expansiones se encuentra en Archaea. De hecho, GDH tiene menos de una copia por genoma en promedio en los

otros taxa, probando que no se encuentra dentro del *core* genómico de estos linajes. Esto contrasta con AroB, que muestra una tendencia opuesta, no es parte del *core* de Archaea, pero muestra copias extra y una presencia mayor que uno en promedio en los otros tres taxa analizados (Figura 2.6, panel b). Los ocho casos mostrados en la figura son todos parte de un clúster biosintético de Cianobacteria, descrito en las secciones posteriores de este trabajo.

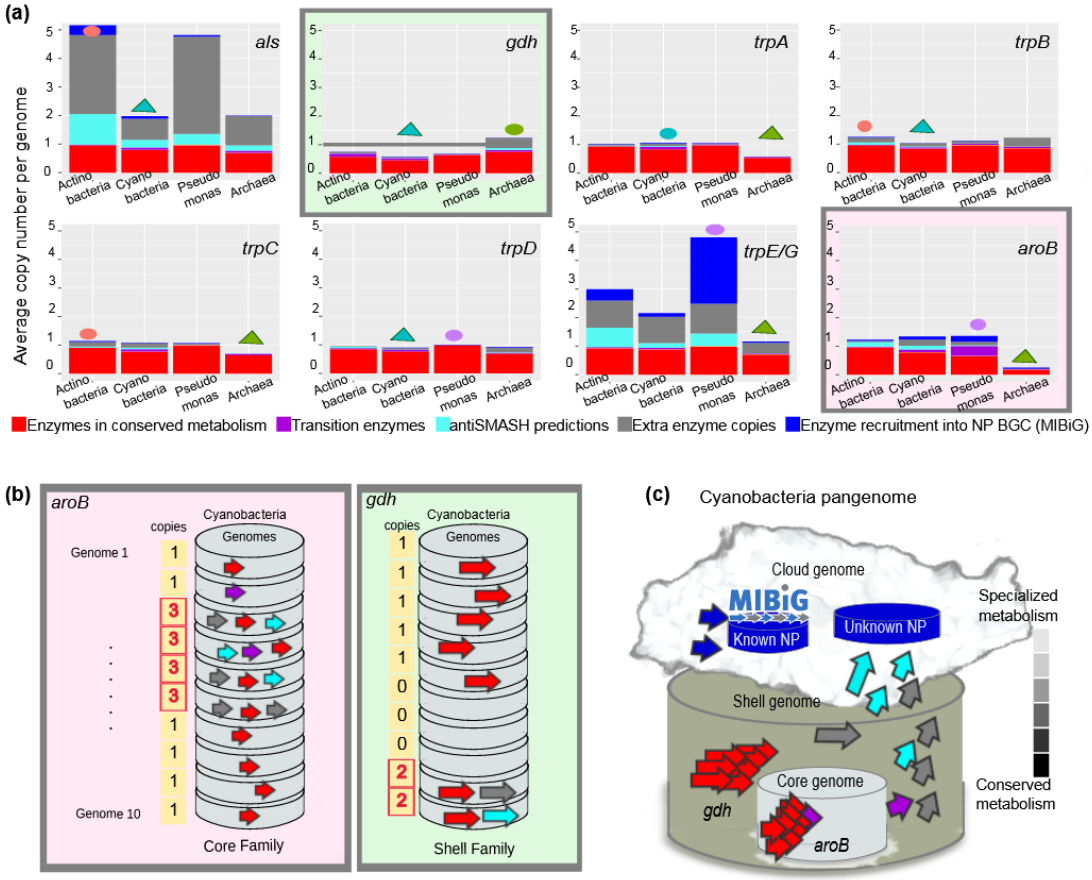

Figura 2.6 Perfiles de expansiones de EvoMining de enzimas conservadas seleccionadas. (a) Patrones de expansión de las ocho familias conservadas cuyas copias adicionales participan en la biosíntesis de escitonemina. La codificación de colores es la siguiente: rojo para el metabolismo conservado, azul para los reclutamientos anotados en MIBiG, cian para las predicciones antiSMASH de metabolismo especializado, púrpura es la intersección entre 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 están el linaje con el mayor número de copias por genoma en promedio, y los círculos en el linaje menos expandido. Aunque Archaea tiende a tener menos expansiones, esta tendencia se revierte en la familia GDH. (b) Se proporción aun ejemplo de un *core* frente a una familia enzimática *shell*. AroB es un EF básico porque tiene al menos una copia por genoma, mientras que GDH es un familia del *shell* debido a su ausencia en tres genomas. A pesar de ser una familia *shell*, GDH tiene expansiones que pueden ser reclutadas en un metabolismo especializado.

(c) Modelo para la nube o genoma variable compuesto parcialmente por enzimas que pertenecen a BGC de productos naturales. En este modelo, el metabolismo conservado se compone de familias tanto del *shell* como del *core* genome. Estas familias pueden sufrir eventos de expansión, y algunas de las copias adicionales son reclutadas para realizar nuevas funciones en el metabolismo especializado.

Con estas observaciones sospechamos que GDH es miembro del *shell genome* (Ver capítulo 1) [Koonin, 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 una familia pertenece al *shell*, podrían suceder casos sobre todo cuando hay mucha variación en un taxón como en el caso de un domino o un phylum en contraposición con taxones conservados como géneros en los que para una cierta familia la mitad de los genomas de un linaje tuviera dos copias y la otra mitad cero. Sin embargo, en el caso de la GDH si es consistente en que está presente en más del 50% de los genomas de cada linaje (Figura 2.6, panel a). Las modas de número de copias también son informativas, una sola copia extra puede ser la que sea reclutada en metabolismo especializado.

En la Figura 2.6 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 cían, 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 cían.

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 *clústeres* biosintéticos escitonemina y el policétido pactamicin [Kudo, 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 Figura 2.7.

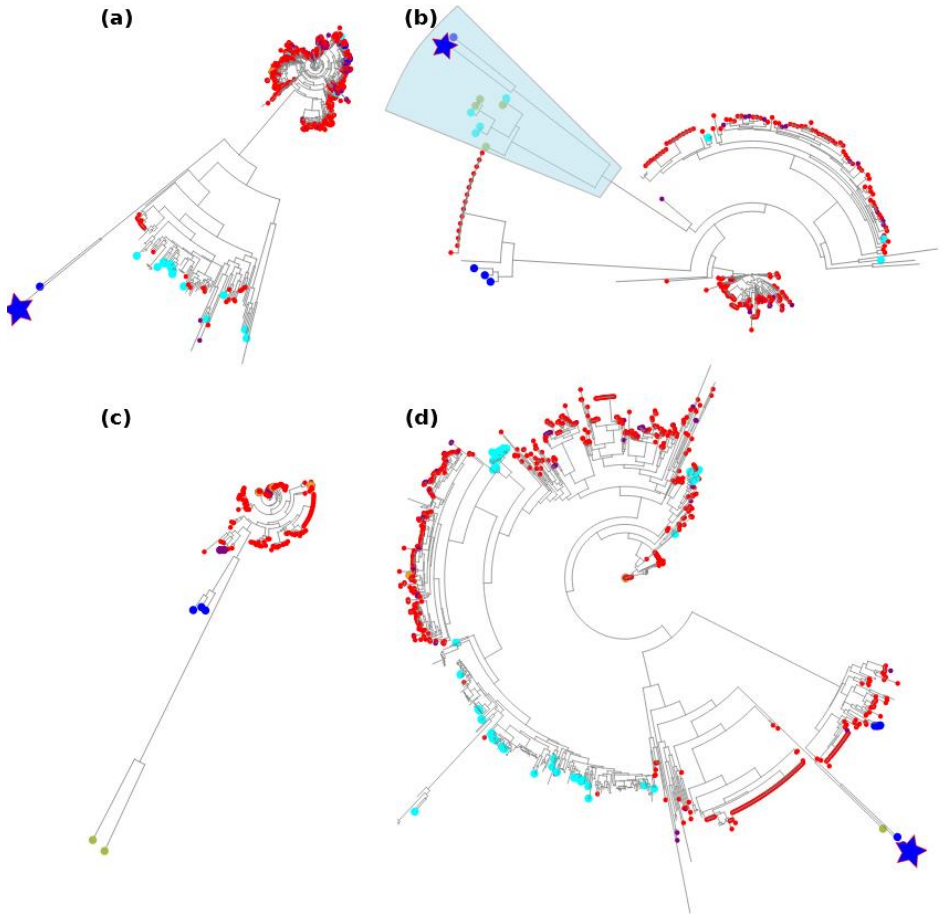

Figura 2.7 Árboles de EvoMining de glutamato deshidrogenasa en cuatro linajes genómicos. (a) Reconstrucciones filogenéticas específicas del linaje que muestran claras diferencias en los perfiles de expansión en Actinobacteria, Cianobacteria, *Pseudomonas* y Archaea. Actinobacteria no tiene predicciones de EvoMining, y a 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 Cianobacteria 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, Cianobacteria y Archaea.

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, de la figura 2.6 paneles a y 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* Figura 2.6, 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, 2018].

En la siguiente sección se analizarán los patrones de expansión reclutamiento de GDH provistas por EvoMining para GDH y se describirán los árboles filogenéticos de cada linaje mostrados en la Figura 2.7, así como un árbol que incluye conjuntamente secuencias de todos los linajes. 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*. Además de GDH se estudiaron las expansiones y los árboles filogenéticos de TrpA, TrpB, TrpC, TrpD, TrpEG, AroB y ALS. Todas ellas pertenecientes a las 42 familias enzimáticas conservadas entre los cuatro linajes y a la vez reclutadas en escitonemina [Balskus, 2008; Soule, 2009] un clúster biosintético de Cianobacteria (Figura 2.8).

2.3.3. GDH y ALS en el clúster 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 [Andersson, 2003; Lilley, 1991] (Figura A.3). GDH cataliza la reacción reversible de desaminación oxidativa de glutamato en α-cetoglutarato y amonio. De acuerdo con el uso de cofactores, la familia GDH puede dividirse en tres 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, 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, 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 ,y en Cianobacteria. Las expansiones están prácticamente ausentes en *Pseudomonas*. En contraste, un número significativo de expansiones es encontrado en Archaea Figura 2.7. 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, 1991]. En Archaea las tres clases de GDH alternan en las ramas del árbol (Figura A.4).

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, 1996], incluyendo muchas predicciones de antiSMASH, y sólo dos marcadas como metabolismo conservado. Estas dos marcas pueden deberse a la pérdida real de una enzima de 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 geranio pirofosfato, un precursor de todos los terpenos y terpenoides [Tholl, 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 Figura 2.7.

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 (Figura 2.7, Tabla A.1). 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, 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* aun 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 gen *scyA* (Figura 2.8, 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. 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.

Al generar el árbol de EvoMining de ALS en Cianobacteria, se observó que *scyA* es un reclutamiento que se localiza en una rama repleta de secuencias de ALS provenientes de *Nostoc sp.*, que fueron etiquetadas como predicciones de EvoMining (c). Estas predicciones incluyen más de veinte organismos conocidos como productores de escitonemina [Balskus, 2008]. Además, ramas cercanas muestran secuencias de ALS que son predicciones de antiSMASH, reforzando la sugerencia de que esta sección del árbol se dedica al metabolismo especializado. Una última rama contiene a los mismos organismos encontrados en el árbol de EvoMining de la familia GDH. Estos organismos son mostrados en el acercamiento de la rama de *scyB* (Figura 2.8, panel b).

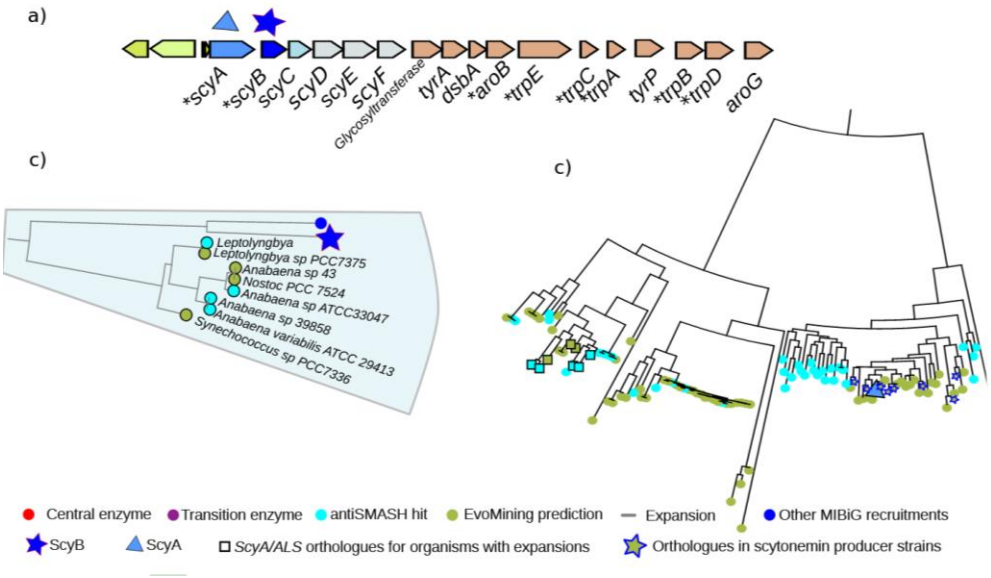

Figura 2.8 Reclutamientos de GDH y ALS por el cluster biosintético de escitonemina. (a) El BGC de escitonemina de *Nostoc punctiformese* compone de genes reguladores (verde), genes que participan en la biosíntesis de escitonemina (azul) y de genes dedicados al suministro de precursores (marrón). Se encontró 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 Cianobacteria 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 Cianobacteria están disponibles en Microreact.

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 clúster de escitonemina. Estas observaciones son una lección para usar EvoMining como herramienta de minería genómica: enzimas cercanas pueden co-diversificarse en ocasiones formando parte de un mismo BGC, pero a la vez estar sujetas a distintas restricciones evolutivas.

2.3.4. El clúster de escitonemina es una familia de clústeres cuyas variantes en los genes accesorios muestran promiscuidad de producto.

Definimos *clúster promiscuo* como una familia de BGC homólogos cuyos miembros sintetizan distintos productos naturales con la misma estructura base, pero con algunas modificaciones que son específicas de cada organismo incluyendo aquellos BGC que en un mismo organismo pueden producir varias moléculas muy similares. Un clúster puede ser promiscuo ya sea por la promiscuidad de alguna de sus enzimas o por la diversidad en los genes accesorios particulares a cada organismo. En oposición tenemos los clústeres u operones del metabolismo central, donde la selección ya ha actuado siempre obtener el mismo producto, por ejemplo, algún aminoácido en particular. Como veremos más adelante el BGC de escitonemina presenta diversidad de productos.

El *clúster* de la escitonemina reportado en MIBiG y mostrado en la Figuras 2.8 y 2.9 en comprende 18 genes [Soule, 2009]. Además de genes reguladores, este BGC incluye a los 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 shiquímico, parecen haber sido reclutadas para proveer de los precursores L-triptófano y prefrenato, que son necesarios para la síntesis de escitonemina. En oposición a las enzimas del operón de triptófano y a AroB que siguen realizando su función de metabolismo conservado aún como 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 (Figura 2.9). ALS une dos piruvatos, transformándolos en S-2-acetolactato [Liu, 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, 2014], mientras que ScyA cataliza una desaminación oxidativa de triptófano. El producto de estas dos enzimas actuando secuencialmente en 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 cómo se llevan a cabo estos últimos pasos [Balskus, 2008].

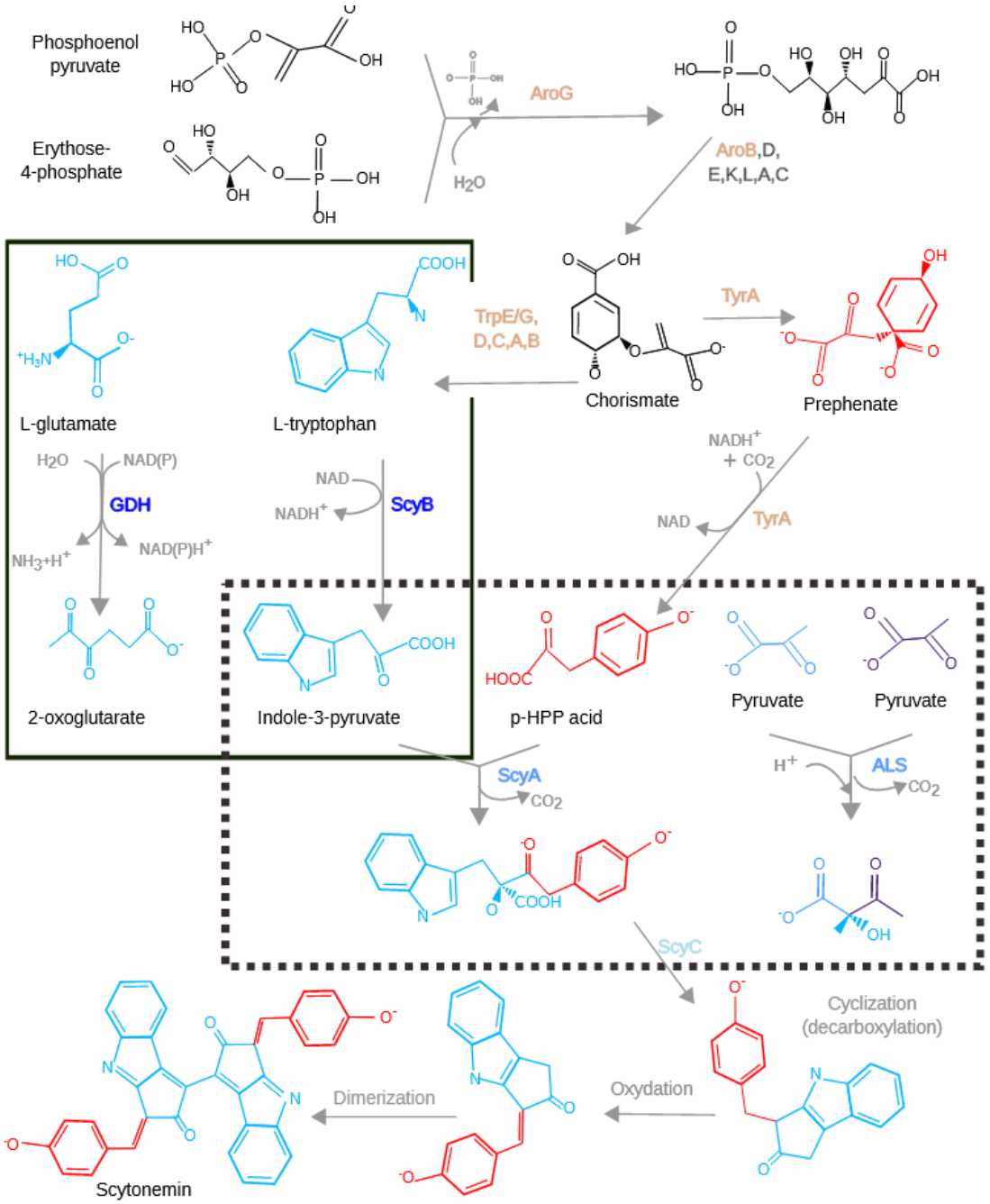

Figura 2.9 Origen metabólico y destino de GDH / ScyA y ALS / ScyB en la biosíntesis de escitonemina. AroG y AroB participan en la síntesis de corismato, un intermediario que se transforma en los precursores que conducen a sustratos de ScyA, es decir, l-triptófano y prefenato. La reacción catalizada por 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.

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 clúster de escitonemina (Figura 2.6 y Figura 2.8). 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 clúster reportado en MIBiG. Los reclutamientos incluyen AroB y todos los genes de la ruta del L-triptófano excepto *trpF*. Estos árboles pueden ser consultados interactivamente, los enlaces de Microreact están en la Tabla 2.1.

Tabla 2.1 Árboles de EvoMining de enzimas relacionadas con la escitonemina en MicroReact

Linaje	Enlace al árbol de EvoMining en Microreact
TrpA	https://microreact.org/project/SyZMprKum?tt=cr
TrpB	https://microreact.org/project/H1jW0rFdX?tt=cr
TrpC	https://microreact.org/project/rkN1THFum?tt=cr
TrpE/G	https://microreact.org/project/rkv20SFOm?tt=cr
TrpD	https://microreact.org/project/H1UuQE0qm?tt=cr
AroB	https://microreact.org/project/SkT1Wp_dm?tt=cr
GDH	https://microreact.org/project/HyjYUN7pQ?tt=cr
ALS	https://microreact.org/project/B11HkUtdm?tt=cr

Los árboles de EvoMining de estas familias incluyen ramas marcadas por reclutamientos que son enzimas que forman parte de las rutas de síntesis de otros pigmentos protectores solares. Entre ellos la shinorina y los aminoácidos de tipo micosporina (MAAs) [Balskus_genetic_2010]. Entre los reclutamientos también están otros no relacionados a la protección solar, como la welwitindolinona [Hillwig, 2014], la ambiguina [Li, 2015] y la fischerindolina [Li, 2017]. Estos resultados ilustran cómo EvoMining puede complementar a antiSMASH mediante la identificación de secuencias que pertenecen a BGC no tradicionales.

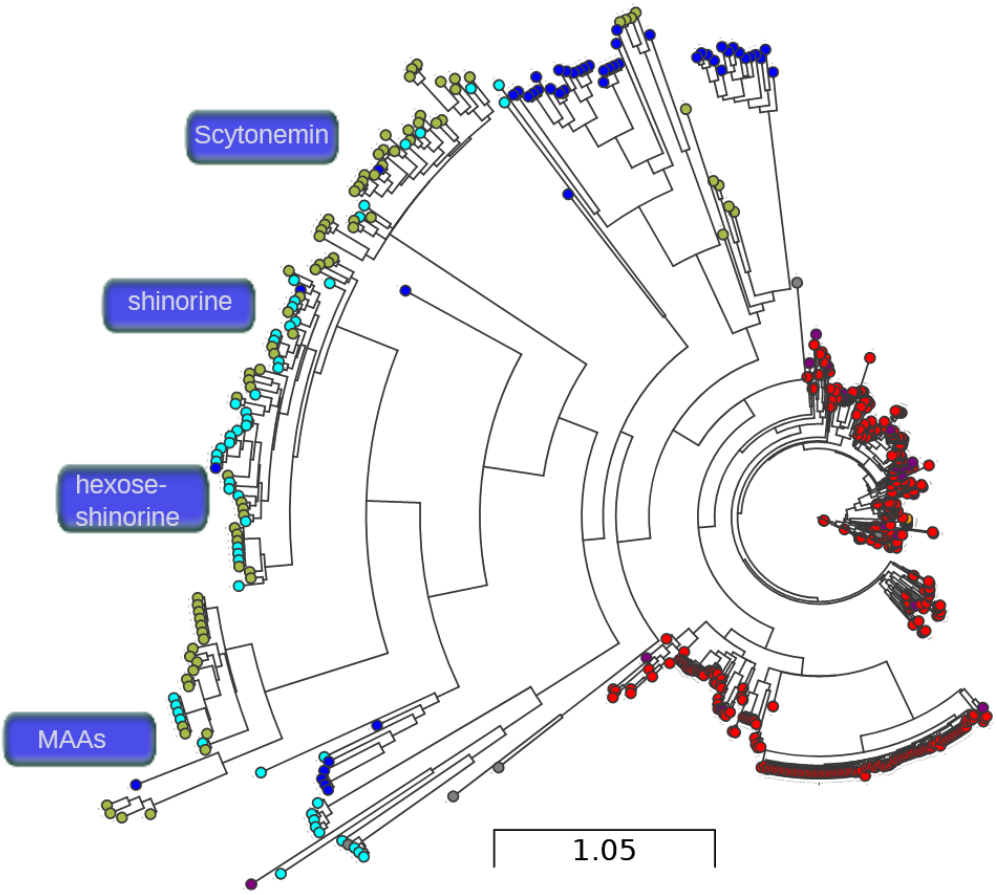

Figura 2.10 Árbol de EvoMining de AroB. El número promedio de copias de esta familia es 1.3, la moda es uno con 29.5 por ciento del total de los genomas superando este umbral. Estas características se reflejan en tres ramas llenas de copias adicionales marcadas ya sea como predicciones de antiSMASH o 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 triptófano, 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 los 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 conclusión, escitonemina es un ejemplo de clúster 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 (Figura 2.11). 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 mono oxigenasas, 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 tirosinasa y o la amidasa pueden derivar en la síntesis de los compuestos 3 y 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, sin embargo, si existen variantes reportadas de la molécula de escitonemina [Grant, 2013]. 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.

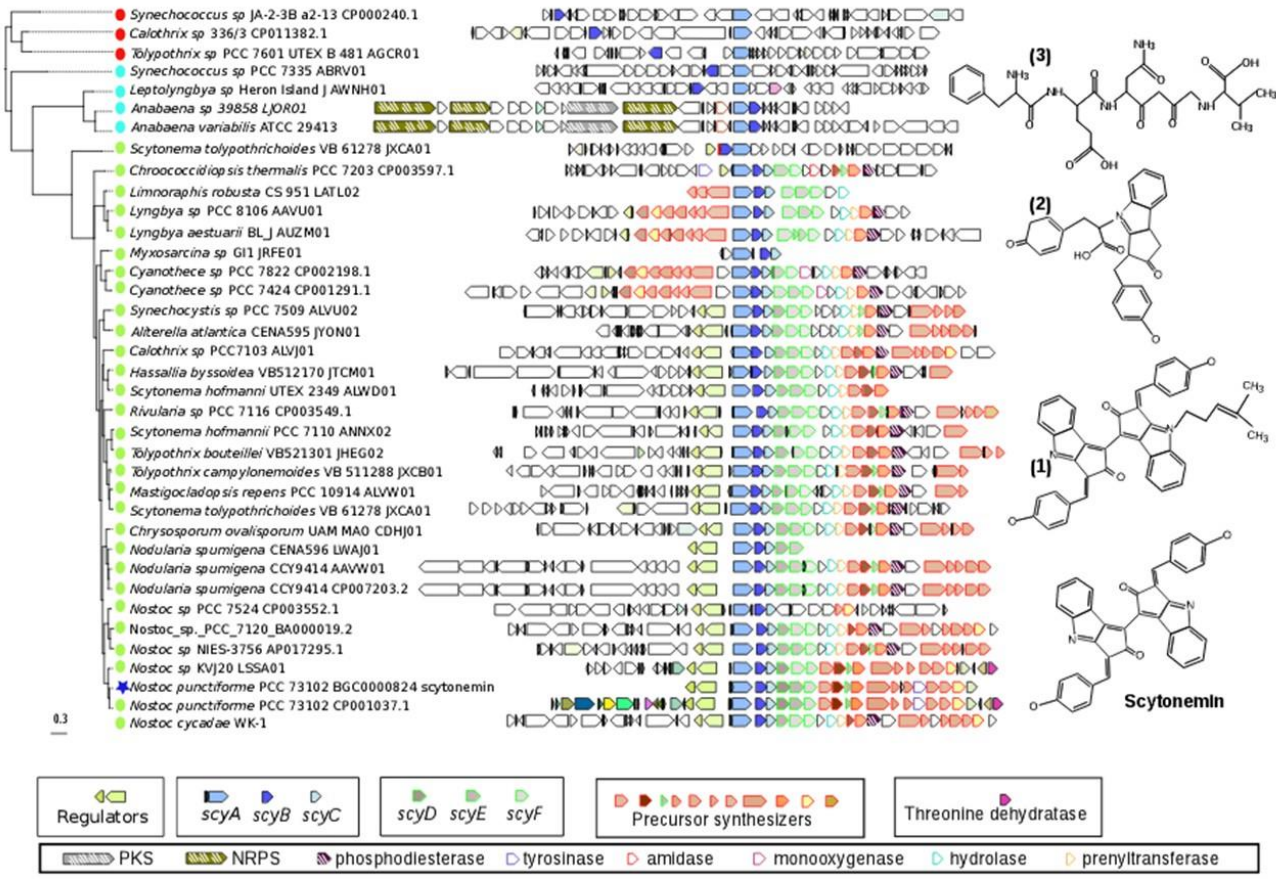

Figura 2.11 El análisis filogenómico de *scyA* y *sycB* mostró una diversidad química en torno al BGC de la escitonemina. Las proximidades genómicas que contienen tanto *scyA* como *scyB* en Cianobacteria se muestran junto a una reconstrucción filogenética utilizando las secuencias de proteínas de estos dos genes.

2.4. EvoMining aplicado a TauD, enzima común de los BGC Rimosamida y Detoxina sugieren otra clase de clústeres con promiscuidad de producto.

Además del ejemplo de las familias del BGC de escitonemina ampliamente discutido en este capítulo, y de las 42 familias de metabolismo conservado apliqué EvoMining para estudiar las expansiones de otras familias enzimáticas. Una de ellas es TauD una enzima del *shell genome* de Actinobacteria que cuya contraparte homóloga en Enterobacteria forma parte del operón de *E. coli* metabolismo de taurina. En Actinobacteria una copia de *tauD* pertenece a este operón que está parcialmente

conservado, mientras que otras copias, cuyo destino metabólico nos muestra EvoMining aparecen en una rama marcada por reclutamientos de MIBiG (Figura 2.12). EvoMining también encuentra expansiones de esta familia en *Pseudomonas* donde además el contexto genómico de la copia secundaria se conserva cerca de una PKS, sugiriendo la pertenencia de *tauD* a BGC de productos naturales. En Actinobacteria *tauD* es parte de 15 BGC, entre ellos los *clústeres* que producen los metabolitos rimosamida y detoxina (Tabla 2.2). Estos metabolitos comparten un *core* molecular y difieren en ornamentaciones. Esta observación y el hecho de que estos BGC comparten más genes además de la rimosamida, nos pueden sugerir que los BGC de Rimosamida y detoxina, también pueden considerarse como parte de una sola clase de BGC que presentan promiscuidad de producto. El estudio de las variantes de rimosamida - detoxina BGC será objeto del siguiente capítulo donde cambiaremos de estudiar las variaciones a nivel secuencia de enzimas a estudiar variantes de clústeres biosintéticos.

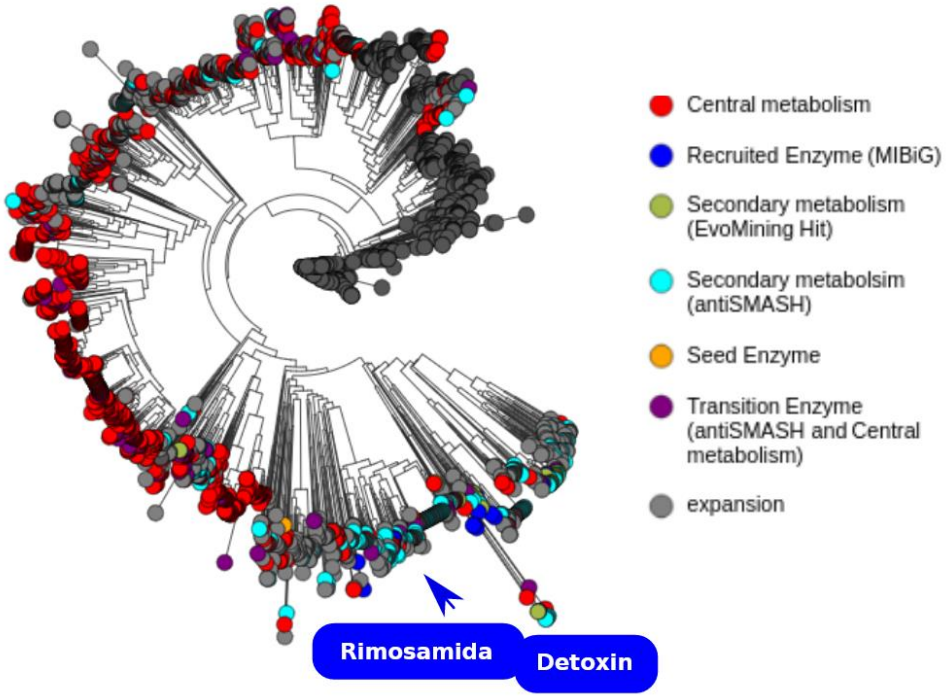

Figura 2.12 Análisis de EvoMining de las expansiones de la famila TauD. En Actinobacteria existe una rama dedicada al metabolismo especializado. Dentro de esta rama hay copias extra en géneros como *Streptomyces*, *Rhodococcus*, *Frankia* y *Amycolatopsis*. Esta figura muestra un clado dentro de las expansiones de la familia que contiene quince homólogos de *tauD* que pertenecen a clusters biosintéticos experimentalmente caracterizados y depositados en MIBiG, incluyendo los de detoxin y rimosamidas. La variedad de BGC mostrada en este clado abre la posibilidad de encontrar variantes moleculares de estas familias

Tabla 2.2 Homólogos de *tauD* en BGC reportados en MIBiG

MIBiG BGC	Compuesto	Clase	Organismo Productor
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

2.5. 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 familias enzimáticas de distintos phyla. Nuestros análisis llevaron a la conclusión de que los patrones de expansión reclutamiento dependen tanto de la familia enzimática como del linaje genómico en el que se analiza. Una consideración importante al usar EvoMining es que el tamaño de genoma correlaciona con el número de copias extra de familias expandidas. Aunque el tamaño de genoma es importante, también encontramos excepciones donde EvoMining pudo predecir enzimas de BGC no tradicionales en genomas relativamente pequeños, sugiriendo que hacen falta más análisis para estudiar esta relación. En este sentido, optamos por comparar linajes genómicos que no solo son altamente divergentes y, en algunos casos, poco conocidos con respecto a la biosíntesis de NP, sino también desproporcionados en cuanto a su resolución taxonómica y distancias. Por lo tanto, es

posible que estos factores hayan impuesto un sesgo al establecer relaciones entre el tamaño del genoma, la tasa de expansión de genes y la diversidad metabólica.

Es interesante observar que las familias más expandidas según los análisis de prueba de concepto anteriores de EvoMining [Cruz-Morales, 2016] fueron asparagina sintasa, 2-dehidro-3-deoxifosfoheptanoato aldolasa y 3-fosfosquimato-1-carboxivinil transferasa, que son las 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 Archaea deben ser exploradas experimentalmente. Estas observaciones enfatizan la naturaleza predictiva de EvoMining.

En este punto conviene enfatizar cómo fue clave el papel de MIBiG [Medema, 2015] para esta versión de EvoMining 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 ningún 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. 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 sugieren la presencia de elementos con promiscuidad. También, EvoMining fue utilizado para ubicar a los homólogos de familias expandidas que probablemente forman parte de BGC. Estos últimos seguramente producen nuevos productos naturales, razón por la cual, abordaremos el estudio de la conservación de estos BGC a partir de estas enzimas en el siguiente capítulo. Se mostró también que el metabolismo se puede expandir a partir de genes de metabolismo conservado considerando tanto el metabolismo *shell*, como el metabolismo *core*. Se ilustró con el ejemplo de escitonemina que en un BGC se encuentran genes provenientes de la expansión de distintas familias de enzimas, que una misma expansión puede ser utilizada para producir distintos análogos con estructuras similares y que los patrones de expansión son específicos de cada familia y linaje. Queda por explorar los

patrones de neofuncionalización de forma exhaustiva con todas las familias del *shell genome,* así como el determinar de los BGC conocidos en MIBiG qué familias del metabolismo conservado fueron las que les dieron origen. En el siguiente capítulo retomaremos el ejemplo de *tauD* para describir el descubrimiento de 3 nuevos BGC cuyo producto natural fue sido demostrado por colaboradores.

Capítulo 3

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

En este capítulo presento CORASON, *CORe Analysis of Syntenic Orthologs to prioritize Natural products biosynthetic gene clusters*, una herramienta para explorar la diversidad en el contenido de los *Clústeres* de Genes Biosintéticos (BGC por sus siglas en inglés) así como su distribución en un linaje proporcionado por el usuario. CORASON es una herramienta en línea de comandos para organizar filogenéticamente la variación de una familia de clústeres, Figura 3.1. Ya que sabemos que hay variantes de BGC que producen productos naturales similares dentro de los linajes genómicos. Otros métodos de minado de genomas han acelerado el descubrimiento de nuevos BGC. Casi cada nueva bacteria que es secuenciada aporta alguna novedad al pangenoma bacteriano conocido. Una fracción de estos genes descubiertos formará parte de variantes de BGC previamente conocidos aportando diversidad a las familias de clústeres biosintéticos. La diversidad genética que existe en las familias de BGC está relacionada con cambios moleculares, incluso pequeñas variantes en un metabolito pueden ocasionar diferencias en su función biológica por lo que es de interés identificar análogos de productos naturales tanto como entender su evolución. Se diseñó CORASON que identifica el *core* génico de un BGC y genera una visualización de las variantes de la familia organizadas filogenéticamente.

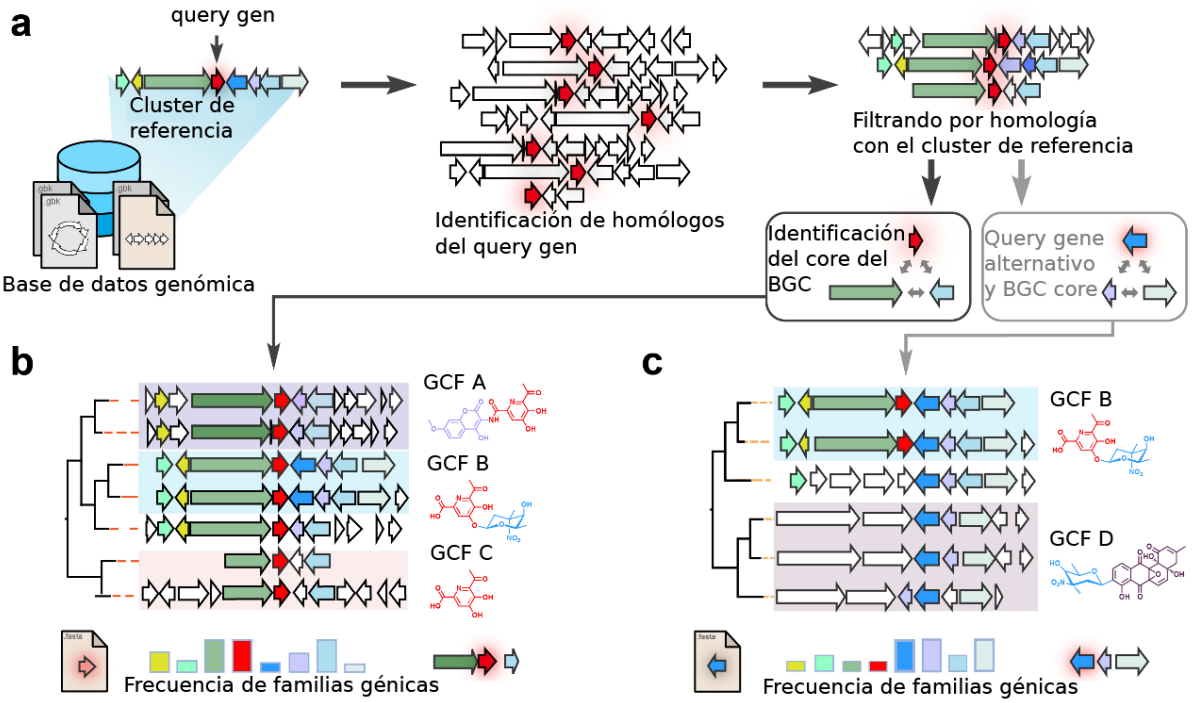

Figura 3.1 La herramienta CORASON localiza familias de *clusters* biosintéticos en un linaje genómico partiendo de un *cluster* y un gen de referencia. Todos los contextos genómicos que contengan ese gen y algún otro gen del cluster de referencia serán encontrados en el linaje seleccionado por el usuario.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

3.1. Algoritmo y características de CORASON

La herramienta CORASON, como se muestra en la Figura 3.1 localiza familias de BGC en un linaje genómico partiendo de un *clúster* y un gen de referencia. Para ello el usuario debe proveer una base de datos de genomas (Genome DB) de un linaje definido con las mismas características y precauciones descritas para EvoMining el capítulo anterior. También debe poner como entrada un BGC y un gen de referencia que sea parte de ese BGC. El algoritmo de CORASON busca los

homólogos de ese gen en cada genoma de la base de datos y luego también detecta a los otros genes del BGC de referencia si es que se encuentran en la vecindad del gen de referencia. Esto lo repite para todas las secuencias del linaje seleccionado por el usuario. Posteriormente, CORASON identifica el *core* génico de la familia con el algoritmo de Orthocore. La información del *core* se utiliza para organizar filogenéticamente todos los miembros de la familia del BGC. Ya que el *core* génico está relacionado con el *core* de la molécula, CORASON permite detectar en la parte variable del BGC a las enzimas accesorias que producen ornamentos i.e. variantes moleculares. Los resultados son un árbol en formato Newick junto con una visualización que permite al mismo tiempo apreciar variación a nivel de presencia-ausencia de genes entre miembros de una familia de BGC, así como la divergencia a nivel de secuencia entre genes conservados entre una variante y el BGC de referencia, mismo que se distingue con un código de color.

CORASON fue diseñado con las siguientes características: (i) Se implementa en una interfaz de línea de comando simple. (ii) Identificación de todos los homólogos de un gen de referencia en una base de datos de genomas. (iii) Identificación del *core* génico del BGC. (iv) Reconstrucción filogenética de la familia de BGC utilizando la información del *core*. (v) Salida visual en formato SVG, que muestra tanto la anotación funcional de los genes como como la distancia respecto a sus ortólogos del *clúster* de referencia.

3.2. Las familias de BGC son variantes del BGC de referencia.

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

Así pues, cuando se habla de un BGC no se debe pensar que este es igual en todos los organismos del linaje, es decir que tiene el mismo contenido génico. Hay variación tanto a nivel de contenido

génico como a nivel de secuencia entre los genes ortólogos. Esta variación produce la promiscuidad de producto, una familia de BGC produce distintos productos a partir de precursores similares. En este trabajo tomamos como BGC de referencia a los anotados en MIBiG, que son de los que se tienen productos reportados con datos experimentales. Para CORASON consideramos como parte de la familia del BGC todas las variantes de BGC que contengan al menos dos genes en común, uno de referencia seleccionado por el usuario y otro cualquiera, pero común con el BGC de referencia.

Como ejemplo de familias de BGC conviene pensar en analogía con los operones. Un operón es el conjunto de genes dedicados a la síntesis de un mismo proceso que regulan de forma coordinada su propia expresión, estos genes suelen ser contiguos y transcribirse simultáneamente desde un solo promotor. Ejemplos de estos operones son los que producen los aminoácidos histidina y triptófano. Estos “BGC” hace millones de años fueron posiblemente parte del metabolismo especializado y debido a su éxito se fijaron en lo que ahora vemos como BGC conservados en ciertos linajes genómicos. En estos casos las fronteras son claras y la variación génica es poca. Aun así, existen otros ejemplos donde puede constatarse variación en las rutas de síntesis de mecanismos centrales de metabolismo procariota. En el caso de histidina y triptófano, como hablaremos en el siguiente capítulo, es la variación a nivel de secuencia y no tanto a nivel de composición génica la que produce diversidad de producto. En oposición a los BGC u operones de metabolismo conservado, están los BGC del metabolismo especializado, así como se encuentran familias con mucha variación génica pueden encontrarse otras muy conservadas.

En las siguientes secciones presento cuatro ejemplos que ilustran como CORASON puede ser utilizado tanto para priorizar nuevos BGC como para ligar la variación génica de familias de BGC con diversidad estructural. Asimismo, se describe cómo este algoritmo es una expansión de las habilidades de EvoMining para encontrar enzimas en el proceso de diversificación funcional. En resumen, CORASON encuentra familias de BGC, y presenta una rápida visualización de sus variantes. Con esta herramienta podemos expandir nuestro conocimiento sobre nuevas variantes químicas mediante la genómica comparativa. CORASON está disponible en su contenedor de Docker en GitHub https://github.com/nselem/corason.

3.3. Aplicaciones de CORASON en Actinobacteria y *Pseudomonas*

En el capítulo anterior las variantes del BGC escitonemina en Cianobacteria fueron encontradas al aplicar CORASON a dicho linaje. En este capítulo presento ejemplos del uso de CORASON para investigar los patrones de conservación/variación en familias de BGC en los linajes Actinobacteria y *Pseudomonas*. Primero, las versiones iniciales de CORASON fueron usadas junto con cromatografía y espectrometría de masas (LC-MS) para estudiar la ecología y evolución de los sideróforos de tipo desferroxiaminas en Actinobacteria [Cruz-Morales, 2017]. En un segundo ejemplo, CORASON fue utilizado para investigar la existencia de variantes en Actinobacteria del clúster productor de arsenolípidos que se encuentra en *Streptomyces lividans*. Finalmente, un tercer ejemplo es presentado: el de los contextos genómicos del metabolismo secundario de *tauD*. Esta enzima proviene de una di oxigenasa involucrada en el metabolismo de taurina en su papel de enzima del metabolismo conservado. En Actinobacteria *tauD* es parte de 15 BGC reportados en MIBiG. Aunque en *Pseudomonas* no existen BGC reportados EvoMining predice expansiones en este linaje genómico y CORASON predice conservación de los contextos genómicos de estas copias extra. Estas vecindades genómicas guardan cierta similitud con variantes de los BGC conocidos en Actinobacteria.

3.4. Detectamos variantes de sideróforos en actinobacterias de vida libre

El hierro es necesario en el metabolismo de muchos seres vivos. Para poder utilizarlo las bacterias han desarrollado moléculas captadoras de hierro llamadas sideróforos. Un ejemplo de sideróforo es la molécula de desferroxiamina sintetizada por los genes *des* en Actinobacteria. Usamos CORASON para identificar variantes del clúster biosintético de desferroxiamina en actinobacterias de cuatro ciénegas (Figura 3.2). En particular, encontramos una variante del BGC que se diferencia del clúster reportado por la ganancia de un miembro de la familia penicilina-amidasa. Al gen que codifica esa función se le llamó *desG* y es responsable de la arilación de desferroxiaminas en actinomicetos acuáticos [Cruz-Morales, 2017]. Ya fue verificado que la variación génica está ligada con la variación molecular.

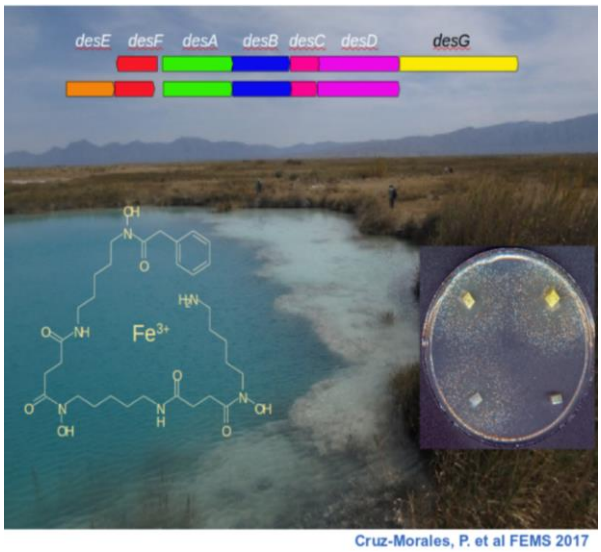

Figura 3.2 Variantes del cluster biosintético de desferroxiamina fueron identificadas por CORASON en Actinobacterias de cuatro ciénegas. Una variante del BGC de desferroxiamina fue identificada. Esta variante se diferencia del cluster reportadopor la ganancia de una penicilin amidasa. A este gen extra se le llamó *desG*.

3.5. Un BGC de metabolismo de arsénico tiene variantes que conforman una familia de BGC

Analizamos los patrones de expansión-reclutamiento de la 3-fosfoshikimato-1-carboxivinyl transferasa (AroA) en el linaje de *Streptomyces* usando EvoMining. Se encontró un cambio de función en esta familia tal que en una de las ramas de expansiones se detectó específicamente una arsenoenol piruvato sintasa. Estudios de mutagénesis y de expresión diferencial génica en presencia de arsénico confirmaron que en *Streptomyces coelicolor* y en *Streptomyces lividans* esta enzima pertenece a un BGC que sintetiza arsenolípidos [Cruz-Morales, 2016]. Mediante curación manual se descubrió por la diversidad en las secuencias de aminoácidos que el contexto genómico de AroA en estos organismos incluye una enzima PKS localizada a seis genes de distancia. antiSMASH predice los BGC tipo PKS de estas enzimas, pero no incluye en ellos a la arsenoenol piruvato sintasa. El valor de EvoMining fue descubrir que esta copia extra de AroA estaba dedicada al metabolismo especializado y por tanto bien podía tener su propio BGC o dada la cercanía con las PKS podía ser parte de estos PKS-BGC. Nos preguntamos si la diversidad de secuencia a nivel de BGC, si se encuentran variantes con distintos patrones de presencia/ausencia en los genes que componen al

BGC de *S coelicolor*. O más aun, quedaba por investigar si el BGC tenía cierto grado de conservación o era exclusivo de estos dos organismos *S. coelicolor y S. lividans*

El contexto genómico conservado en las expansiones de AroA puede apreciarse en la Figura 3.3. La arsenoenol piruvato sintasa fue descubierta en un árbol de EvoMining como parte de una rama de expansiones de AroA en Actinobacteria. El contexto genómico de la arsenoenol piruvato sintasa de *S. coelicolor* tiene un *core conservado* en Actinobacteria. Este BGC está dedicado a la síntesis de metabolitos secundarios de tipo arsenolípidos. Primero se identificaron en otras secuencias genómicas contextos que contengan el homólogo de AroA y algún otro gen de su vecindad en *S. coelicolor*. A continuación, se ordenaron manualmente los contextos obtenidos y se pudo identificar al menos cuatro diferentes subclases de BGC. La primera clase mostrada en un rectángulo morado no contiene PKS o NRPS, la segunda enmarcada en verde contiene una PKS-NRPS híbrida. Los otros dos subgrupos sólo contienen una PKS, en un caso está río arriba y en otro río abajo de AroA. La tercera clase incluye una PKS a la izquierda y a más de cinco genes de distancia de la arsenoenol piruvato sintasa; finalmente la última clase contiene una PKS a la derecha a sólo un gen de distancia de esta enzima. Así, se muestra que existen variantes de un BGC que contienen un *core* común, presumiblemente pueden producir variantes de arseno compuestos.

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

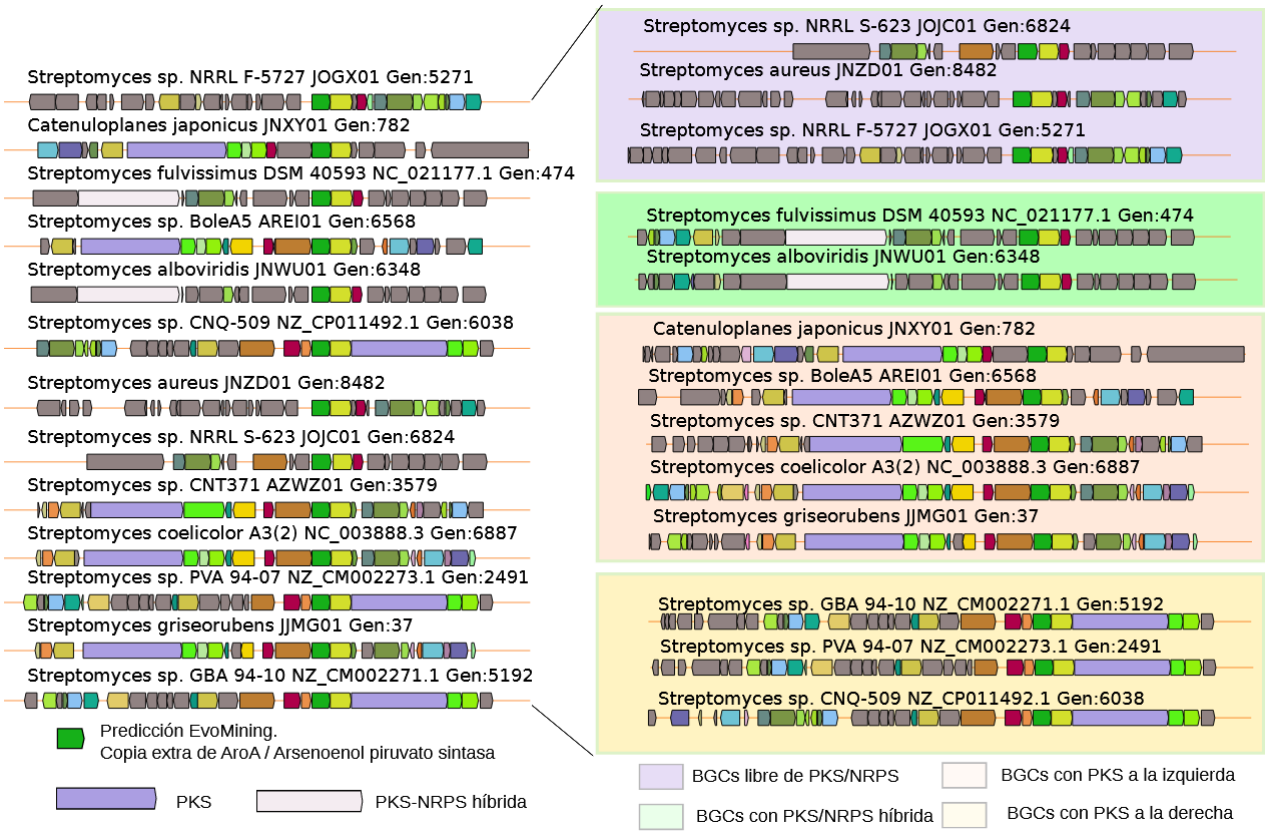

Figura 3.3 El contexto genómico conservado en las expansiones de AroA sugiere que es una familia de BGC con promiscuidad de productos. El contexto genómico de la arsenoenol piruvato sintasa de *S. coelicolor* tiene un *core* conservado en Actinobacteria que produce arsenolípidos. Se identificaron cuatro clases de BGCs. La primera clase en morado no contiene PKS o NRPS, la segunda enmarcada en verde contiene una PKS-NRPS híbrida. La tercera clase incluye una PKS a la izquierda ya más de cinco genes de distancia de la Arsenoenol piruvato sintasa y finalmente la última clase contiene una PKS a la derecha y a sólo un gen de distancia de esta enzima.

3.6. Combinamos CORASON y BiG-SCAPE para mejorar la clasificación de BGC y logramos predecir dos nuevos compuestos de la familia rimosamida - detoxina que fueron caracterizados experimentalmente

BIG-SCAPE es una herramienta bioinformática para clasificar un conjunto de BGC en familias de acuerdo con el contenido, conservación y distribución de sus dominios [Navarro-Munoz, 2018]. Un

dominio es una región conservada de una proteína que evoluciona y funciona independientemente del resto de la proteína. La identificación de dominios es particularmente importante en los BGC porque las modificaciones químicas que catalizan los genes de los BGC frecuentemente están codificadas en dominios que pueden o no ser parte del mismo gen conservado pero que de cualquier forma repiten la misma reacción. Es decir, un dominio con la función conservada puede realizar el mismo paso de una ruta biosintética formando parte de genes que no son homólogos. BIG-SCAPE usa la conservación de dominios (no de genes) para agrupar BGC en diferentes familias de *clústeres*, sin embargo, no permite una visualización eficiente su evolución. Por ello se utilizó una parte del algoritmo de CORASON que al proporcionar un algoritmo para ordenar la diversidad dentro de la familia usando el *core* de dominios conservados Figura 3.4. Así se consigue proponer la filogenia de los BGC y en ocasiones conectar mediante la evolución a familias de BGC aparentemente separadas por BiG-SCAPE.

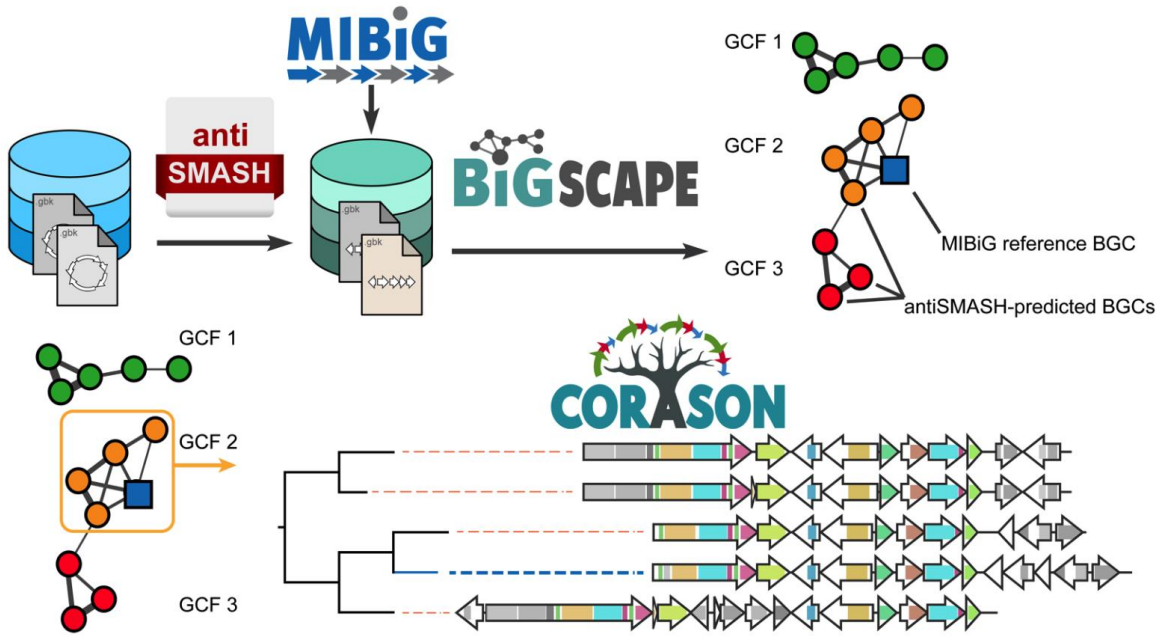

Figura 3.4 CORASON permite ordenar visualizar las familias de genes propuestas por BIG SCAPE. BIG SCAPE toma BGC que pueden provenir de MIBiG, de AntiSMASH y los clasifica en familias (GCF1, GCF2 y GCF3) que comparten dominios conservados. Con CORASON encontramos el *core* de dominios para hacer la filogenia de los BGC con ellos y visualizar su evolución.

3.6.1 Identificamos nuevos productos variantes de la familia de BGC Rimosamida - Detoxina en Actinobacteria integrando BiG SCAPE y CORASON

Ya que el TauD fue sugerida por un análisis de EvoMining como una familia con expansiones reclutadas a productos naturales de distintas subfamilias en Actinobacterias, en esta sección utilizamos CORASON y BIG-SCAPE para analizar una base de datos de miles de genomas. Con estas herramientas se organizó la diversidad biosintética de las familias de BGC detoxina y rimosamida [Mcclure, 2016]. El análisis reveló diversidad tanto en los géneros de los organismos que contienen a esta familia, y en la composición genética del BGC. Entre los géneros con alguna variante del BGC están *Amycolatopsis*, *Streptomyces*. El *core conservado* de los BGC detoxina y rimosamida está compuesto por una NRPS, una NRPS/PKS híbrida, y un homólogo de *tauD*. En *E. coli* tauD se encuentra en el operón *tauABCD*. La ruta de síntesis de rimosamida difiere de la de detoxina porque tiene una NRPS adicional, que codifica para una modificación del *core* de molécula detoxina/rimosamida con isobutirato y glicina.

El hecho de que el gen *tauD* estuviera presente en todos los miembros de la familia captó nuestra atención Figura 3.5. El TauD pertenece a la superfamilia de enzimas hidroxilasas dependientes de Fe (II)/α-cetoglutarate. En particular *tauD* codifica una taurina dioxigenasa dependiente de α-ketoglutarato involucrada en la asimilación de sulfito por la liberación oxigenolítica del aminoácido taurina. Interesantemente, esta familia también está presente en linajes como hongos, bacterias y plantas. Dichas enzimas catalizan hidroxilaciones, desaturaciones, expansiones y formaciones de anillos entre otras transformaciones químicas. A la fecha, el rol de TauD en la biosíntesis de los metabolitos detoxina y rimosamida aún es desconocido, se ha sugerido que es responsable de la oxidación de la prolina observada en algunos análogos.

Para identificar variantes de los BGC relacionados a detoxina y rimosamida dentro de la rama de metabolismo especializado los 1175 BGC que contenían un homólogo de *tauD* se pasaron por un análisis combinado de BiG-SCAPE/CORASON. Se usó *tauD* como gen de referencia en CORASON ya que es el único gen miembro del ‘BGC core’ que está presente en todos los genomas. Es importante notar que el *core* del BGC podría contener hasta 3 genes porque la NRPS, y la NRPS-PKS híbrida quedan fuera en este ejemplo debido a que algunos genomas no están completamente secuenciados y justamente hay huecos a los extremos de los contigs que contienen estos BGC. Este es el caso de los organismos *Streptomyces humi*, *Streptomyces spectabilis* y *Amycolatopsis vancoresmycina*.

3.6.2. CORASON sugiere que las familias detoxina y rimosamida pertenecen a un amplio grupo de familias dedicadas a la síntesis de péptidos.

El análisis de CORASON reveló que las familias de los BGC detoxina y rimosamida identificados en BiG-SCAPE eran parte de una familia expandida de BGC de biosíntesis de péptidos que incluía clados inexplorados del phylum Actinobacteria (Figura 3.5). La organización filogenética de los BGC provista por CORASON, reveló familias de BGC que fueron omitidas debido al umbral utilizado por el algoritmo de agrupamiento de BiG-SCAPE. Esto debido a la cercanía de genes biosintéticos detectados por antiSMASH como suficientemente diferentes como para ser clasificados en otras familias.

Hipotetizamos que las moléculas de la familia detoxina codificadas por los BGC en los clados inexplorados contendrán cierta novedad química relacionada con las variaciones genéticas. Afortunadamente, 40 de las 152 cepas identificadas como portadoras de un BGC estaban representadas en nuestros datos metabolómicos de 363–cepas por LC-MS/MS. Los análisis de redes moleculares de estos datos indicaron la presencia de tres detoxina conocidas, cuatro rimosamida conocidas y otras 103 variantes de detoxina o rimosamida, el basto universo químico sugerido por el análisis BiG-SCAPE/CORASON.

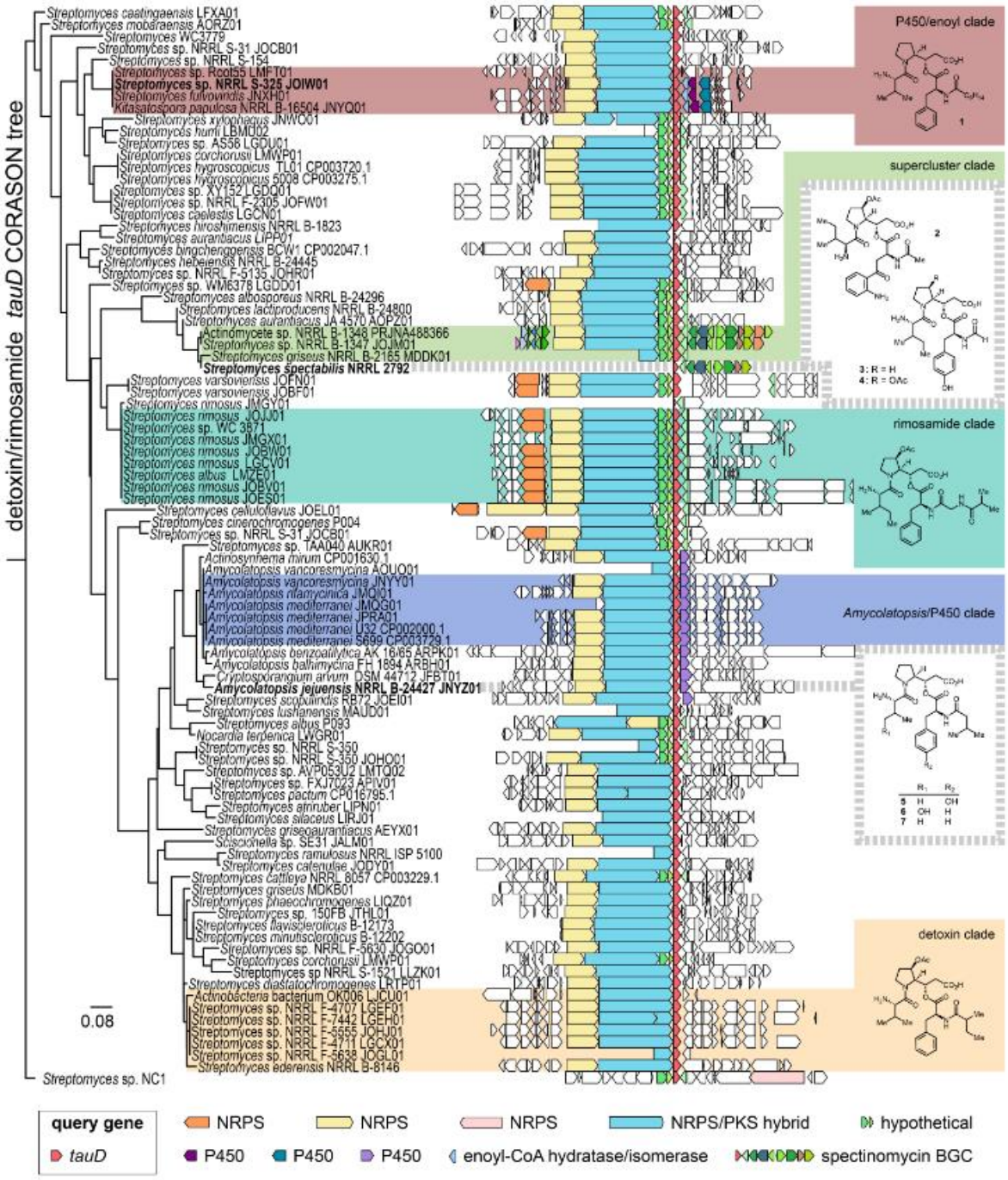

Figura 3.5 Visualización de la Familia Expandida de BGC de TauD detectada porCORASON y sus 5 subfamilias de BGC detectadas por BIG SCAPE. A la derechase muestra la molécula producida por cada subfamilia.

3.6.3. El árbol de CORASON de las familias rimosamida/detoxina muestra casos de diversidad génica que correlaciona con novedad química.

Tres de los clados que codifican para detoxina BGC fueron identificados por BiG-SCAPE dentro del árbol de CORASON capturaron nuestro interés (ver las cajas coloreadas de la Figura 3.5). En esta sección se describe el trabajo experimental realizado por Michael Mulloney del grupo de colaboradores de West Chicago, de los tres clados y específicamente de los organismos que seleccioné como candidatos a presentar diversidad química.

3.6.3.1 El clado P450/enoil agrega una heptanamida al core molecular detoxina/rimosamida

El primer clado es el 'P450/enoyl clade' contiene genes como el citocromo P450 y una enoil-CoA hidratasa/isomerasa dentro de cada uno de sus BGC. Este clado está marcado en rojo en la Figura 3.5. Análisis de datos por tándem MS de extractos de *Streptomyces* sp. NRRL S-325, que se encuentra dentro de este clado, llevó al descubrimiento de la detoxina S1. Este nuevo análogo contiene una cadena lateral de heptanamida, una estructura única entre las detoxinas y rimosamidas cuya instalación posiblemente depende de la enzima enoil - CoA hidratasa/isomerasa.

3.6.3.2 El super clado espectinomicina / detoxina - rimosamida produce al menos cinco variantes de detoxina.

El segundo clado de interés, fue nombrado el 'supercluster clade' (Figura 3.5, en verde claro), comprende los BGC con genes de detoxina adyacentes al clúster que produce espectinomicina. El clúster de espectinomicina (MIBiG BGC0000715) contiene en su periferia al gen *tauD* como se muestra en la línea gris punteada de la Figura 3.5. La secuencia del clúster de espectinomicina depositada en MIBiG es la única secuencia disponible de *Streptomyces spectabilis* NRRL 2792. Como no se sabe que tauD participe en la síntesis de espectinomicina se hipotetizó que pueden existir los genes del clúster de detoxina al lado de los genes del BGC de espectinomicina en *S. spectabilis* NRRL 2792. Adquirimos esta cepa para determinar si el análisis de CORASON podía ayudar a la predicción de detoxina basado solamente en la presencia del gen de referencia, pero en ausencia completa de la secuencia del BGC de detoxina. El análisis en tándem de espectrometría de masas de extracto *S. spectabilis* NRRL 2792 reveló la presencia de cinco compuestos tipo detoxina.

Los tiempos de retención de iones y los patrones de fragmentación de los últimos dos compuestos también fueron observados en extractos de *Streptomyces* sp. NRRL B-1347 parte del clado del super clúster. Esto confirma la habilidad de CORASON para guiar un descubrimiento mediante la utilización de la filogenia a pesar de lo limitado de los datos en la cepa NRRL-2792. El análisis de LC-MS de cultivos de NRRL-2792 suplementados con isótopos estables etiquetados de aminoácidos corroboraron las predicciones estructurales basadas en los análisis de la cepa cercana *Streptomyces* sp. NRRL B-1347. Aunque los datos de MS fueron insuficientes para desenmascarar esta estructura, el compuesto 2 fue producido por *S. spectabilis* NRRL 2792 en suficiente abundancia para el aislamiento y la elucidación estructural por NMR.

3.6.3.3. El clado *Amycolatopsis P450* produce cinco variantes de detoxina.

El tercer clado que se estudió es al que pertenecen las familias rimosamida - detoxina, contiene BGC casi enteramente provenientes del género *Amycolatopsis*. Este clado está marcado en morado en la Figura 3.5. Este clado de BGC también contienen un gen P450 único entre los BGC del árbol, así que fue llamado el clado '*Amycolatopsis* /P450 clade'. Aunque no se contaba con datos metabolómicos de las cepas del clado de BGC definidos por BiG-SCAPE como una *Gene Cluster Family* (GCF), la visualización filogenética de CORASON permitió la selección de una cepa de *Amycolatopsis* de la que se tenían datos metabolómicos con un BGC muy similar, y que también contiene el gen P450 (Figura 3.5, línea gris cerca del clado *Amycolatopsis*/P450). Los análisis de los datos de tándem MS de extracto fermentado de *Amycolatopsis jejuensis* NRRL B-24427 reveló isómeros de detoxinas P1 que contienen tirosina, P2 mostrando fenilalanina y una valina hidroxilada, así como la detoxina P3, un análogo cercano libre de hidroxilación. Además, se consiguió validar la asignación de aminoácidos observados en los patrones de fragmentación de MS/MS mediante el uso de experimentos de incorporación de isótopos estables etiquetados de aminoácidos.

3.7 CORASON permitió explorar la promiscuidad de familias de BGC de enzimas divergentes de metabolismo central.

Nuestros resultados ilustran como BiG-SCAPE puede identificar conjuntos de BGC relacionados, en un gran número de secuencias de genomas. Además, al usar CORASON para reconstruir las filogenias de BGC para ordenar visualmente la evolución de un clúster biosintético y su diversidad proveen herramientas poderosas para el descubrimiento de nuevos clados de BGC que codifican en

consecuencia para nueva química. Respecto a los BGC detoxina/rimosamida, CORASON mostró habilidad para ayudar a minar bases de datos genómicas y descubrir siete nuevas detoxinas. Específicamente, la organización de las variantes de los BGC facilitó la identificación de las correspondientes variaciones en la estructura química -la presencia de un enoil-CoA hidratasa/isomerasa corresponde a la familia de amida ácido graso detoxina S1 y la presencia de un gen P450 corresponde a la presencia de hidroxilaciones en detoxinas P1–P3.

Estos resultados demuestran una forma alternativa en la que la minería genómica permite identificar cómo se genera la diversidad de funciones ahora a nivel de BGC. Al identificar cómo es la evolución de los genes conservados de los *clústeres* y diferenciarlos de los cambios en los genes accesorios de los rodean fue posible guiar el descubrimiento de nuevos productos naturales. También este análisis nos permitió observar que existen distintos niveles de variación que generan diversidad química y que pueden ser entendidos como promiscuidad. Dicha variación se puede originar debido a la divergencia a nivel de secuencia de dominios proteicos pasando por variación a nivel de la variación entre genes de la misma familia enzimática y llega hasta grupos de genes que varían en cuanto a presencia/ausencia de algunos los genes que los componen. Por si esto fuera poco, además detectamos que hay familias de BGC que a su vez forman superfamilias de BGC o que están compuestas por subfamilias de BGC y cuya diversidad a todos esos niveles correlaciona con la diversidad de los productos naturales que producen.

Capítulo 4

La familia PriA/HisA

PriA es la familia de enzimas de Actinobacteria homóloga a la familia HisA existente en Enterobacteria, Cianobacteria, *Pseudomonas* y Archaea. Según las definiciones de este trabajo PriA es una familia promiscua ya que se ha encontrado que varios miembros de PriA *in vitro* tienen la capacidad de catalizar tanto la reacción correspondiente a su homólogo HisA en la síntesis de histidina como la isomerización que cataliza TrpF cuya reacción está involucrada en la producción de triptófano. Es decir, se ha mostrado mediante cinéticas enzimáticas y análisis de complementación genética que en varias Actinobacterias PriA participa a la vez en las rutas de síntesis de histidina y triptófano mediante la isomerización de un anillo de cinco carbonos (Figura 4.1). Los primeros miembros caracterizados como promiscuos en esta familia fueron *Streptomyces coelicolor* y *Mycobacterium tuberculosis* [Barona-Gomez, 2003]. La mayoría de las actinobacterias han perdido el gen *trpF,* aunque su actividad es esencial. Además, conservan el resto del operón de triptófano, por lo que se cree que la promiscuidad de PriA está extendida en un gran subconjunto de Actinobacteria. En la ruta de histidina PriA isomeriza el sustrato ProFAR en PRFAR, realizando la función HisA. En la ruta de triptófano PriA lleva a cabo la isomerización de PRA en CdRP catalizando la reacción de TrpF en la ruta del triptófano. La vecindad genómica de PriA en *Streptomyces coelicolor* y en otras Actinobacterias contiene genes tanto del operón de histidina como del de triptófano.

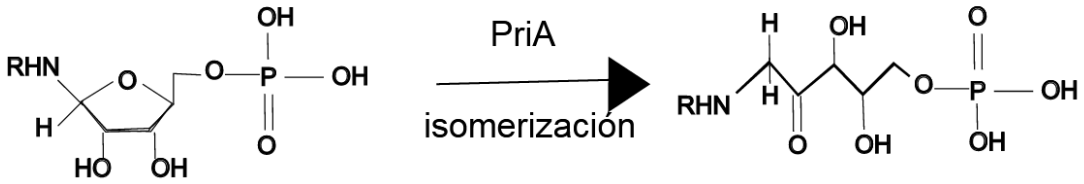

Figura 4.1 La reacción que cataliza PriA es una isomerización donde abre un anillo de cinco carbonos

En este capítulo usamos como modelo de una familia promiscua la familia de PriA/HisA porque además PriA ha mostrado un gradiente funcional en Actinobacteria. Esta variación divide a la familia en varias subfamilias según su capacidad catalítica en los sustratos ProFAR y PRA. La evolución de las subfamilias de PriA y sus funciones es compleja y muestra dinámicas distintas en cada linaje (ver introducción). Por ejemplo, incluye a los de miembros de PriB, la subfamilia ubicada en el género *Streptomyces* con baja capacidad de catálisis para la función TrpF. Varios *Streptomyces* con un ortólogo en la familia PriB, se diferencian de otras Actinobacterias en que contienen en su genoma un gen *trpF* localizado fuera del contexto genómico inmediato de los operones de histidina y triptófano. Otra subfamilia de PriA es subHisA, que ha perdido totalmente la actividad TrpF, existen miembros de subHisA en *Corynebacterium* y en *Actinomyces*. Finalmente, en Actinomyces también encontramos la subfamilia *subTrpF* que ha perdido la actividad de HisA.

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

4.1. La familia PriA cambió su función y sus patrones de promiscuidad en los cuatro linajes analizados

4.1.1. Las expansiones no son condición necesaria para la promiscuidad

Para explorar PriA en diversos linajes genómicos se utilizaron las herramientas EvoMining y CORASON descritas en los capítulos previos. Se investigaron las expansiones de la familia PriA en los linajes Actinobacteria, Cianobacteria, *Pseudomonas* y Archaea. En la Figura 4.2 se muestra el número promedio de copias por genoma en los linajes genómicos seleccionados. En Actinobacteria, donde se sabe que PriA es promiscua no se detectaron copias extra. Según EvoMining en Actinobacteria no hay expansiones, prácticamente todas

las copias son reconocidas como de metabolismo central (rojo o morado) aunque algunas PriA además son marcadas por antiSMASH como parte de algún clúster biosintético (morado). En cambio, en Cianobacteria, *Pseudomonas* y Archaea la figura muestra en negro las copias extra de las que no se conoce su destino metabólico. El caso de Archaea es llamativo porque las copias de metabolismo central llegan en promedio hasta .5 copias por genoma, es decir muchos genomas de Archaea no cuentan con una copia de PriA, y en cambio, contrario a Actinobacteria, un 50% de las copias es marcado en negro, es decir varios de los genomas que tienen al menos una copia de PriA en realidad tienen dos copias. Esta figura muestra que en Actinobacteria PriA constituye un ejemplo de familia promiscua mayoritariamente distribuida con una sola copia por genoma por lo que es evidente que para que una familia sea promiscua no es imperativo tener copias extra con marcas de reclutamiento en metabolismo especializado. Aunque las copias extra suelen ser una indicación de promiscuidad, no son una condición necesaria. Tanto EvoMining como CORASON mostraron que existen excepciones de organismos donde PriA tiene doble copia, tanto en Actinobacteria como en otros linajes genómicos.

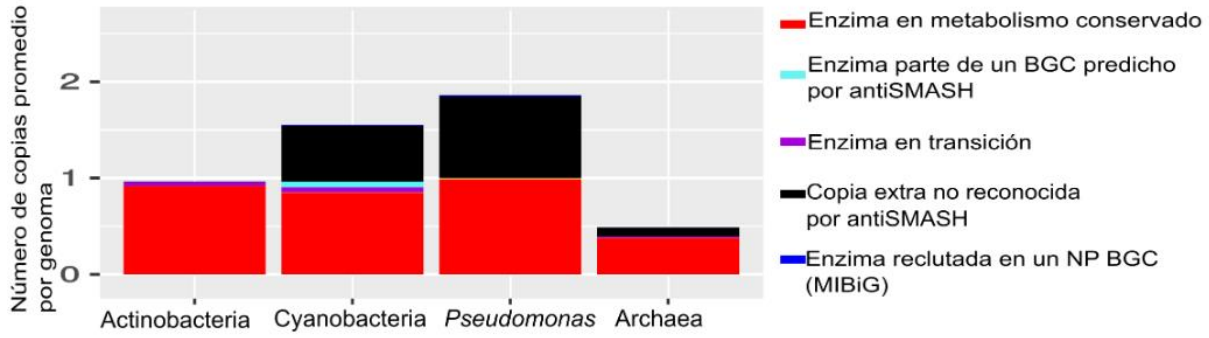

Figura 4.2 Número promedio de copias por genoma de PriA en Actinobacteria, Cianobacteria, *Pseudomonas* y Archaea. Los colores muestran el destino metabólico asignado a cada copia según EvoMining. En rojo están los BBH a las enzimas semilla de metabolismo central. En morado las enzimas de metabolismo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

4.1.2. Las expansiones condujeron a la creación de la subfamilia HisF

Después del conteo de número de copias promedio, se analizaron los árboles de PriA de EvoMining coloreados de acuerdo con el número de copias, Figura 4.3. En Actinobacteria la mayoría de las hojas son verdes mostrando que existe sólo una copia por genoma en ese organismo. Sin embargo, existen varias hojas de color amarillo, lo que indica que hay dos copias en algunos genomas. En Cianobacteria, *Pseudomonas* y Archaea en contraste con Actinobacteria, se muestran una mezcla entre organismos que poseen una (verdes) o dos copias (amarillos) de PriA. Sin embargo, al analizar detalladamente los árboles producidos por EvoMining en los distintos linajes, tanto en Cianobacteria como en *Pseudomonas* la copia extra está en una rama divergente y muy poblada del árbol de esta Familia Expandida, lo que indica que esas copias extra son en realidad miembros de otra subfamilia enzimática que aún guarda cierta similitud de secuencia con PriA. Esta segunda copia está en su mayoría anotada por RAST como imidazol glicerol fosfato sintasa ciclasa en ambos linajes. En Archaea, sin embargo, diversas especies de la clase Methanomicrobia sí tienen dos copias cercanas a PriA.

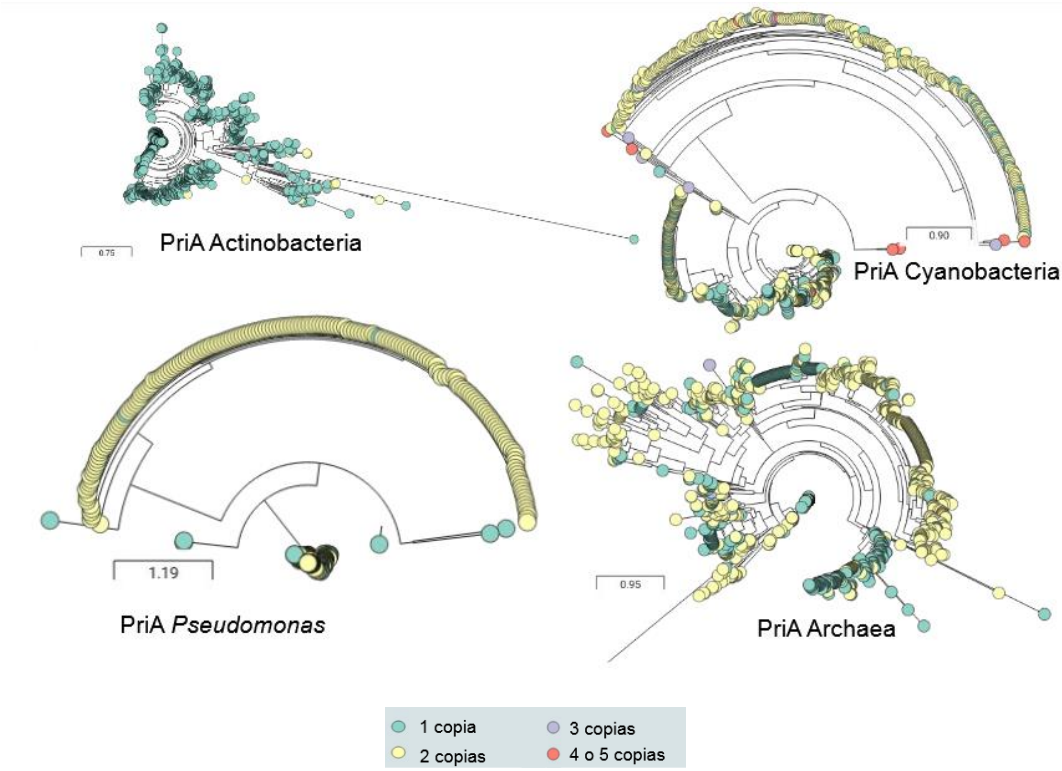

Figura 4.3 Número de copias de PriA en Actinobacteria, Cianobacteria, Pseu-domonas y Archaea. En Actinobacteria, donde se ha comprobado su promiscuidadenzimática la moda en el número de copias es uno (verde). En Cianobacteria yArchaea se observa una mezcla de hojas verdes y hojas amarillas, donde las ama-rillas son nodos que pertenecen a organismos con dos copias. También en estoslinajes algunos organismos poseeen tres copias (morado) y cuatro o cinco (rojo). EnPseudomonas la mayoría de los organismos muestran dos copias.

4.1.3. Cada linaje tiene distintos destinos metabólicos de PriA

Después de explorar cuáles organismos tienen expansiones de PriA, analizamos el posible destino metabólico de las copias extra de la familia como se muestra en la Figura 4.4. El árbol de Actinobacteria está poblado de hojas rojas, es decir de PriA dedicadas al metabolismo conservado, en este caso relacionado a las rutas de Histidina y Triptófano. Sin embargo, hay algunas hojas grises, como es el caso de los dos *Serenicoccus*. Es posible que estas PriA puedan tener funciones alternativas o ser otros eventos con cambio de promiscuidad. Además, en Actinobacteria la PriA de *Janibacter hoilley*, la rama más larga del árbol es muy divergente. Esto se debe a que existe una fusión de PriA con HisH. La fusión de genes bacterianos es un mecanismo común para formar proteínas multidominio [Pasek, 2006]. De hecho, dentro del operón de histidina existen fusiones reportadas HisNB y HisIE con proteínas bifuncionales en ciertas proteobacterias [Fani, 2005]. La fusión aquí presentada no parece ser un artefacto de anotación ya que hay otros genomas de *Janibacter* con secuencias de PriA más grandes que el promedio. Los árboles que se produjeron por EvoMining están disponibles para exploración interactiva en Microreact en los enlaces de la Tabla 4.1.

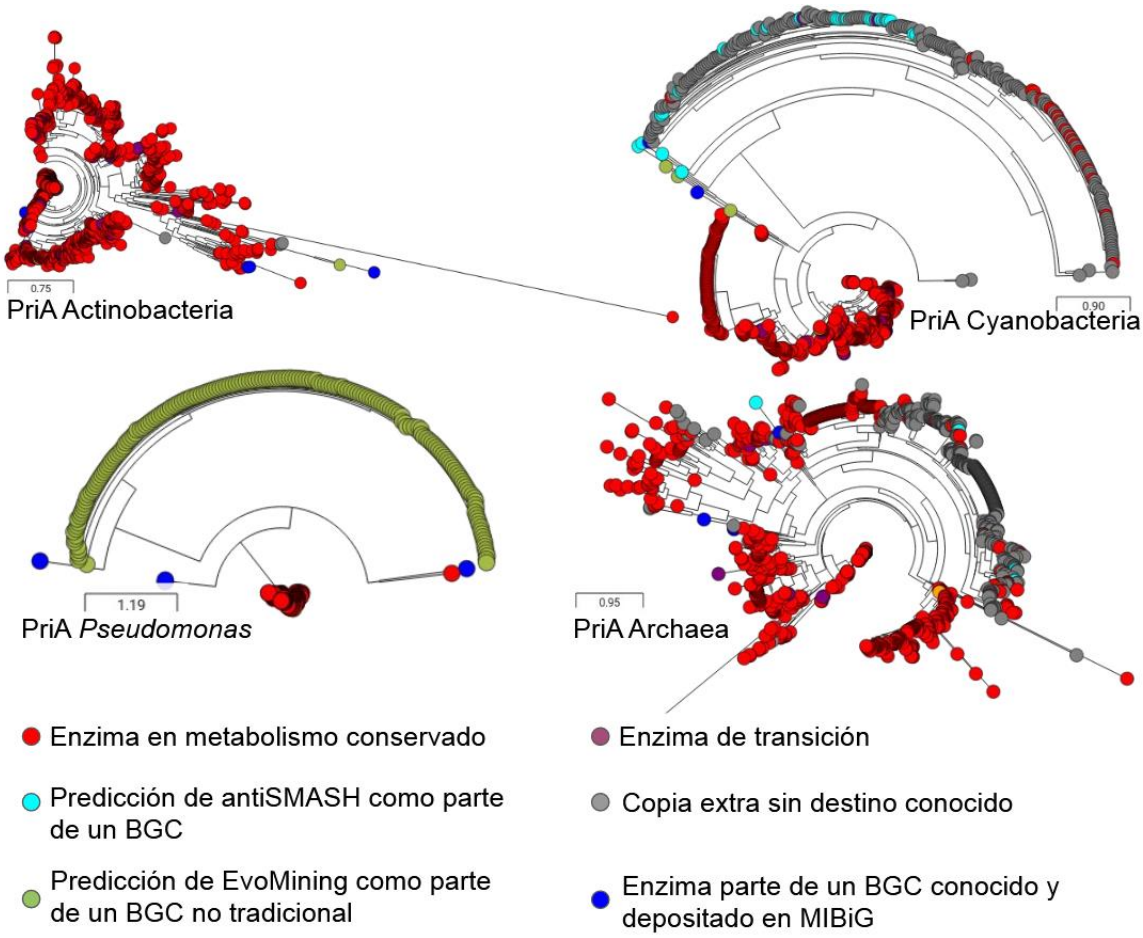

Figura 4.4 Árboles de destino metabólico de PriA en Actinobacteria, Cianobacteria, Pseudomonas y Archaea según EvoMining

Tabla 4.1 Árboles de EvoMining de PriA/HisA en MicroReact

LINAJE	ENLACE AL ÁRBOL DE EVOMINING EN MICROREACT
ACTINOBACTERIA	https://microreact.org/project/7g2IGfkv9
CIANOBACTERIA	https://microreact.org/project/qF6jWRMox
PSEUDOMONAS	https://microreact.org/project/ydff6DWqs
ARCHEA	https://microreact.org/project/lg-m9Cm6f

En Cianobacteria hay pocas predicciones de EvoMining, lo que se muestra en escasas hojas verdes que además no están localizadas cerca de su reclutamiento la HisA de saxitoxina (Figura 4.4, derecha-arriba), el BGC proveniente de Cianobacteria. Varias de las copias extra de Cianobacteria corresponden a HisF, la subunidad enzimática imidazol glicerol fosfato

sintasa parte también del operón de histidina. HisA proviene de una duplicación de un ancestro común de la mitad de su tamaño y HisF proviene de la duplicación de HisA [Fani, 1994 y 1997]. Por ello puede considerarse que EvoMining en efecto está encontrando una expansión de HisA que consiste en la familia HisF en el linaje Cianobacteria, lo mismo ocurre en *Pseudomonas*. En su momento HisF fue parte del *dispensable genome*, no todos los organismos poseían esta expansión, posteriormente, al parecer antes del último ancestro común de Archaea, Bacteria y Eucaria, la síntesis de histidina fue fijada [Fondi, 2009]. Este es un caso donde lo que alguna vez fue *dispensable genome* / metabolismo secundario, se ha vuelto *shell genome* / metabolismo conservado y donde se muestra que en estos linajes Cianobacteria y *Pseudomonas* EvoMining aún es capaz detectar esa marca de expansión. De hecho, se ha reconstruido una HisF ancestral con el objetivo de probar si retenía la actividad de HisA, es decir si poseía promiscuidad enzimática [Merkl, 2016]. LUCA-HisF mostró no ser promiscua para los sustratos de HisA ni tampoco para su similar el sustrato TrpF.

En *Pseudomonas* hay una gran población de predicciones de EvoMining, pero el árbol tiene similitudes con Cianobacteria, la rama divergente corresponde en su mayoría a copias de HisF. De hecho, el reclutamiento que hace que el árbol de *Pseudomonas* tenga una rama verde está anotado funcionalmente no como una HisA, sino como una HisF parte de un BGC que produce un lipopolisacárido en la Proteobacteria *Legionella pneumophila*. En Archaea, la rama central con una mezcla de hojas grises y rojas contiene también copias de HisF provenientes de genomas de *Sulfolobus* (Figura 4.4, izquierda-abajo). Al verificar en el árbol interactivo en Microreact (enlace de la Tabla 4.1) el número de copias presente en organismos con un homólogo clasificado como HisF se comprobó que los *Sulfolobus* sólo poseen una copia en este árbol. Sin embargo, tanto al explorar manualmente los genomas de *Sulfolobus*, como al revisar la literatura sobre metabolismo de histidina en Archaea se encontró que sí se han encontrado homólogos de PriA en genomas de *Sulfolobus* [Fondi, 2009]. Por tanto, este árbol sugiere que, aunque existe tanto PriA como su expansión HisF en el género *Sulfolobus*, la copia HisF es la más parecida a la semilla con que se generó este árbol. Además, la similitud de secuencia del homólogo de HisA no fue suficiente para recuperar copias de HisA en este experimento computacional. Algunas de las hojas grises del árbol sin destino metabólico conocido serán exploradas en la siguiente sección.

4.1.4. Homólogos de PriA han sido reclutados a clústeres de metabolismo especializado

Se encontraron algunos homólogos reclutados a BGC conocidos, marcados en azul en el árbol, que son miembros de la familia expandida PriA/HisA en estos linajes genómicos (Figura 4.4). La información de dichos BGC está listada en la Tabla 4.2. Entre ellos se encuentran dos toxinas de Cianobacteria [Moustafa, 2009], un lipopolisacárido producido por una Proteobacteria y un BGC productor de cloro-pentostatina producido en Actinobacteria [Gao, 2017].

Tabla 4.2 Reclutamientos de expansiones de PriA en MIBiG

COMPUESTO	ORGANISMO	ORGIEN	CLASE
PENTOSTATIN	*Actinomadurasp*. ATCC 39365	Actinobacteria	Otros
SAXITOXINA	*Cylindrospermopsis raciborskii* T3	Cianobacteria	Alcaloide
TOXINA	*Dolichospermum circinale* AWQC131C	Cianobacteria	Otros
LIPOPOLISACÁRIDO	*Legionella pneumophila*	Proteobacteria	Sacárido

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

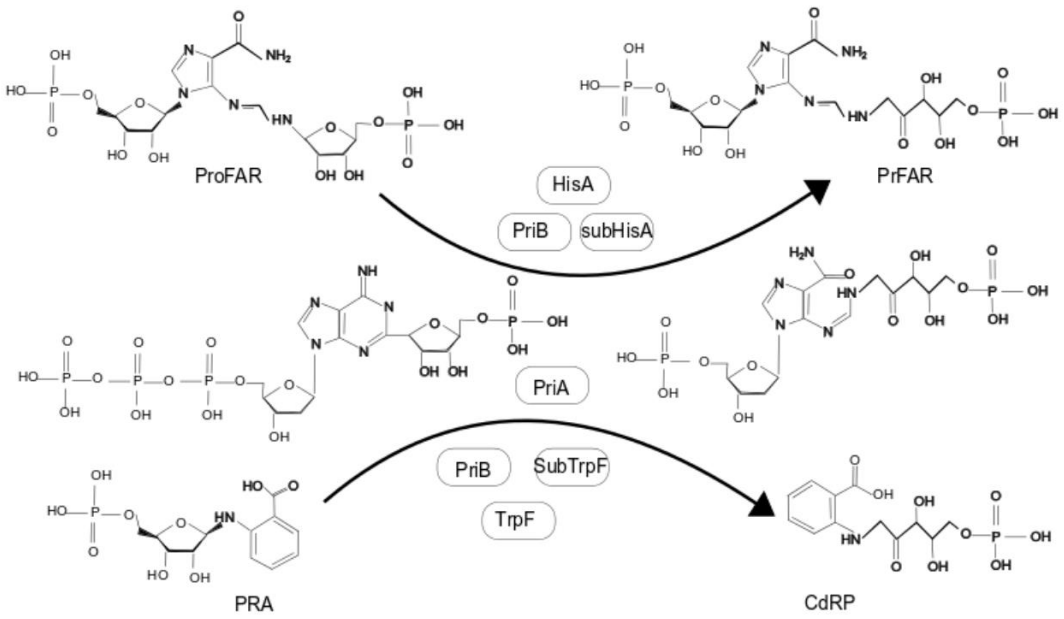

Figura 4.5 PriA participa en la síntesis del antibiótico ada en Actinomadura. Lossustratos nativos de Pria, ProFAR y PRA son isomerizados de manera muy similara un paso en la ruta de síntesis de ada

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

4.2.1. Contextos genómicos de HisA en Actinobacteria

En Actinobacteria, se observó que todos los *Streptomyces* tienen el clúster de PriA parcialmente conservado con respecto al BGC de *Streptomyces coelicolor*. Ejemplos de ello son *S. roseous*, *S. sviceus*, *S. sp C* y *S. Mg1* donde genes tanto de la ruta de histidina como de triptófano rodean a PriA. Otros como *S rimosus*, *S HmicA12* y *S. sp CT34* tienen los genes de triptófano más alejados (Figura 4.6). Como ya se describió en la sección de EvoMining, el único organismo de este género con una copia extra de PriA es *Streptomyces* CT34. Esta copia parece deberse a transferencia horizontal dado que su mejor hit en NCBI proviene de una *Lentzea*. Aun así, parece ser un homólogo lejano ya que tuvo 50% de identidad en 98% de cobertura con respecto a la copia de *Lentzea*. Otro caso interesante en Actinobacteria es

Actinomadura, ya que CORASON muestra que el clúster *ada* no está conservado en ellas (datos no mostrados). Además, también en Actinobacteria CORASON muestra en *Sporichthya polymorpha* DSM 43042 una PriA precedida por una NRPS, una enzima por excelencia de productos naturales. En otros organismos antiSMASH predice que PriA forma parte de clústeres putativos, por ejemplo, en *Modestobacter marinus* NC_0179551, *Geodermatophilus obscurus*, y en *Streptacidiphilus jeojiense*. Los contornos de los genes reconocidos por antiSMASH como parte de un BGC son marcados en azul en las figuras de CORASON.

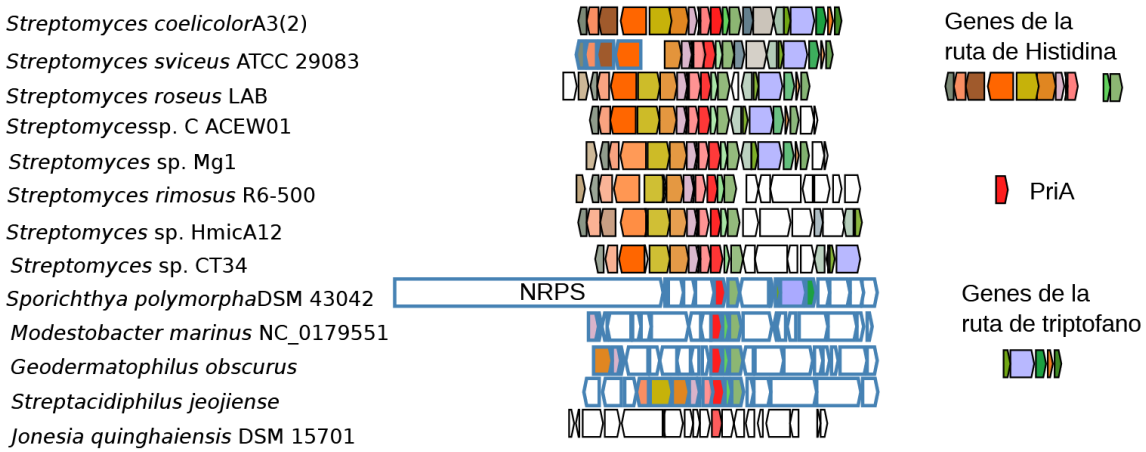

Figura 4.6 Contextos de PriA en Actinobacteria

4.2.2. Contextos genómicos de HisA en Archaea

En cuanto al contexto de HisA en Archaea la Figura 4.7 muestra que existen contextos como los de *Thermococcus kodakarensis KOD1* y *Thermococcus sp JCM 11816* donde *hisA* está rodeada de genes tanto de histidina como de triptófano. Sin embargo, esta configuración conjunta no es la generalidad. En Archaea, se conocen tres escenarios para la organización de los genes de síntesis de histidina. En el primero los genes de síntesis están dispersos en el genoma. El segundo escenario es la existencia de suboperones pequeños como *hisHAF*. Finalmente, el tercer escenario es el operón con la mayoría de los genes juntos, aunque no siempre en el mismo orden [Fondi, 2009].

En efecto en Euryarchaeota están presentes los tres escenarios, en *Archaeoglobales* y *Halobacteriales* los genes del operón *his* están distribuidos en el genoma, como muestran los contextos de *hisA* en *Halorubrum tebenquichense* y *Halonotius sp.* Sin ningún otro gen del operón cerca. En *Thermosarcinales* donde ya se conocía la existencia del suboperón *hisGBA*, se observó que éste se encuentra conservado en la mayoría de los miembros del orden como ejemplifica el contexto correspondiente a *Methanosarcina lacustris*. En *Thermococcales* se conocía la configuración *hisGDBHAFIEC* que es confirmada por los tres *Thermococcus* de la Figura 4.7, donde además se muestran dos contextos acompañados del su operón de triptófano *trpCDEGF*, resultando un contexto total de *hisGDBHAFIECXtrpCDEGF*. En el phylum Thaumarchaeota se habían estudiado sólo dos organismos y estos presentaban la configuración *hisGDCXBHAI*. En contraste en este trabajo encontramos en el orden *Nitrosopumilales* la configuración *hisDCXBHAFIE*, donde además la *hisX* es una fosfatasa homóloga a la *hisX* presente en *T. kodakarensis*. Este patrón es común en Thaumarchaeota, aunque también se muestra la existencia del suboperón *hisAF* en *Aigarchaeota,* así como la configuración de genes dispersos en el genoma en *Thaumarchaeota archaeon JGI OTU-1*. En Crenarchaeota *hisCGABFDEHI* está presente según la anotación funcional de RAST, aunque no todos los genes aparecen en color en la figura debido a la baja identidad de secuencia que conservan con respecto al contexto de referencia de *T. kodakarensis*. En *Pyrobaculum neutrophilum* está parcialmente conservado el operón con la configuración *hisCEDFBA*, nuevamente con algunos genes con baja identidad de secuencia respecto a *T. kodakarensis*. La configuración genes dispersos en el cromosoma en el phylum Crenarchaeota es mostrada en el contexto de *Ignicoccus islandicus*. Finalmente fue analizado el contexto de *hisA* en Bathyarchaeota un phylum donde no se contaba con información previa. En Bathyarchaeota se conserva como en Thaumarchaeota el ordenamiento *hisGDCXBHAI*, aunque existen otras configuraciones como muestra *Bathyarchaeota archaeon B25*.

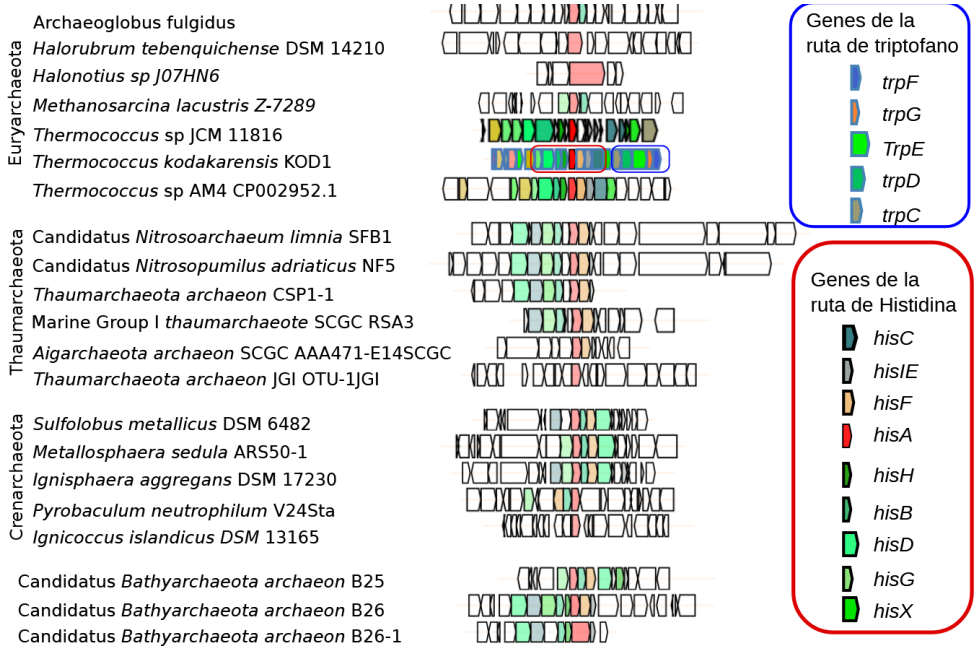

Figura 4.7 Contextos de PriA en Archea

Otra característica interesante de la visualización de los contextos de HisA en Archaea es que los tamaños de *hisA* son más variados que en Actinobacteria. Esta característica se debe tanto a variación intrínseca del tamaño de *hisA* como a diversas fusiones. Por ejemplo, en *Halonotius sp J07HN6 hisA* se encuentra fusionada con *flaJ* una proteína relacionada con la construcción de flagelos en Archaea. En Candidatus *Bathyarchaeota archaeon B26-1 hisA* se muestra fusionada con *hisF*. En conclusión, no encontramos una configuración conservada que prevalezca en todo Archaea. Esto puede deberse a que Archaea es un dominio, no un phylum como Actinobacteria, y por tanto hay una mayor distancia entre los organismos que la conforman.

4.2.3. Contextos genómicos de HisA en saxitoxina

Como se mostró en la sección de análisis de expansiones de HisA con EvoMining en Cianobacteria, *hisA* aparece dentro del clúster reportado como productor de saxitoxina en MIBiG. El contexto que rodea al reclutamiento de *hisA* en el clúster de saxitoxina no está conservado en Cianobacteria según la visualización de contextos producida por CORASON. Tomando como semilla la secuencia de HisA del BGC de saxitoxina y como referencia el

clúster de saxitoxina, se muestra que, si bien los reguladores del BGC saxitoxina sí están conservados en la vecindad genómica de HisA en otros organismos, este no es el caso para las enzimas biosintéticas del *clúster*. Además, en el lado izquierdo de la Figura 4.8 se muestra que la HisA del BGC saxitoxina no está ubicada en una rama de PriA divergente, al contrario, está embebida en la parte más conservada. Por estas razones es posible que PriA esté en la orilla del BGC de saxitoxina y más bien no participe en la síntesis de este compuesto o bien que mantenga su función primaria y a la vez tenga una función aún desconocida teniendo un papel en la regulación de la producción de saxitoxina.

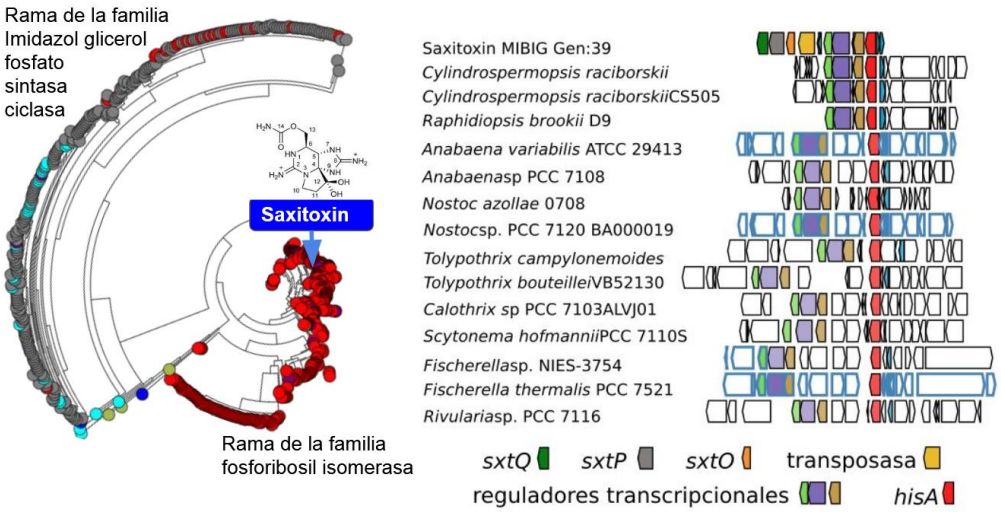

Figura 4.8 HisA en saxitoxin, un clúster de Cianobacteria.

4.3. Evolución molecular y estructural de PriA

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

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

En esta sección analizamos cómo cambia la capacidad catalítica de PriA sobre un sustrato mientras se varía la del otro. Para ello se utilizaron datos de mutantes de subHisA de *Corynebacterium diphteriae*. Estas mediciones de cinéticas enzimáticas fueron obtenidas del trabajo de tesis de Lianet Noda [Noda, 2012]. A partir de la secuencia original que se mostró es una subHisA, se realizaron mutantes con el objetivo de alcanzar la promiscuidad, es decir de convertir la enzima subHisA en una PriA. Se comenzó con diferentes mutantes puntuales adicionando una mutación cada vez, hasta llegar a una con 11 mutaciones. En esta colección de mutantes varias ganaron la función de PRA isomerasa, a distintos niveles. La que alcanzó mayor actividad PRA isomerasa fue la 9.3, una variante con nueve mutaciones. En estos datos quedaba pendiente la exploración de los caminos mutacionales, es decir cómo es el camino desde una mutante sencilla hasta una múltiple ¿cuántas rutas son posibles? ¿Existe alguna tendencia en ciertos momentos de la ruta sobre el incremento/decremento de alguna de las dos funciones?

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

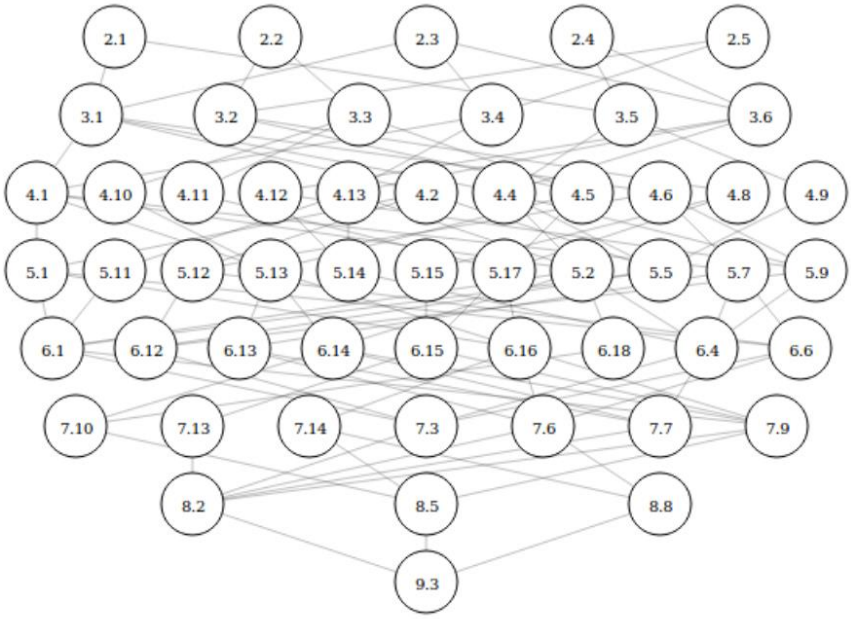

Figura 4.9 Rutas desde una subHisA hasta una variante con 9 mutaciones. En cada círculo el primer dígito indica el número de mutaciones.

Al analizar todas las rutas que llevan a 9.3 se descubrió que no existe en ellas una trayectoria en la todos los pasos incrementen la actividad de TrpF. Por otra parte, se observa que a pesar de que la primera mutación siempre decrece la actividad de HisA, en pasos subsecuentes se observa que esta actividad se incrementa, lo que muestra que hay epistasis positiva o amortiguadora. Esto es que los efectos negativos que de las mutaciones sobre SubHisA pueden ser menos negativos o hasta positivos cuando las mismas mutaciones ocurren en una proteína que ya tiene otras mutaciones. Como ejemplo, en la Figura 4.10 se muestran las rutas donde cada mutante mantiene un nivel mínimo de actividad de ProFAR isomerasa ($\frac{K_{cat}}{K_m} PriA_{ProFAR} \geq .004$). En azul sólido se ven los incrementos en PRA y en rojo sólido los incrementos en ProFAR. Las líneas punteadas indican que la actividad decreció en ese paso de la ruta. Entre una y cuatro mutaciones el azul sólido es predominante, es decir se incrementa la actividad para PRA, pero entre 4 y 5 mutaciones ningún paso incrementa la actividad de PRA y en cambio sí se incrementa la actividad para ProFAR, esta

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

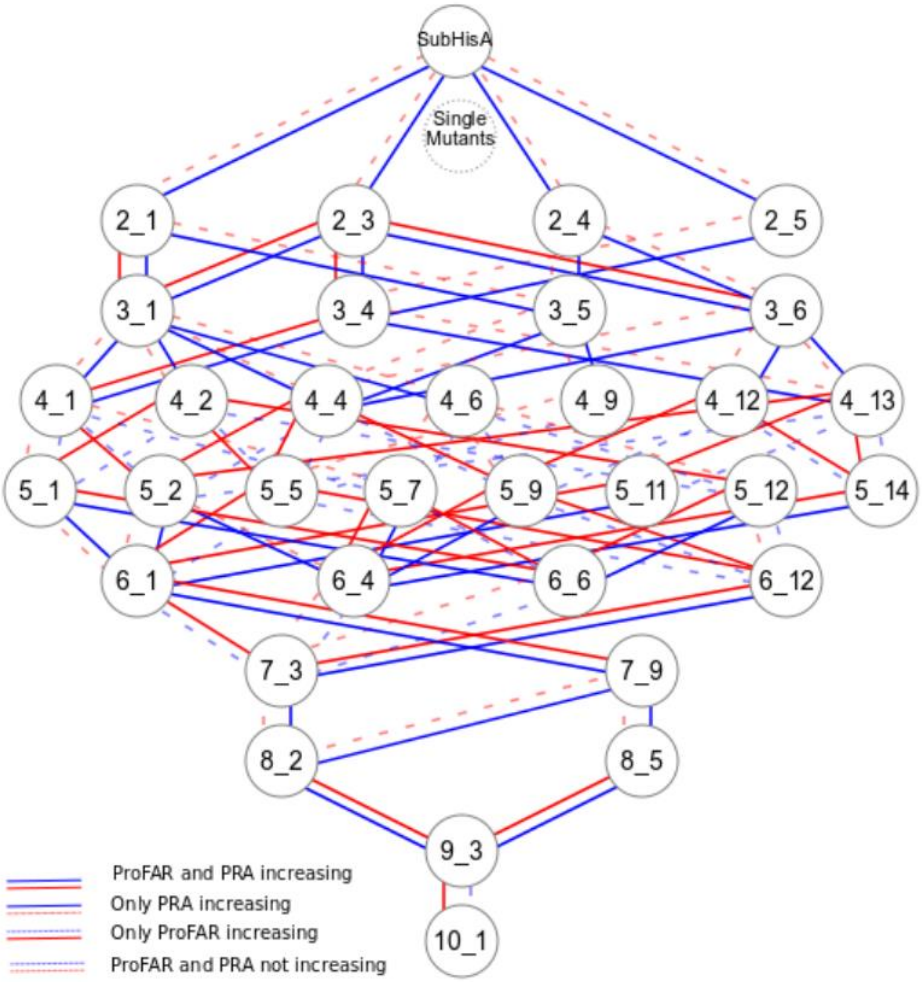

Figura 4.10 Rutas de evolución dirigida para ganar la función PRA. En la figura se muestra como la mayoría de los pasos que incrementan una función hacer decrecer la capacidad catalítica de la otra. El primer número de cada nodo indica cuántas mutaciones tiene esa variante y el segundo número es el identificador de esa combinación específica.

4.3.2. Los residuos con covariación en el registro evolutivo de PriA permiten una reconstrucción aproximada de su estructura tridimensional

El estudio de la evolución de PriA en la sección anterior nos proporcionó el aprendizaje de que para adquirir una actividad el camino no es estrictamente creciente. En cambio, suele haber pasos donde alguna de las dos actividades baja. En esta sección utilizaremos el registro evolutivo resultado de millones de años, tomaremos miles de homólogos de PriA para inferir la estructura tridimensional de una secuencia. EVcouplings es un método que considera las secuencias génicas existentes como experimentos exitosos de la naturaleza. Con esta información obtiene la covariación entre pares de aminoácidos de las secuencias existentes en el registro evolutivo. Los pares fuertemente relacionados se denominan acoplamientos, estos acoplamientos a menudo están cerca físicamente en la estructura terciaria de la proteína. Se ha demostrado que muchas proteínas contienen suficientes acoplamientos distribuidos ampliamente en toda la secuencia, de forma que con ellos es posible la reconstrucción de su estructura tridimensional [Marks, 2011]. En esta sección se aplicará EVcouplings para reconstruir la estructura tridimensional de PriA.

Las diferencias a nivel estructural pueden amplificar la información proporcionada por variaciones a nivel de secuencia. Una métrica común de distancia entre dos estructuras de proteínas es el *root mean squared deviation* (RMSD). Esta distancia se obtiene después de alinear las estructuras calculando el promedio de las distancias al cuadrado entre los átomos de carbono primarios [Kufareva, 2012]. La información de la covariación entre residuos es suficiente para que el método EVcouplings genere estructuras con un error de 2.7–4.8 Å respecto a la estructura cristalográfica conocida [Marks, 2011]. Por este motivo se decidió implementar EVcouplings y aplicarlo a la familia PriA. Este método es de difícil instalación ya que requiere varias dependencias. Por ello, desarrollé un contenedor Docker donde dependencias, software y base de datos quedan instalados. Este desarrollo fue incluido por los autores de EVcouplings como sugerencia de instalación. El contenedor Docker implementa el *EVcouplings python framework* [Hopf, 2019] que comprende cinco etapas para estudiar el análisis de coevolución de residuos de una familia de proteínas. Dichas

etapas son i) Alineado, ii) análisis de acoplamiento, iii) plegamiento basado en acoplamientos iv) análisis de mutación y v) comparación con estructuras conocidas.

EVcouplings fue aplicado a la secuencia de PriA de *Streptomyces coelicolor* obtenida de la base de datos Uniprot con identificador HIS4_STRCO. Los parámetros de EVcouplings se dejaron con su configuración inicial incluyendo el umbral de recuperación para las secuencias utilizadas en el alineamiento. Con este alineamiento, utilizando la información de los acoplamientos de sus aminoácidos, se obtuvo el modelo *PriAEV* de la estructura tridimensional de PriA de *S. coelicolor*.

Ahora bien, 1VZW es una estructura de PriA de *S. coelicolor* obtenida mediante cristalografía [Kuper, 2005]. En esta tesis calculé que el RMSD entre PriA (1VZW) y su homólogo lejano HisF (2A0N proveniente de *Thermotoga maritima* es de 7.34 Å. Así mismo la distancia obtenida entre PriA (1VZW) y el modelo de TrpF 5LHE de *Thermococcus kodakaraensis* es de 12.088 Å. Por tanto, cualquier modelo exitoso de PriA debe tener un RMSD menor que estas distancias con respecto al modelo 1VZW. La distancia entre la estructura PriAEV y HisF es de 5.603 Å y el RMSD entre PriAEV y TrpF es de 12.554 Å. Además, la distancia entre estas 1VZW y PriAEV medida utilizando el RSMD (pymol 2.2.3) entre ellas es de 3.362 Å. Así pues, PriAEV consigue diferenciarse de HisF y TrpF y estar a una distancia de 3.362 Å de 1VZW. Esta distancia es menor que el 3.730 Å el RMSD entre la estructura de PriA 1VZW de *S. coelicolor* y la estructura de PriA *Streptomyces sviceus ATCC 29083* con identificador de PDB 4TX9. Además, las estructuras cristalográficas tienen una resolución de 1-3 Å como puede verse en la Tabla 4.3, por tanto, un RMSD de 3 Å es aceptable para decir que las estructuras son similares. Sin embargo, los RMSD entre otras estructuras cristalográficas de *Streptomyces coelicolor* son menores a 2Å, por ejemplo, el (RMSD (2VEP, 1VZW) = 0.449Å, RMSD (1VZW, 2X30) = 1.929 Å, RMSD (1VZW,5DN1) =2.802 Å). Así pues, EVcouplings obtuvo una estructura de *PriA* de *S. coelicolor* similar a la cristalográfica 1VZW, pero aparentemente no mejor que otras estructuras cristalográficas de *S. coelicolor*. La comparación de la estructura PriAEV lograda por EVcouplings con respecto a 1VZW es mostrada en la Figura 4.11.

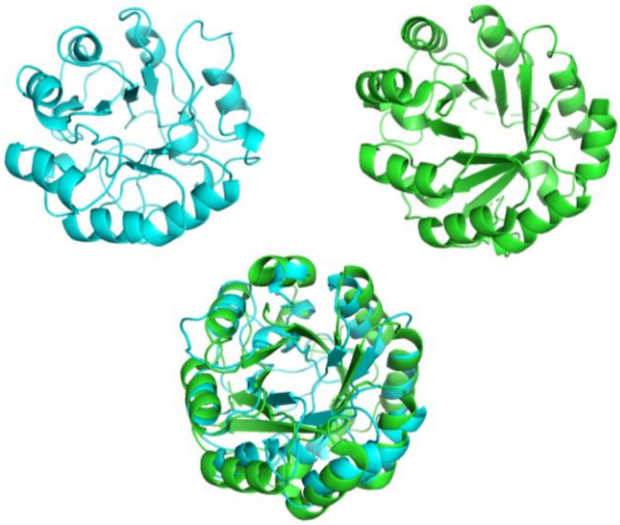

Figura 4.11 Comparación de estructura tridimensional de PriA generada porEVcouplings con una estructura cristalográfica de Streptomyces coelicolor. En azulse muestra la estructura generada considerando acoplamientos de los aminoácidosde PriA según el registro evolutivo. En verde está la estructura 1vzw obtenidaexperimentalmente. En la parte inferior de la figura se muestran las dos estructurasalineadas. Su RMSD es de 3.362 Å

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

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

Tabla 4.2 Cambios covariados de aminoácidos relevantes en el estudio de mutaciones dirigidas

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

Una forma gráfica de ver esta información se muestra en la Figura 4.12. Esta figura es simétrica porque los acoplamientos son simétricos. Los puntos negros indican acoplamientos entre pares de aminoácidos. Entre más cerca de la diagonal hay más acoplamientos porque son residuos cercanos en la secuencia lineal de la proteína. En las líneas naranjas muestro cómo el aminoácido 21D tiene un acoplamiento con el 173T, que está muy cerca del 175D que ha sido asociado con la actividad de isomerización de PRA, pero no de ProFAR. A futuro, para obtener resultados más precisos sobre la covariación de residuos en Actinobacteria, se debe proveer un alineamiento exclusivo de Actinobacteria.

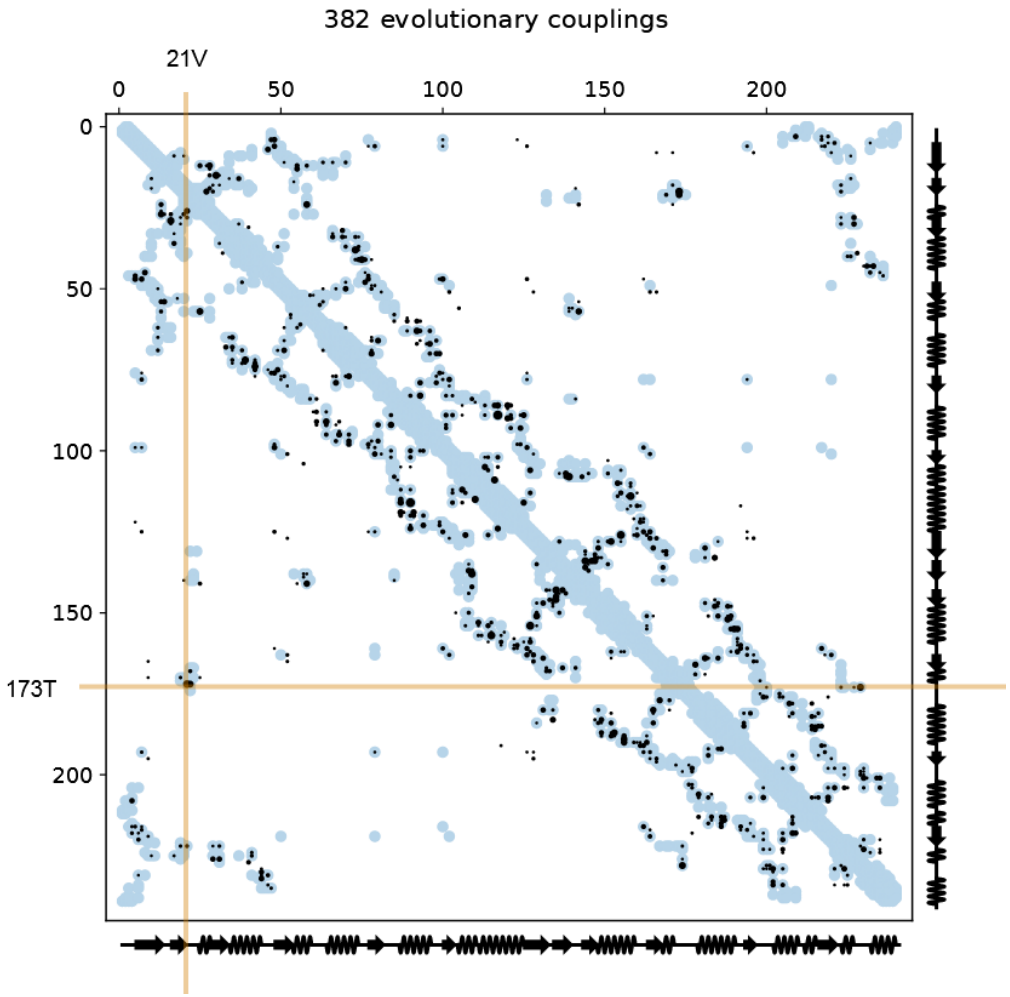

Figura 4.12 Visualización de los acoplamientos en la familia PriA

Finalmente se investigó también la variación individual de los aminoácidos catalíticos D11, D130 [Due, 2011] y D175 reportados en *Mycobacterium tuberculosis* [Verduzco-Castro, 2016]. Estas posiciones corresponden en *S coelicolor* a D11, D131 y D171. En este trabajo se encontró que existe un conjunto de secuencias donde estos residuos presentan variantes, Figura 4.13. Los homólogos más relevantes de PriA/HisA abordados en este capítulo junto con las secuencias seleccionadas divergentes en los residuos mencionados fueron alineadas para mostrar la diversidad de las familias PriA/HisA a nivel de aminoácidos. Aunque existen variantes de 131D fuera del phylum Actinobacteria, como en *Pseudomonas*, Cianobacteria, Enterobacteria, Proteobacteria y Chloroflexi también en Actinobacteria se encontraron 159 *Corynebacterium*, 49 *Streptomyces* y 2 *Actinokineospora* con un residuo diferente al aspártico en la posición 131. Están depositados en la base pública NCBI hasta

enero de 2019, 349 genomas públicos de *Streptomyces* 2019, así pues, aproximadamente un 14% tiene una variante en la posición 131. Entre estos *Streptomyces* están el ya mencionado *Streptomyces* CT 34 que posee dos copias de PriA una de ellas muy divergente y el *Streptomyces rimosus* del que se habló en el capítulo anterior porque sintetiza rimosamidas. Es llamativo que las variantes de 131D se encuentran principalmente en *Corynebacterium* y en *Streptomyces* que corresponden a los géneros donde se han ubicado a la familia subHisA y a la familia PriB respectivamente. La variabilidad mostrada en estos géneros podría estar relacionada con la existencia de estas familias. Otros los *Streptomyces* con una variante en 131D son *S. oceani*, *S. scabiei*, *S. fradiae*, *S. rimosus*, *S. sp.* CT34 y *S.* MUSC 14 y *S. D11H*. De ellos los tres últimos son además los únicos *Streptomyces* con la variante D11H que comparten con *Actinokineospora auranticolor*. De hecho, las tres posiciones conservadas en la mayoría de las PriA presentan variantes en estos cuatro organismos, por lo que sería interesante caracterizarlos bioquímicamente. Entre los organismos con doble copia de PriA están *Serinicoccus marinus*, *Serinicoccus profundi* y *Ornithinimicrobium pekingense*, ambas copias fueron incluidas en el alineamiento. Estas seis secuencias de PriA tienen aspártico en las tres posiciones conservadas, las diferencias que muestran en otras posiciones indican que las copias primarias y las secundarias de cada organismo se agrupan entre sí. La secuencia de *Actinomadura sp.* ATCC 39365 involucrada en el BGC *ada* es similar a la PriA de *S. coelicolor* por lo que posiblemente PriA de *S. coelicolor* también podría presentar esta actividad. Se incluyó en el análisis una secuencia reconstruida del ancestro común de HisA en Bacteria [Plach, 2016], esta secuencia muestra el aspártico en las tres posiciones. En Cianobacteria la HisA del BGC saxitoxina es prácticamente igual a la de *Cylindrospermopsis raciborskii* CS50 y ambas conservan el aspártico en las tres posiciones. En Archaea las PriA más grandes encontradas *Halorubrum tebenquichense* y *Halonotius sp J07HN6* tienen la variante en D171E, que comparten con los *Streptomyces* y la *Actinokineospora* con variante D11H y con los *Streptomyces* con la variante D130E. Finalmente también en Archaea *Thermococcus sp JCM 11816* que comparte el contexto genómico con genes de triptófano presenta los tres residuos conservados.

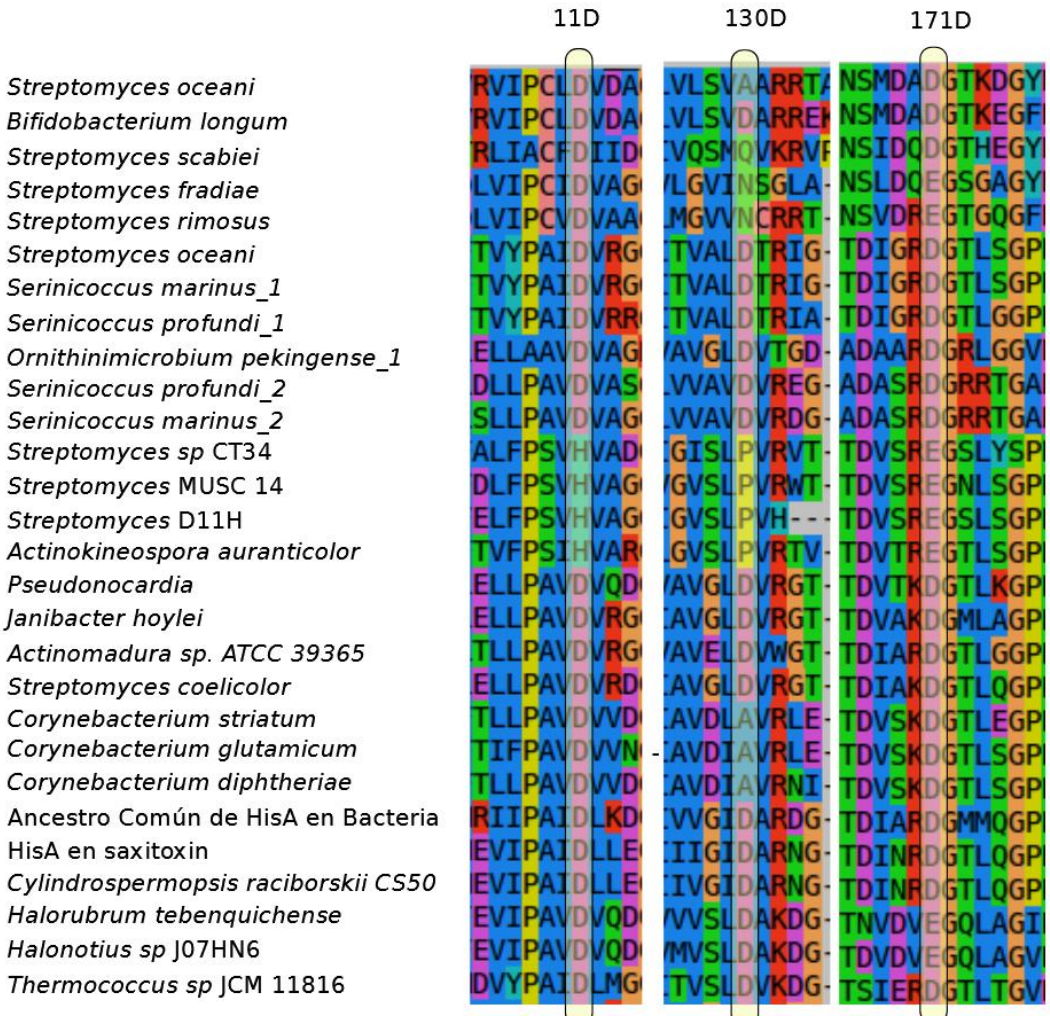

Figura 4.13 Miembros de PriA que poseen variantes en los residuos catalíticos D11, D130 y D171

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

Además de los sustratos conocidos ProFAR y PRA en los que PriA es capaz de realizar una isomerización, es posible que PriA pueda ser promiscua en otros sustratos. De hecho, como se vio en la sección de EvoMining de este capítulo PriA parece participar en la síntesis del antibiótico pentostatina (ada BGC). Tomando este ejemplo como inspiración, se buscaron sustratos químicamente parecidos a ProFAR y PRA. Esta sección buscará probar la afinidad de sustratos parecidos a los nativos de PriA para posteriormente probar alguno en copias

selectas de PriA provenientes de diversos organismos. Veinte sustratos (S1, S2, … S20) fueron recolectados tanto de la literatura [Adams, 2014, Due, 2011, Reisinger, 2014, Verduzco-Castro, 2016] como de predicciones quimio informáticas [Jffryes, 2015]. Estos sustratos son mostrados en, Figura 4.14, Los sustratos nativos son PRA (S3) y el sustrato PROFAR es S7. Por otra parte, S13-S16 son sustratos activados por luz. Los sustratos S17 (PRAP) y S18 (Compuesto V) se encontraron en la literatura, mientras que S6 (GMP), S11 (GTP) y otros fueron sugeridos por el grupo del Dr. Chris Henry debido a sus similitudes quimio informáticas. En la siguiente sección, posteriormente a la selección de sustratos seleccionaremos las secuencias de PriA con las cuales se realizará el docking enzima-sustrato.

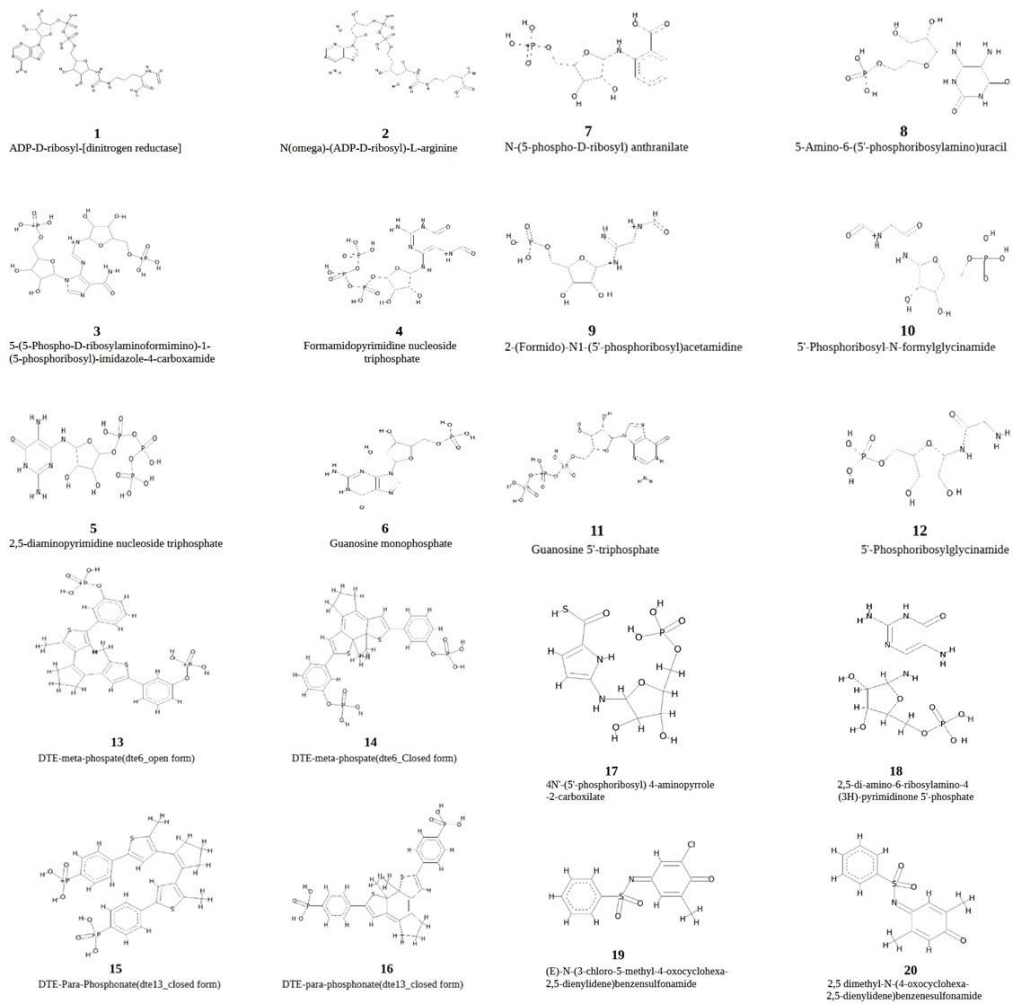

Figura 4.14 Sustratos químicamente similares a los de PriA

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

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

Para realizar el docking se requieren estructuras cristalográficas. Cuando existían estructuras cristalográficas de la secuencia específica de PriA del organismo se utilizó dicha estructura. En caso contrario, el grupo del Dr. Carrillo-Tripp generó estructuras homólogas utilizando la técnica de modelado por homología mediante el software Rosetta. Para cada secuencia decidí usar como plantilla la estructura tridimensional del homólogo de PriA más cercano que sí contara estructura con estructura cristalográfica. Los organismos con estructura cristalográfica de alguna familia relacionada a PriA están descritos en la Tabla 4.3.

Para este análisis fueron las secuencias con estructura(s) cristalográfica(s) seleccionadas fueron: para la familia HisA la Enterobacteria *Salmonella enterica* (PDB:5AHE), para la familia subHisA la Actinobacteria *Actinomyces urogenitalis* (PDB:4X2R), la estructura que representa a subTrpF es la de *Arthrobacter aurescens* (PDB:4WD0). En cuanto a la familia PriA, varias estructuras cristalográficas de *Mycobacterium tuberculosis* (Mtub

PDB:2Y88,2Y89,2Y85,3ZS4) y *Streptomyces coelicolor* (Scoe PDB:2VEP,2X30,1VZW) fueron incluidas. Las estructuras de PriB consideradas fueron las de *Streptomyces sviceus* (PDB:4U28,4TX9) y las de *Streptomyces sp* Mg1 (4W9T, 4X9S). Finalmente, la estructura cristalográfica de TrpF corresponde a *Jonesia denitrificans* (PDB:4WUI) Tabla 4.4.

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

Tabla 4.3 Estructuras cristalográficas disponibles de PriA y familias relacionadas

Organismo	**PDB**	**Familia más relevante**	**Resolución Å**	**Fecha**
Salmonella enterica	5AHE	HisA	1.7	2015
Salmonella enterica	5AB3	HisA	1.8	2016
Salmonella enterica	5ABT	HisA	1.65	2016
Salmonella enterica	5AC7	HisA	1.9	2016
Salmonella enterica	5AC8	HisA	1.7	2016
Salmonella enterica	5AC6	HisA	1.99	2016
Salmonella enterica	5A5W	HisA	1.6	2015
Salmonella enterica	5AHF	HisA	2.2	2012
Thermotoga maritima	2W79	HisA	1.85	2008
Thermotoga maritima	1QO2	HisA	1.85	2000
Streptomyces sp. Mg1	4X9S	PriB	1.6	2014
Streptomyces sviceus	4TX9	PriB	1.6	2014
Streptomyces sviceus	4U28	PriB	1.33	2014
Streptomyces sp. Mg1	4W9T	PriB	1.57	2014
Arthrobacter aurescens	4WD0	subTrpF	1.5	2014
Streptomyces coelicolor	5DN1	PriA	1.95	2015
Streptomyces coelicolor	1VZW	PriA	1.8	2004
Streptomyces coelicolor	2VEP	PriA	1.8	2007
Streptomyces coelicolor	2X30	PriA	1.95	2010
Mycobacterium tuberculosis	2Y85	PriA	2.4	2011
Mycobacterium tuberculosis	2Y88	PriA	1.33	2011
Mycobacterium tuberculosis	2Y89	PriA	2.5	2011
Mycobacterium tuberculosis	3ZS4	PriA	1.9	2012

Actinomyces urogenitalis	4X2R	SubHisA	1.05	2014
Corynebacterium efficiens	4AXK	SubHisA	2.25	2013
Thermococcus kodakaraensis	5LHE	TrpF	1.85	2016
Thermococcus kodakaraensis	5LHF	TrpF	1.75	2016
Thermus thermophilus	1V5X	TrpF	2	2003
Thermotoga maritima	1DL3	TrpF	2.7	1999
Thermotoga maritima	1LBM	TrpF	2.8	2002
Thermotoga maritima	1NSJ	TrpF	2	1996
Jonesia denitrificans	4WUI	TrpF	1.09	2014
Pyrococcus furiosus	4AAJ	TrpF	1.75	2012

Tabla 4.4 Secuencias selectas de PriA y familias relacionadads para el análisis de docking

Abreviatura	Organismo
Save	*Streptomyces avellaneus*
Scar	*Streptomyces carneus*
Spur	*Streptomyces purpeofuscus*
Smeg	*Streptomyces megasporus*
Sfrad	*Streptomyces fradiae*
Svar	*Streptomyces varsoviensis*
Satra	*Streptomyces atratus*
Srim	*Streptomyces rimosus* R6-500
SCT34	*Streptomyces sp.* CT34
S1813	*Streptomyces sp.* NRRL S-1813
Ssul	*Streptomyces sulphureus* DSM 40104
Scla	*Streptomyces clavuligerus* ATCC 27064
Stsu	*Streptomyces tsukubaensis* NRRL18488
Sbik	*Streptomyces bikiniensis*
Sven	*Streptomyces venezuelae* ATCC 10712
Sful	*Streptomyces fulvoviridis*
Scal	*Streptomyces californicus*
Sbaa	*Streptomyces baarnensis*
Salb	*Streptomyces albus* J1074
Siak	*Streptomyces iakyrus*
Sgha	*Streptomyces ghanaensis* ATCC 14672
Sbic	*Streptomyces bicolor*
Sipo	*Streptomyces ipomoeae* 91-03
Sbot	*Streptomyces bottropensis* ATCC 25435
SspC	*Streptomyces sp.* C
SMg1	*Streptomyces sp.* Mg1

Sxan	*Streptomyces xanthophaeus*
Skat	*Streptomyces katrae*
Slav	*Streptomyces lavendulae subsp. lavendulae*
Sery	*Streptomyces erythrochromogenes*
Sniv	*Streptomyces niveus* NCIMB 11891
Ssvi	*Streptomyces sviceus* ATCC 29083
Sfla	*Streptomyces flaveus*
Save	*Streptomyces avermitilis* MA-4680 = NBRC 14893
Scoe	*Streptomyces coelicolor* A3(2)
ArTC1	*Arthrobacter aurescens* TC1
Sglob	*Streptomyces globisporus* C-1027
Sgri	*Streptomyces griseolus*
S34	*Streptomyces sp.* CT34 (paralogo)

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

Con las estructuras de las enzimas seleccionadas de PriA se realizaron simulaciones de docking respecto a los veinte sustratos descritos. El procedimiento puede ser consultado en Docking Protocols. En la Figura 4.15 presento una visualización de los resultados obtenidos. En esta figura se marcan en color los organismos de los que se tiene una caracterización bioquímica, en azul están las secuencias correspondientes a TrpF, en rojo las de HisA, en rosa las secuencias de PriA y subHisA y en verde las de PriB y subTrpF. Con un asterisco fueron marcadas las secuencias de las que se cuenta con estructura cristalográfica. En el caso de las secuencias que no contaban con estructura cristalográfica, la estructura que sirvió como base está anotada después de la abreviatura del organismo. Las abreviaturas corresponden a los organismos de la Tabla 4.4. Los cuadrados grises son datos que faltaron de simular. Entre más azul es mayor la afinidad del sustrato por la secuencia, entre más rojo menor la afinidad de la enzima por este sustrato. Se requieren más controles para este experimento, pero una tendencia general es que la columna correspondiente a S11 (GTP) es el sustrato que tiene mayor afinidad. Por ello, aunque la afinidad no tiene por qué corresponder con capacidad catalítica, se decidió investigar la capacidad de secuencias de

PriA para catalizar alguna reacción sobre el GTP, esa investigación es el objeto de la última sección de este capítulo.

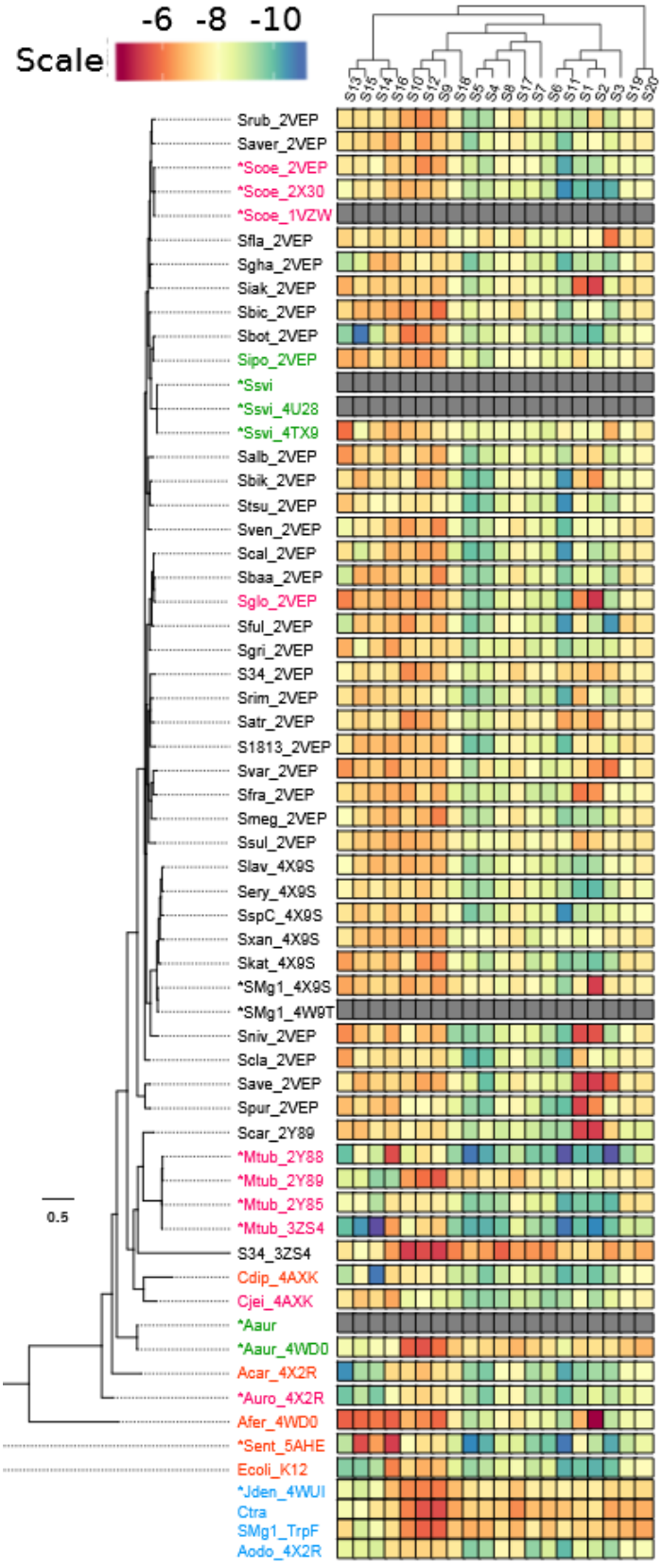

Figura 4.15 *Heatplot* del *docking* de enzimas relacionadas a PriA y sus posibles sustratos

4.5 PriA en cinéticas enzimáticas no tradicionales.

Después de la exploración genómica de PriA/HisA se trabajó en caracterizaciones experimentales no tradicionales de PriA. No se tuvo el tiempo ni la experiencia para tener réplicas de los resultados descritos a continuación, sin embargo, se incluyen en esta tesis porque pueden ser un buen comienzo para luego retomar este trabajo. Los protocolos fueron cuidadosamente descritos y se encuentran incluidos en los anexos. Este capítulo se enfoca en el montaje de dos experimentos, el primero sugerido por la sección anterior: *i)* Cinéticas de PriA en GTP. Mientras que el segundo está motivado por la curiosidad sobre el cambio en la capacidad catalítica según se provea primero un sustrato u otro, este experimento consiste en la *ii)* Medición simultánea de la actividad de PriA sobre ProFAR y PRA.

Lo primero que se realizó fue la sobreexpresión heteróloga de PriA. La secuencia de PriA fue clonada en cepas de *E. coli* V68 donde se indujo la sobre expresión. A partir de este cultivo productor de PriA se purificó la proteína. Se corroboró el éxito de esta actividad mediante un gel de proteína en la que puede verse expresión en la barra correspondiente al tamaño de PriA Figura 4.16. La proteína purificada fue almacenada a -80° en esferas de proteína selladas con nitrógeno líquido para su mejor preservación, tal y como se describe en el anexo de los procedimientos. Con este *stock* de proteína se realizaron cinéticas enzimáticas *in vitro*.

Figura 4.16 Gel de proteína donde se muestra la banda correspondiente a PriAdespués de ser sobre-expresada y purificada

4.5.1 PriA puede metabolizar GTP

Como se sugirió en la sección anterior PriA puede tener actividad sobre GTP. Ahora bien, la actividad de PriA en ProFAR es medida en un ensayo estandarizado mediante la detección de cambio de fluorescencia. Para esta medición por fluorimetría se coloca la enzima en placas de 96 pozos (Nuc 96-Well Optical Bottom Plates), se agrega el sustrato y utilizando un lector de fluorescencia de placas (TECAN infinite M1000) se mide el cambio de fluorescencia en el tiempo. Una variación de este ensayo permitió obtener información sobre la actividad PriA sobre dGTP. Los parámetros utilizados para la medición fueron tomados del trabajo no publicado de Verduzco-Castro en este laboratorio (excitación a 255 nm y emisión a 334 nm). Para comenzar esta investigación se utilizó primero dGTP en lugar de GTP esto debido a que el dGTP es un sustrato siempre disponible en cualquier laboratorio ya que es necesario para la reacción de PCR. Así pues, ensayos preliminares de actividad de PriA sobre este dGTP se realizaron en la proteína purificada de PriA de *Streptomyces coelicolor*, activa en ambos sustratos nativos utilizando como control su mutante inactiva que sustituye en la posición 11 el ácido aspártico por una alanina (D11A). La Figura 4.17 muestra el resultado de este experimento donde se aprecia un decrecimiento mayor en la fluorescencia del pozo que contiene la proteína activa respecto del pozo que contiene a la mutante.

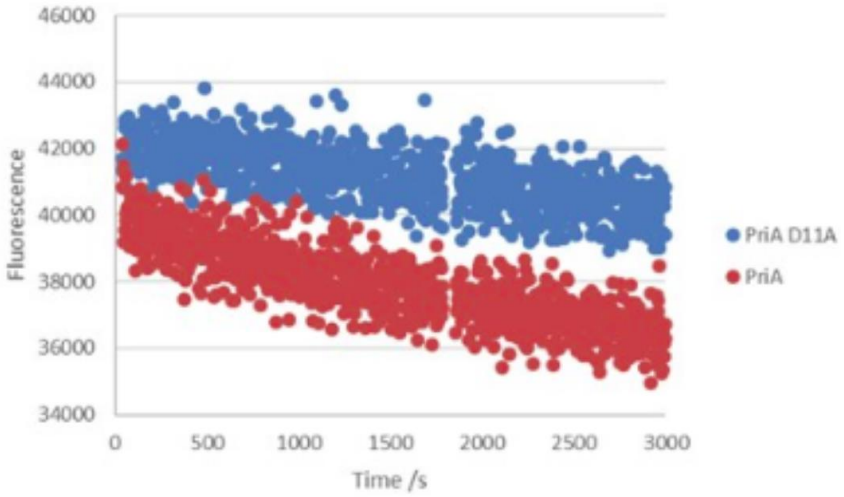

Figura 4.17 PriA de *S. coelicolor* y su mutante no funcional sobre dGTP

Este experimento de detección de actividad de dGTP fue ampliado incluyendo ahora proteínas provenientes de otros organismos. Para ello se utilizaron los stocks disponibles de

PriA provenientes de *Streptomyces roseus*, *Thermomonospora curvata* y *Mycobacterium smegmatis*; los stocks de PriB provenientes de *Streptomyces sp C*, *Streptomyces sviceus* y de *Streptomyces sp* Mg1 y finalmente el stock de TrpF de *Jonesia denitrificans*. El ensayo descrito arriba fue realizado en estas enzimas y el resultado es mostrado en Figura 4.18. En la parte inferior derecha vemos que como debe ser el control negativo, el pozo con Buffer y sustrato no muestra ningún cambio en la fluorescencia. Sin embargo, no es claro que exista una tendencia claramente decreciente en los pozos que contienen enzima. Posiblemente la pendiente que se observa en los datos de *Thermomonospora curvata*, indique cierta actividad sobre dGTP. Este comportamiento también puede deberse a que las proteínas llevaban tiempo almacenadas. Los experimentos deben ser repetidos con proteína recién purificada.

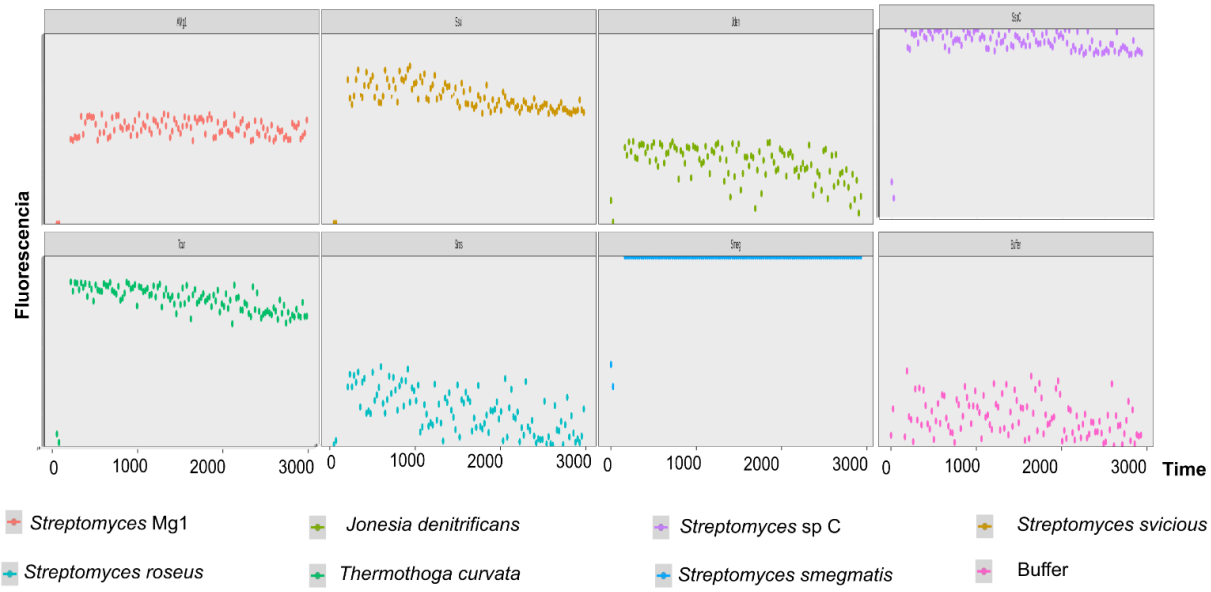

Figura 4.18 Actividad de PriA y familias relacionadas de distintos organismossobre dGTP. Este experimento no muestra disminución de fluorescencia en lasenzimas seleccionadas lo que implicaría que no existe actividad importante sobredGTP, excepto poiblemente en el caso de Thermomonospora curvata

Finalmente se probó con los mismos parámetros la actividad de PriA de *Streptomyces coelicolor* sobre GTP. Se comprobó que la enzima estuviera activa mediante ensayos enzimáticos exitosos en sus otros dos sustratos. Sin embargo, no se pudo detectar cambio en la fluorescencia de GTP debido a la actividad de PriA. Es posible que la PriA de *S coelicolor* no pueda catalizar ninguna reacción sobre el GTP por lo que una variante que a futuro podría aportar a este experimento es utilizar análogos fluorescentes de GTP.

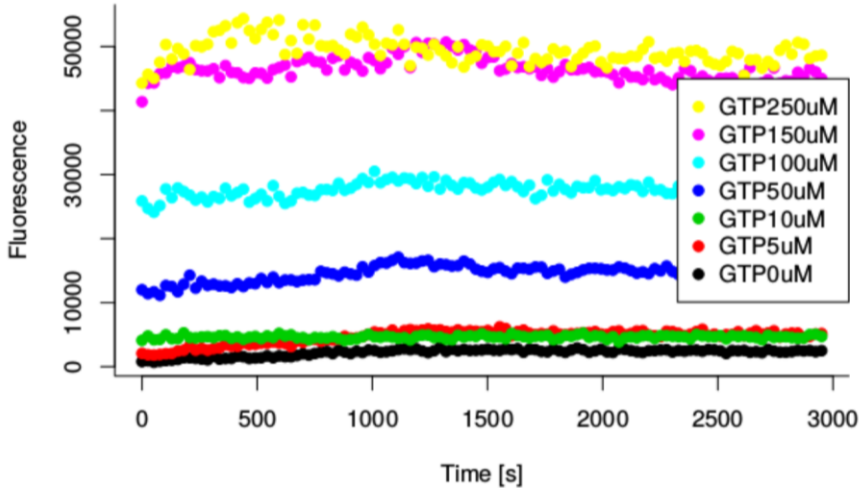

Figura 4.19 PriA de *S. coelicolor* y su mutante no funcional sobre dGTP

4.5.2 Cinéticas simultáneas para PRA y ProFAR

Otra pregunta que me surgió al investigar la promiscuidad enzimática es si esta es una propiedad de la población o de cada una de las enzimas existentes en la célula. En PriA esta pregunta puede traducirse a ¿Cómo saber si cada molécula de PriA cataliza PRA y ProFAR *in vivo*? o si en cambio, existe una subpoblación que cataliza la isomerización de PRA y otra subpoblación de moléculas que cataliza isomerización de ProFAR. Una buena aproximación *in vitro* a este problema sería poder medir el comportamiento de una sola molécula en un medio con ambos sustratos. Una serie de estos experimentos mostraría si existen subpoblaciones dedicadas a la catálisis de uno u otro substrato, o bien si todas las enzimas de una población son igualmente promiscuas. La existencia de subpoblaciones podría deberse a interacciones de la enzima con el sustrato, como por ejemplo que agregar PRA primero al medio sesgue el comportamiento de la enzima y la haga mejor para catalizar PRA que ProFAR o viceversa.

Para empezar a abordar este problema quisimos medir la actividad de una población de PriA sobre PRA y ProFAR. Por separado, ambos ensayos están montados, pero nunca se ha intentado un ensayo simultáneo. Las constantes catalíticas de PriA para la isomerización de ProFAR se mide utilizando absorbancia mientras que el ensayo enzimático de isomerización de PRA se realiza mediante la medición de fluorescencia. En esta sección se empezó a

desarrollar un método para medir simultáneamente las actividades de PRA y ProFAR. Primero intentamos medir ambas actividades secuencialmente, pero monitoreando que la señal de la otra actividad no se viera afectada. Para esto agregamos la enzima entre los segundos 100 y 400 en los pozos que ya tienen todo para que se lleve a cabo la catálisis de PRA y observamos cómo la fluorescencia baja a todas las concentraciones mientras que la absorbancia se mantiene casi constante (Figura 4.20). Luego cerca del segundo 1500 agregué el sustrato ProFAR y se observa que la absorbancia cae mientras que la fluorescencia en los mismos pozos permanece casi constante. El siguiente paso fue medir simultáneamente la actividad de una población de PriA sobre PRA y ProFAR y aunque si lo medimos no pudimos optimizarlo lo suficiente para medir las constantes catalíticas (datos no mostrados). Las pruebas piloto sugieren que es posible cuantificar la catálisis de ProFAR y PRA simultáneamente y en el caso de los homólogos que también toman a GTP se podrían medir incluso las tres actividades catalíticas para identificar si hay interacciones entre los sustratos.

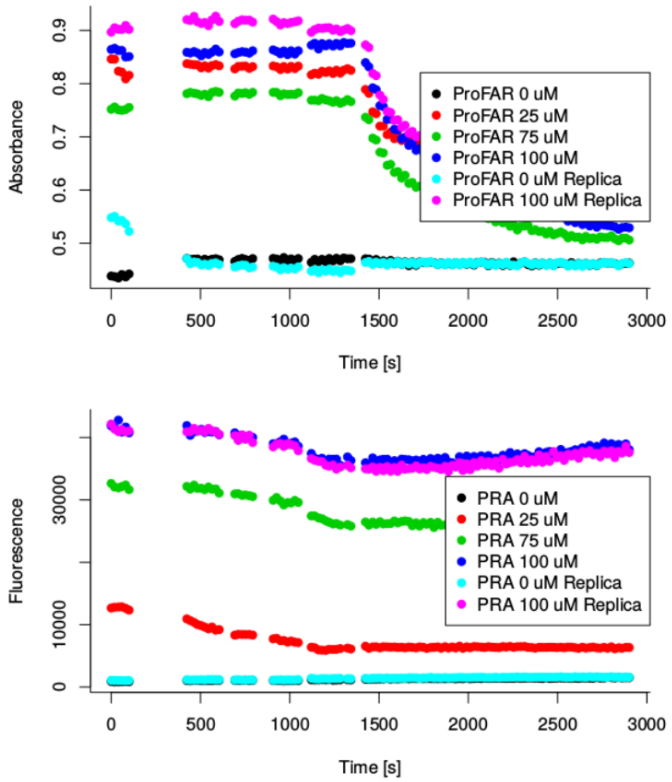

Figura 4.20 Se pueden medir las dos actividades de PriA simultáneamente. Semidió secuencialmente la actividad de ProFAR isomerasa y de PRA isomerasa enla misma población de enzimas PriA. Se cuantificó la absorbancia (arriba) y lafluorescencia (abajo) durante casi 3000 segundos. entre los segundos 100 y

400 seagregó la enzima a un buffer que ya tenía PRA y al segundo 1400 se agregó elsustrato ProFAR a la misma preparación.

En este capítulo estudiamos la familia PriA en varios niveles tanto su variación enzimática como la variación del contexto genómico. Vimos que PriA no habría sido sugerida por EvoMining como una familia promiscua ya que no tiene una marca de expansión por número de copias, y sin embargo sí es promiscua en Actinobacteria. EvoMining nos proporcionó en Cianobacteria y Archaea un ejemplo donde recupera la historia de HisF como expansión ancestral de HisA, mostrando que no todas las expansiones van al metabolismo secundario, algunas expansiones son reclutadas por el metabolismo conservado. Sin embargo, vale la pena reflexionar que en su momento HisF fue metabolismo especializado, ya que sólo los organismos con la expansión de PriA poseían esta copia que posteriormente fue fijada por la evolución. De esta forma tenemos un caso donde es el metabolismo “secundario” el que alimenta al metabolismo conservado. Otras estrategias deben abordarse para identificar familias promiscuas donde no sea claro un proceso de expansión. Finalmente encontramos las variantes más interesantes de PriA en los linajes seleccionados y analizamos residuos importantes de cada una de ellas. Finalmente se intentó caracterizar tanto nuevas actividades para PriA como la medición simultánea de las actividades nativas, aunque se avanzó en este camino, estos experimentos deben ser replicados para mayor validez.

Perspectivas

La promiscuidad ha redefinido cómo entendemos la función enzimática. Se ha transitado de considerar a todas las enzimas como altamente especializadas a entender que muchas de ellas pueden llevar a cabo más de una función metabólica. La promiscuidad parece parte del proceso evolutivo, provee material para que una posterior duplicación conlleve a la formación de nuevas familias enzimáticas. Este trabajo trató de entender la promiscuidad a diferentes escalas, incluyendo la diferenciación entre la existencia de familias promiscuas y la de ortólogos promiscuos miembros de dichas familias. Ahora bien, las enzimas suelen formar parte de clústeres biosintéticos productores de metabolitos. También se generalizó el concepto de promiscuidad enzimática al de promiscuidad de una familia de BGC, es decir consideré ahora a una familia de BGC como una unidad de síntesis y me pregunté si podía ser promiscua en el sentido de que variantes del BGC, ya sea en la secuencia de los genes del core o en el contenido de genes accesorios, produjeran variantes del metabolito. Es decir, se planteó la existencia de promiscuidad de producto en clústeres biosintéticos. Con estos tres niveles en mente: enzima, familia enzimática, y clúster biosintético se desarrollaron Orthocore, EvoMining y CORASON, tres herramientas de genómica comparativa con el objetivo de entender mejor la promiscuidad enzimática en los niveles descritos.

Con Orthocore pude resolver la filogenia del orden *Actinomycetales* utilizando los genes del *core conservado* y esto permitió entender que en dicho orden los patrones de especiación están relacionados con la pérdida y ganancia de promiscuidad de la familia PriA. Además, puede utilizar Orthocore para encontrar genes marcadores de la especie *Clavibacter michiganensis* que permiten diagnosticar la presencia de esta especie en plantas de tomate. Una lección de este capítulo es que el *core genome* no sólo contiene genes del core metabólico, también puede contener genes productores de metabolitos especializados, y este contenido incrementará entre más cercanos sean filogenéticamente los organismos del linaje seleccionado. Una vez desarrollado Orthocore quedaban por investigar las familias enzimáticas contrarias al *core conservado* es decir aquellas que sí tienen marcas de cambio en el número de copias de su gen codificante en al menos algunos organismos de un linaje genómico. El cambio ya podía ser observado porque aplicando Orthocore pueden organizarse filogenéticamente organismos de genomas cercanos, por ejemplo, los del género *Streptomyces*.

Así pues, para estudiar no sólo los patrones de expansión sino también los de neofuncionalización, en particular los de reclutamiento a metabolismo especializado desarrolle a continuación la herramienta EvoMining. Las familias señaladas por EvoMining como expandidas y reclutadas son

candidatos a ser familias promiscuas, y los miembros intermedios entre ortólogos dedicados a la función primaria y ortólogos dedicados al metabolismo especializado son candidatos, aunque no exclusivos a ser miembros promiscuos de la familia. En esta sección entendí que las familias del *shell genome* también tienen reclutamientos y expansiones, que este comportamiento no es exclusivo de ciertas familias del *core genome*. Después de esta observación queda pendiente realizar un estudio con todas las enzimas del *shell genome* de un linaje para entender cómo son las tendencias generales por ejemplo por subsistema metabólico. Con EvoMining desarrollé dos ejemplos donde en efecto enzimas de metabolismo conservado (*shell o core*) mostraban miembros reclutados en el metabolismo especializado. El primer ejemplo fueron familias del clúster de síntesis de escitonemina en Cianobacteria. A la fecha este BGC no ha sido predicho por otros métodos bioinformáticos ya que está compuesto exclusivamente por familias que provienen de expansiones del metabolismo primario y no posee ninguna enzima catalogada como exclusivamente dedicada al metabolismo especializado. El segundo ejemplo es la familia TauD común en los BGC de Rimosamida y Detoxina. Este ejemplo tiene expansiones y reclutamientos tanto en Actinobacteria como en *Pseudomonas*. El estudio de las variantes de estos BGC fue resuelto con el desarrollo de CORASON, una herramienta especializada en identificar variantes de contextos genómicos.

Así pues, desarrollé CORASON para ver el continuo de variantes de un BGC en un linaje genómico. Con esta herramienta, al observar la amplia variación en cuanto a presencia y ausencia de genes accesorios en Actinobacteria de la familia rimosamida - detoxina BGC pensé que esta familia de clústeres debía ser promiscua. Además, esta observación estaba soportada por el hecho de que rimosamida y detoxina compartían un core molecular común posiblemente relacionado con el core génico de la familia, pero tenían las moléculas se diferenciaban en ciertos ornamentos, probablemente debidos a los genes accesorios de la familia. Propuse pues que esta familia de BGC es promiscua, y que al seleccionar genomas de diversos clados del árbol del BGC se deberían encontrar nuevas variantes moleculares. Exitosamente, nuestro grupo de colaboradores pudo caracterizar dos nuevas variantes al core molecular rimosamida - detoxina comprobando así la amplia promiscuidad de esta familia.

En este punto quedó pendiente tener una medida de la diversidad de los clústeres, pienso que, así como un pangenoma de un linaje se puede clasificar como abierto o cerrado según pueda o no saturarse el número de familias génicas al agregar más genomas del linaje, este concepto podemos generalizarlo a una medida de la variabilidad genética del *pan clúster*. Entre más familias de genes distintos aparezcan en la vecindad del core del clúster mayor será la apertura de esta familia de BGC. Clústeres muy conservados con poca variación en los genes accesorios tendrán un “pancluster” cerrado, como ejemplo de este caso está el operón de histidina en ciertas clases de Archaea. Las

mismas medidas de apertura de pangenoma pueden aplicarse al *pancluster*, entre más abierto sea un clúster es posible que presente mayor promiscuidad por producto por concepto de variación en sus genes accesorios. Otra posibilidad para que un BGC sea promiscuo a pesar de tener muy poca variación en los genes accesorios es que alguna de las enzimas conservadas sea promiscua.

Finalmente, estudiamos diversos aspectos de la familia PriA, una familia promiscua en Actinobacteria que en ese phylum no pasó por un proceso de duplicación reciente y por tanto no sería sugerida por EvoMining como una familia promiscua. En este último capítulo vimos el ejemplo de HisF como resultado de una expansión de HisA y su posterior reclutamiento en el clúster de síntesis de histidina. Con ello EvoMining nos revelaba que el destino metabólico de una nueva función no tiene que permanecer en el metabolismo especializado, ese destino también puede ser fijado en metabolismo conservado. Además, también en Actinobacteria encontramos una PriA que parece catalizar una reacción de un sustrato parecido a sus sustratos nativos en un clúster de síntesis de la pentostatina. Esta secuencia es similar a la de *S. coelicolor* por lo que posiblemente hemos dado con una nueva ganancia de promiscuidad de PriA que podemos probar en los cristales de proteína ya producidos. Con CORASON vimos que, a pesar de la nueva disponibilidad de genomas, los patrones del operón his observados en los primeros estudios de Archaea se mantienen. Después estudiamos PriA a nivel de aminoácidos, montamos un método para aprovechar el registro evolutivo para generar estructuras tridimensionales de PriA. Quedó pendiente generar estructuras para todas las secuencias de la familia PriA para buscar diferencias que correlacionen con las subfamilias. Con las tres herramientas EvoMining, CORASON y EVcouplings seleccioné homólogos interesantes de PriA, ya sea porque tienen dos copias en su genoma de origen, porque poseen contextos genómicos atípicos o bien porque tienen variantes en los aminoácidos catalíticos. Queda pendiente la caracterización bioquímica de esta selección.

Además, una mezcla entre Orthocore y EvoMining me permitió descubrir que PriA no sólo no está expandida en Actinobacteria, sino que ni siquiera es parte del *core genome*. Debido a la ausencia tanto de PriA como del resto de los genes de la ruta ciertos grupos de Actinobacteria parecen ser auxótrofos para histidina. Entre ellos están *Atopobium*, *Molibuncus* y *Tropheryma*, así como ciertos *Bifidobacterium* y *Actinomyces* e incluso algunos *Corynebacterium*. Los géneros *Molibuncus* y *Tropheryma* carecen de la mayoría de los genes de síntesis tanto de histidina como de triptófano. Los *Corynebacterium* en cambio sí poseen genes de síntesis de triptófano, lo que sugiere que podría estar subfuncionalizándose reteniendo sólo la función TrpF. Esto es una novedad para *Corynebacterium* donde ya se conoce la ocurrencia de homólogos pertenecientes a subHisA, pero no se ha detectado hasta ahora la presencia de la familia subtrpF. Además, en *Saccharomonospora halophila* , *Streptomyces sp.* NRRL B-2790, *Streptomyces sp.* NRRL S-1777, *Streptomyces sp.*

NRRL WC-3549 y *Streptomyces sp.* NRRL WC-3704 tampoco fue encontrada PriA y sí genes de los operones de histidina y triptófano. En los últimos tres se localiza una doble copia de HisF. Esto puede deberse tanto a problemas técnicos de secuenciación como a una verdadera observación biológica donde se sugiere que HisF puede ser multifuncional rescatando al menos la función de HisA y posiblemente la de TrpF.

Posteriormente en PriA parece que detecté una nueva interacción con el sustrato dGTP, aunque estos experimentos no son conclusivos ya que no fueron replicados. Finalmente, una pregunta obligada es si, así como la promiscuidad varía entre ortólogos de una familia, también varía entre cada una de las moléculas provenientes del mismo gen del mismo organismo. Es decir, si la promiscuidad es una propiedad de cada molécula, o si es una propiedad de una población de moléculas, donde unas se dedican a un sustrato y otras a la catálisis de otro. Esta pregunta se debe abordar con una técnica de medición molécula por molécula. Di un primer paso intentando medir la actividad de una población de PriA simultáneamente tanto sobre ProFAR como sobre PRA.

Considero que con los desarrollos pude abordar la búsqueda de familias de enzimas candidatas a tener cambios en promiscuidad acotándome a aquellas cuyo cambio de estado era debido a un proceso de neofuncionalización que dejó como huella un cambio en el número de copias respecto a organismos cercanos. CORASON como herramienta de vecindad genómica ayudó a identificar cambios en las vecindades en el sentido de que no sólo la secuencia de la copia secundaria de la enzima era divergente de la copia central, sino también la vecindad genómica de la nueva copia era divergente de la vecindad genómica de la copia primaria. Estas herramientas, además de tratar la promiscuidad enzimática me ayudaron también a minar el pangenoma de linajes genómicos en busca de nueva química en clústeres biosintéticos de metabolitos secundarios.

Apéndices

Protocolos para usar Orthocore, myRAST, fastOrtho, Clavigenomics, y BPGA

Anotación genómica con el Docker myRAST

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

Descargar myRAST Docker distribución

Una vez con Docker instalado en la computadora, se hace pull al Docker myRAST.

docker pull nselem/myrast
Abrir myRAST en la terminal

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

Usar myRAST

-Ejemplo subir un archivo fasta

svr_submit_RAST_job -user -passwd -fasta -domain Bacteria -bioname “Organism name” -genetic_code 11 -gene_caller rast

-Para bajar un archivo de anotación genómica funcional:

svr_retrieve_RAST_job table_txt > $ID.txt

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

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

cut -f1 Rast_ID | while read line; do svr_retrieve_RAST_job $line amino_acid > $line.faa ; done
Formatos de RAST para descargar archivos

Puedes cambiar el formato table_txt por el que tú necesites.

Tabla Formatos de descarga disponibles en myRAST

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

Orthocore

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

Material suplementario de EvoMining

Tabla A1 Función y número de copias promedio en los cuatro linajes taxonómicos de las enzimas de las 42 familias de la base de datos de enzimas de EvoMining

Table S1. Function and average copy number in enzyme families in Enzyme DB

Key	Enzyme Family	*Average copy per genome in each database*				*Maximum*	Minimum
		Actino bacteria	*Cyano bacteria*	*Pseudo monas*	*Archaea*		
A1	Acetylornithine aminotransferase	8.24	2.72	12.03	3.12	Pseudomonas	Archaea
B1	Glutamine synthetase	3.21	1.17	6.46	1.29	Pseudomonas	Cyanobacteria
C1	Anthranilate synthase component 1	2.58	2.01	2.42	1.1	Pseudomonas	Archaea
D1	Fumarate hydratase	1.93	0.92	3.61	0.45	Pseudomonas	Archaea
E1	Acetolactate synthase large subunit	4.81	1.87	4.69	1.97	Actinobacteria	Cyanobacteria
F1	Aspartate transaminase	3.38	2.52	5.53	2.96	Pseudomonas	Archaea
G1	Imidazole glycerol phosphate synthase H	1.90	2.05	2.89	1.81	Pseudomonas	Archaea
A2	Dihydro picolinate synthase	2.09	0.9	3.1	1.01	Pseudomonas	Cyanobacteria
B2	Dihydroxy acid dehydratase	1.68	0.96	2.73	0.75	Pseudomonas	Archaea
C2	Diaminopimelate decarboxylase	1.68	1.02	2.43	0.54	Pseudomonas	Archaea
D2	Cysteine synthase	2.84	2.34	2.71	0.89	Pseudomonas	Archaea
E2	Acetylglutamate kinase	0.98	0.92	1.84	0.41	Pseudomonas	Archaea
F2	3-isopropylmalate dehydrogenase	1.43	1.05	2.45	1.76	Pseudomonas	Cyanobacteria
G2	Citrate synthase	2.14	0.92	1.97	0.71	Actinobacteria	Archaea
A3	Glycine hydroxymethyltransferase	1.76	0.89	2.24	0.91	Pseudomonas	Archaea
B3	Phosphoribosyl isomerase A	0.95	1.53	1.81	1.12	Pseudomonas	Archaea
C3	Fumarate reductase iron sulfur subunit	1.81	0.47	1.04	0.57	Actinobacteria	Cyanobacteria
D3	Glutamate 5 semialdehyde dehydrogenase	0.96	1.43	1.05	0.26	Cyanobacteria	Archaea
E3	Glutamine 2 oxoglutarate aminotransferase	1.18	0.23	1.46	0.5	Pseudomonas	Cyanobacteria
F3	3 dehydroquinate synthase	1.17	1.22	1.13	0.2	Pseudomonas	Archaea
G3	Pyruvate kinase	1.34	1.44	1.6	0.6	Pseudomonas	Archaea
A4	Imidazoleglycerol phosphate dehydratase	0.94	0.87	0.94	0.74	Archaea	Cyanobacteria
B4	Glutamate synthase	0.96	1.08	0.96	0.19	Cyanobacteria	Archaea
C4	Isopropylmalate isomerase large subunit	1.21	1.03	1.2	1.83	Archaea	Cyanobacteria
D4	Ornithine carbamoyltransferase	1.15	0.9	1.63	1.23	Pseudomonas	Cyanobacteria
E4	Argininosuccinate lyase	1.05	0.92	1.27	0.75	Actinobacteria	Archaea
F4	N acetylglutamate synthase	0.99	0.93	0.97	0.3	Pseudomonas	Archaea
G4	glutamate dehydrogenase	0.74	0.56	0.65	1.23	Archaea	Cyanobacteria
A5	Pyrroline 5 carboxylate reductase	0.97	0.89	0.99	0.39	Pseudomonas	Archaea
B5	Threonine synthase	1.67	1.48	1.21	1.79	Archaea	Pseudomonas
C5	Tryptophan synthase alpha	1.00	1.03	1	0.56	Cyanobacteria	Archaea
D5	Indole-3-glycerol phosphate synthase	1.10	1.06	1.02	0.67	Actinobacteria	Archaea
E5	Ribose phosphate pyrophosphokinase	1.23	0.93	1	0.84	Actinobacteria	Archaea
F5	Histidinol dehydrogenase	1.03	1.06	1	0.75	Cyanobacteria	Archaea
G5	Argininosuccinate synthase	1.06	0.92	0.99	0.77	Actinobacteria	Archaea
A6	Enolase	1.18	0.93	1.05	0.89	Actinobacteria	Archaea
B6	N acetyl gamma glutamyl phosphate reductase	1.02	0.93	0.97	0.75	Actinobacteria	Archaea
C6	Phosphoribosylanthranilate isomerase	1.12	0.87	0.98	0.48	Actinobacteria	Cyanobacteria
D6	Tryptophan synthase beta	1.22	1.02	1.08	1.22	Actinobacteria	Cyanobacteria
E6	Ketoacid reductoisomerase	1.00	0.89	0.98	0.8	Pseudomonas	Archaea
F6	Phosphoglycerate kinase	0.98	0.92	0.96	0.87	Pseudomonas	Archaea
G6	Anthranilate phosphoribosyltransferase	0.95	0.88	0.97	0.89	Pseudomonas	Cyanobacteria

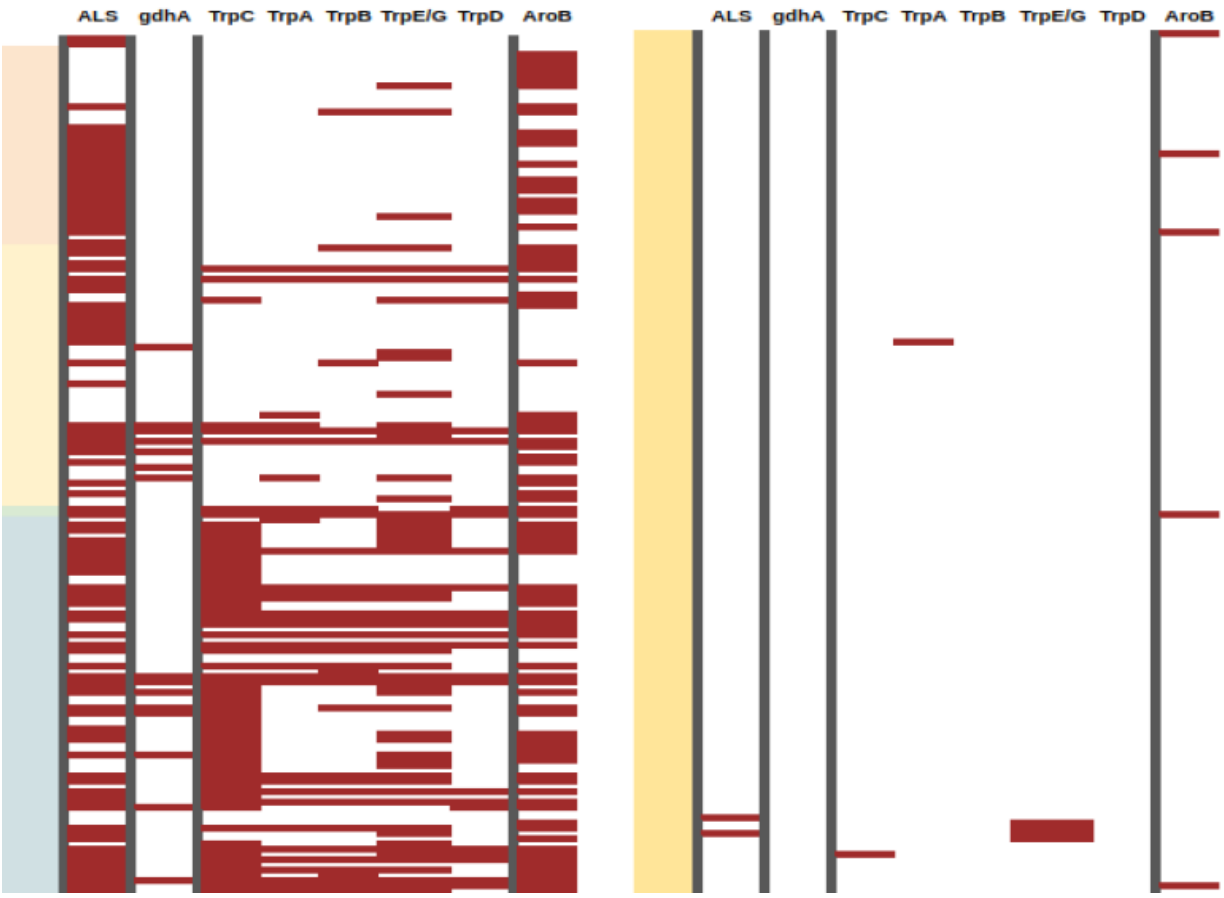

Figura A.1: HeatMap de EvoMining de enzimas relacionadas a escitonemina en el linaje Cianobacteria

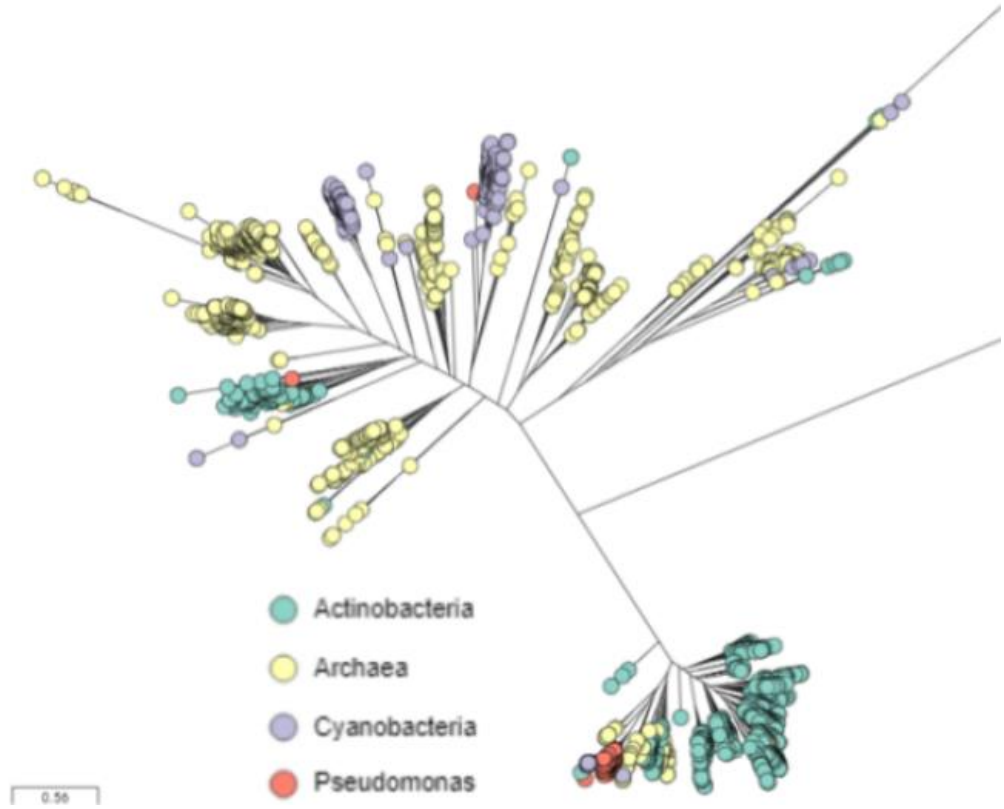

Figura A.2: Árbol de GDH en Actinobacteria, Cyanobacteria, *Pseudomonas* y Archaea donde se muestra la transferencia horizontal ocurrida en esta familia.

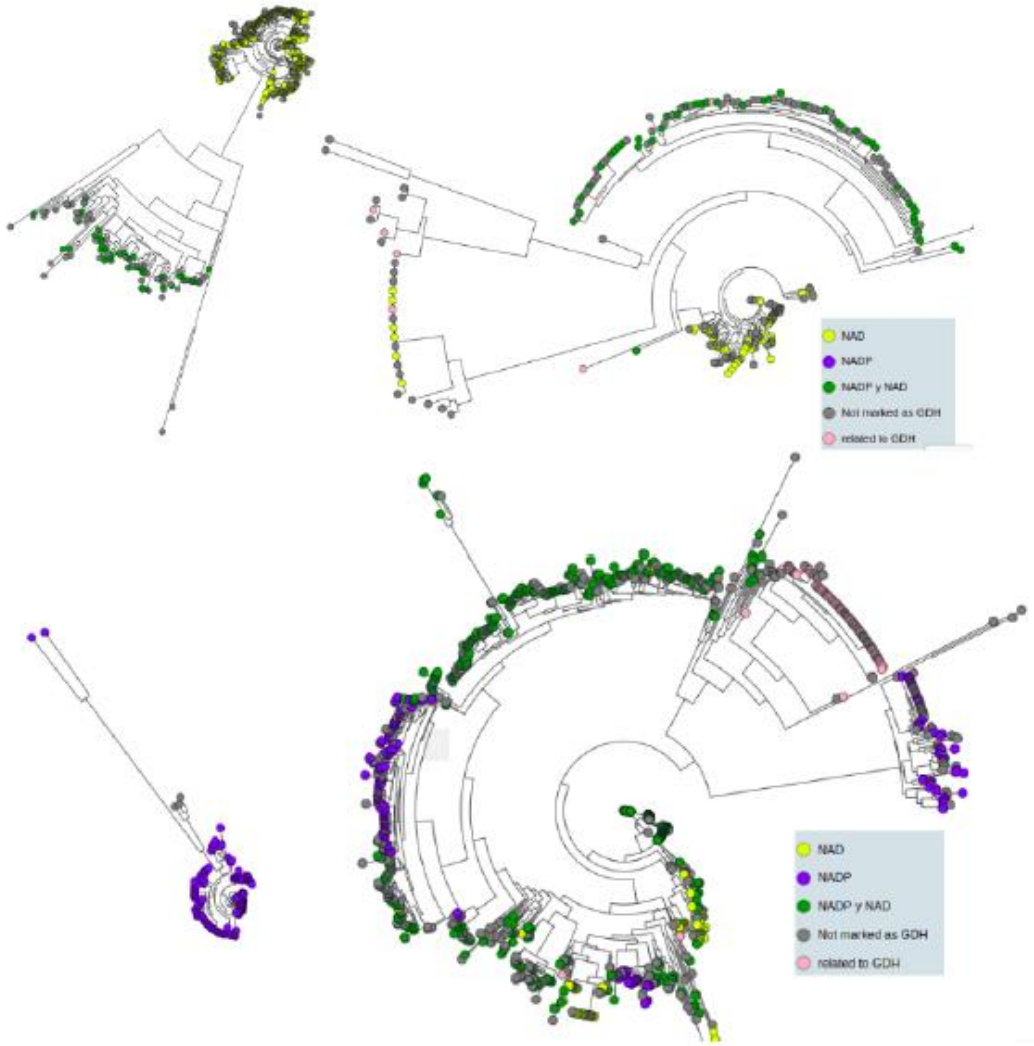

Figura A.3: Árbol de GDH en los cuatro linajes donde se muestra la alternancia en el uso de cofactores.

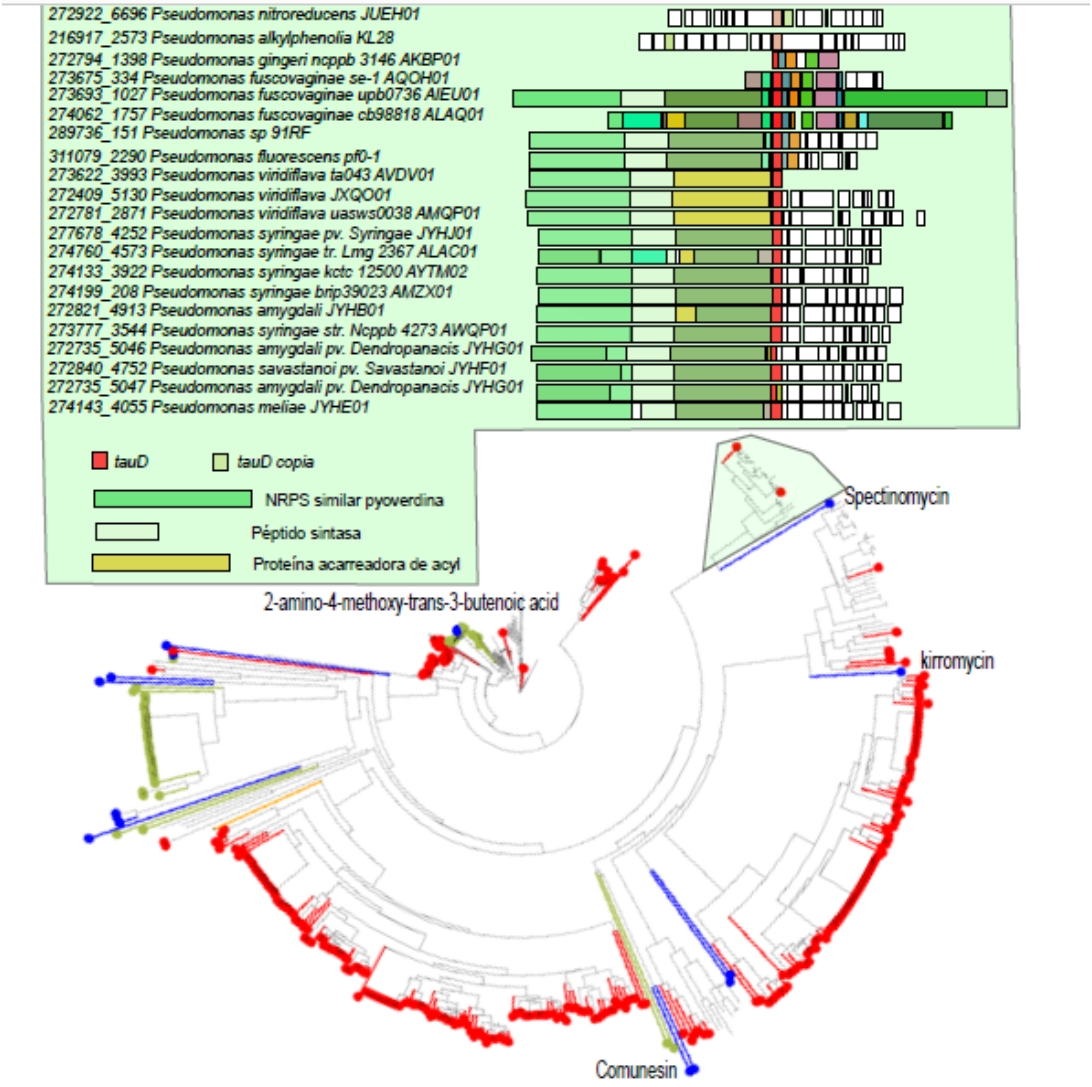

Figura A.5: En Pseudomonas el gen *tauD* tiene un contexto que incluye genes de metabolismo especializado. Variantes de este contexto están distribuidas en varios organismos.

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

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

El tercer apéndice, comandos de Docker y Git y R

Docker

-Create a new repository
docker build . -t evomining
docker push nselemevomining

Restart docker and free all ports
sudo service docker restart

list containers
docker ps -a

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

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

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

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

uninstall docker from ubuntu (Fresh start)
sudo apt-get purge docker-engine
sudo apt-get autoremove --purge docker-engine
rm -rf /var/lib/docker # This deletes all images, containers, and volumes

Run Evomining container using nselem/newevomining image
docker run -i -t -v /home/nelly/GIT/EvoMining/:/var/www/html/EvoMining/exchange -p 80:80
nselem/newevomining /bin/bash

Start evomining inside this container
perl startevomining

Visualize a tree
http://10.10.100.234/EvoMining/cgi-bin/color_tree.pl?9&&/var/www/html/EvoMining/exchange/CianosBBH_MiBIG_DB.faa_CIANOS file 9.new must be on folder volume CianosBBH_MiBIG_DB.faa_CIANOS

Find a perl module
perl -MList::Util -e'print $_ . " => " . $INC{$_} . "\n" for keys %INC' EvoMining notes
Gblocks only runs inside folder /var/www/html/EvoMining

Git

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

Connect GitHub and DockerHub

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

Apéndice 3 Dinámica molecular vs datos experimentales

Tabla B1 Actividad de PriA en S3 PRA

organism	Family	K_M	k_{cat}	$\frac{k_{cat}}{K_M}$	Pre MD	Pos MD	Reference
Afer	HisA	$1,1 \pm 0,2$	$0,05 \pm 0,001$	0.045	-10.1	-12.3	Noda-García L et al 2015
Ecoli	HisA	1,6	4,9	3.1	-9.9	-16	Henn-Sax et al. (2002)
Sent	HisA	$17,0 \pm 0,1$	$7,8 \pm 2,4$	$4,5Ö10^5$	-10.3	-20.1	Söderholm A et al (2015)
Aaur	PriB	$2,1 \pm 0,5$	$1,8 \pm 0,2$	0.9	-7.4		verduzco-castro 2016
Sipo	PriB	$3,8 \pm 0,2$	$0,82 \pm 0,02$	0.21	-8.2	-14.7	verduzco-castro 2016
SspC	PriB	$11,4 \pm 3,4$	$2,53 \pm 0,74$	0.22	-8.5	-12.7	verduzco-castro 2016
SMg1	PriB	$13,2 \pm 3,4$	$0,92 \pm 0,19$	0.069	-8	-15.2	verduzco-castro 2016
Ssvi	PriB	$3,9 \pm 0,89$	$0,69 \pm 0,04$	0.18	-8.2	-16.7	verduzco-castro 2016
Scoe	PriA	$3,6 \pm 0,7$	$1,3 \pm 0,2$	0.4	-8.4	-15	Noda-García et al (2010)
Sglob	PriA	$4,2 \pm 0,8$	$0,74 \pm 0,03$	0.18	-9.2	-16.7	verduzco-castro
Mtub 2Y85	priA	190,23	$0,012 - 9,7$				Due et al 2011
Mtub 3ZS4	priA	?	$-9,9$				Due et al 2011 (To be published)
Auro	priA	$4,0 \pm 0,9$	$0,2 \pm 0,03$	0.04	-9.2		Vazquez-Juarez (2016)
Cjei	PriA	$2,3 \pm 0,2$	$0,9 \pm 0,08$	0.39	-8.5		Noda-García et al (2013)
Cdip	subHisA	$4,4 \pm 0,5$	$2,6 \pm 0,3$	0.59	-9.2		Noda-García et al (2013)
SMg1 TrpF	TrpF3	-	-	-	-6.9	-9.6	verduzco-castro 2016
Jden	TrpF3	-	-	- -7.2	-9.4	$16,8 \pm 3,3$	Verduzco-Castro E et al 2016
Acar	SubHisA	0.02					
Aodo	SubTrpF	-	-	-			

Tabla B2 Actividad de PriA en S7 ProFAR

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

Referencias

1. Jensen. Enzyme Recruitment in Evolution of New Function. Annual Review of Microbiology [Internet]. 1976 [cited 2017 Feb 8];30(1):409–25. Available from: http: //dx.doi.org/10.1146/annurev.mi.30.100176.002205

2. Nam H, Lewis NE, Lerman JA, Lee D-H, Chang RL, Kim D, et al. Network Context and Selection in the Evolution to Enzyme Specificity. Science [Internet]. 2012 Aug [cited 2017 Feb 9];337(6098):1101–4. Available from: http://science. sciencemag.org/content/337/6098/1101

3. Patrick WM, Quandt EM, Swartzlander DB, Matsumura I. Multicopy Suppression Underpins Metabolic Evolvability. Molecular Biology and Evolution [Internet]. 2007 Dec [cited 2017 Feb 9];24(12):2716–22. Available from: https://academic.oup.com/mbe/article/24/12/2716/978039/MulticopySuppression-Underpins-Metabolic

4. Pandya C, Farelli JD, Dunaway-Mariano D, Allen KN. Enzyme Promiscuity: Engine of Evolutionary Innovation. Journal of Biological Chemistry [Internet]. 2014 Oct [cited 2017 Feb 9];289(44):30229–36. Available from: http://www.jbc.org/ content/289/44/30229

5. Khersonsky O, Tawfik DS. Enzyme promiscuity: A mechanistic and evolutionary perspective. Annual Review of Biochemistry. 2010;79:471–505.

6. Copley SD. Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. Current Opinion in Chemical Biology [Internet]. 2003 Apr [cited 2017 Feb 8];7(2):265–72. Available from: https://www.sciencedirect.com/science/ article/pii/S1367593103000322

7. Hult K, Berglund P. Enzyme promiscuity: Mechanism and applications. Trends in Biotechnology [Internet]. 2007 May [cited 2017 Feb 8];25(5):231–8. Available from: https://www.sciencedirect.com/science/article/pii/ S016777990700073X

8. O'Brien PJ, Herschlag D. Catalytic promiscuity and the evolution of new enzymatic activities. Chemistry & Biology [Internet]. 1999 Apr [cited 2017 Feb 9];6(4):R91– R105. Available from: http://www.sciencedirect.com/science/article/pii/ 158 Apéndice D. References S1074552199800337

9. Barona Gómez F, Hodgson DA. Occurrence of a putative ancient like isomerase involved in histidine and tryptophan biosynthesis. EMBO reports [Internet]. 2003 Mar [cited 2017 Feb 8];4(3):296–300. Available from: http://embor.embopress. org/content/4/3/296

10. Risso VA, Gavira JA, Gaucher EA, Sanchez Ruiz JM. Phenotypic comparisons of consensus variants versus laboratory resurrections of precambrian proteins. Proteins: Structure, Function, and Bioinformatics [Internet]. 2014 Jun [cited 2017 Feb 9];82(6):887–96. Available from: http://onlinelibrary.wiley.com/doi/10. 1002/prot.24575/abstract

11. Kumari V, Shah S, Gupta MN. Preparation of Biodiesel by Lipase-Catalyzed Transesterification of High Free Fatty Acid Containing Oil from Madhuca indica. Energy & Fuels [Internet]. 2007 Jan [cited 2017 Feb 8];21(1):368–72. Available from: http://dx.doi.org/10.1021/ef0602168

12. Li C, Henry CS, Jankowski MD, Ionita JA, Hatzimanikatis V, Broadbelt LJ. Computational discovery of biochemical routes to specialty chemicals. Chemical Engineering Science [Internet]. 2004 Nov [cited 2017 Feb 8];59(22–23):5051–60. Available from: https://www.sciencedirect.com/science/article/pii/ S0009250904006669

13. Glasner ME, Gerlt JA, Babbitt PC. Evolution of enzyme superfamilies. Current Opinion in Chemical Biology [Internet]. 2006 Oct [cited 2017 Feb 9];10(5):492–7. Available from: https://www.sciencedirect.com/science/article/pii/ S1367593106001177

14. Baier F, Copp JN, Tokuriki N. Evolution of Enzyme Superfamilies: Comprehensive Exploration of Sequence–Function Relationships. Biochemistry [Internet]. 2016 Nov [cited 2017 Feb 8];55(46):6375–88. Available from: http://dx.doi.org/10.1021/ acs.biochem.6b00723

15. Bloom JD, Romero PA, Lu Z, Arnold FH. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. Biology Direct [Internet]. 2007 [cited 2017 Feb 8];2:17. Available from: http://dx.doi.org/10. 1186/1745-6150-2-17

16. Nath A, Atkins WM. A Quantitative Index of Substrate Promiscuity. Biochemistry [Internet]. 2008 Jan [cited 2017 Feb 9];47(1):157–66. Available from: http://dx. doi.org/10.1021/bi701448p

17. Zou T, Risso VA, Gavira JA, Sanchez-Ruiz JM, Ozkan SB. Evolution of Conformational Dynamics Determines the Conversion of a Promiscuous Generalist into a Specialist Enzyme. Molecular Biology and Evolution [Internet]. 2015 Jan [cited 2017 Feb 9];32(1):132–43. Available from: https://academic.oup.com/mbe/article/32/ 159 1/132/2925568/Evolution-of-Conformational-Dynamics-Determines

18. Firn RD, Jones CG. A Darwinian view of metabolism: Molecular properties determine fitness. Journal of Experimental Botany [Internet]. 2009 Mar [cited 2017 Feb 8];60(3):719–26. Available from: https://academic.oup.com/jxb/article/ 60/3/719/452667/A-Darwinian-view-of-metabolism-molecular

19. Weng J-K, Noel JP. The Remarkable Pliability and Promiscuity of Specialized Metabolism. Cold Spring Harbor Symposia on Quantitative Biology [Internet]. 2012 Jan [cited 2019 Jan 24];77:309–20. Available from: http://symposium.cshlp.org/ content/77/309

20. Jia B, Cheong G-W, Zhang S. Multifunctional enzymes in archaea: Promiscuity and moonlight. Extremophiles [Internet]. 2013 Mar [cited 2017 Feb 8];17(2):193–203. Available from: http://link.springer.com/article/10.1007/s00792-012- 0509-1

21. Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, Tawfik DS. The 'evolvability' of promiscuous protein functions. Nature Genetics [Internet]. 2005 Jan [cited 2017 Feb 8];37(1):73–6. Available from: http://www.nature.com/ng/ journal/v37/n1/full/ng1482.html

22. Dean AM, Thornton JW. Mechanistic approaches to the study of evolution. Nature reviews Genetics [Internet]. 2007 Sep [cited 2017 Feb 8];8(9):675–88. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2488205/

23. Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nature Biotechnology [Internet]. 2009 Feb [cited 2017 Feb 9];27(2):157–67. Available from: http://www.nature.com/nbt/journal/v27/n2/ full/nbt1519.html

24. Hopkins AL. Drug discovery: Predicting promiscuity. Nature [Internet]. 2009 Nov [cited 2017 Feb 8];462(7270):167–8. Available from: http://www.nature.com/ nature/journal/v462/n7270/full/462167a.html

25. Nath A, Zientek MA, Burke BJ, Jiang Y, Atkins WM. Quantifying and Predicting the Promiscuity and Isoform Specificity of Small-Molecule Cytochrome P450 Inhibitors. Drug Metabolism and

Disposition [Internet]. 2010 Dec [cited 2017 Feb 9];38(12):2195–203. Available from: http://dmd.aspetjournals.org/content/ 38/12/2195

26. Eichborn J von, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R. PROMISCUOUS: A database for network-based drug-repositioning. Nucleic Acids Research [Internet]. 2011 Jan [cited 2017 Feb 9];39(suppl_1):D1060–6. Available from: https://academic.oup.com/nar/article/39/suppl_1/D1060/ 2506056/PROMISCUOUS-a-database-for-network-based-drug

27. Zhang W, Dourado DFAR, Fernandes PA, Ramos MJ, Mannervik B. Multidimensional epistasis and fitness landscapes in enzyme evolution. Biochemi- 160 Apéndice D. References cal Journal [Internet]. 2012 Jul [cited 2017 Feb 9];445(1):39–46. Available from: http://www.biochemj.org/content/445/1/39

28. Carbonell P, Faulon J-L. Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics [Internet]. 2010 Aug [cited 2017 Feb 8];26(16):2012–9. Available from: https://academic.oup.com/bioinformatics/article/26/16/ 2012/215921/Molecular-signatures-based-prediction-of-enzyme

29. Nagao C, Nagano N, Mizuguchi K. Prediction of Detailed Enzyme Functions and Identification of Specificity Determining Residues by Random Forests. PLOS ONE [Internet]. 2014 Jan [cited 2017 Feb 8];9(1):e84623. Available from: http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0084623

30. Cheng X-Y, Huang W-J, Hu S-C, Zhang H-L, Wang H, Zhang J-X, et al. A Global Characterization and Identification of Multifunctional Enzymes. PLoS ONE [Internet]. 2012 Jun [cited 2017 Feb 8];7(6). Available from: http://www.ncbi. nlm.nih.gov/pmc/articles/PMC3377604/

31. Noda-García L, Camacho-Zarco AR, Medina-Ruíz S, Gaytán P, CarrilloTripp M, Fülöp V, et al. Evolution of Substrate Specificity in a Recipient's Enzyme Following Horizontal Gene Transfer. Molecular Biology and Evolution [Internet]. 2013 Sep [cited 2017 Jan 31];30(9):2024–34. Available from: https://academic.oup.com/mbe/article/30/9/2024/1000280/Evolutionof-Substrate-Specificity-in-a-Recipient

32. Verdel-Aranda K, López-Cortina ST, Hodgson DA, Barona-Gómez F. Molecular annotation of ketol-acid reductoisomerases from Streptomyces reveals a novel amino acid biosynthesis interlock mediated by enzyme promiscuity. Microbial Biotechnology [Internet]. 2015 Mar [cited 2017 Feb 9];8(2):239–52. Available from: http: //onlinelibrary.wiley.com/doi/10.1111/1751-7915.12175/abstract

33. Plach MG, Reisinger B, Sterner R, Merkl R. Long-Term Persistence of Bifunctionality Contributes to the Robustness of Microbial Life through Exaptation. PLOS Genetics [Internet]. 2016 Jan [cited 2017 Feb 13];12(1):e1005836. Available from: http://journals.plos.org/plosgenetics/article?id=10.1371/ journal.pgen.1005836

34. Copley SD. An Evolutionary Biochemist's Perspective on Promiscuity. Trends in biochemical sciences [Internet]. 2015 Feb [cited 2017 Feb 8];40(2):72–8. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4836852/

35. Espinosa-Cantú A, Ascencio D, Barona-Gómez F, DeLuna A. Gene duplication and the evolution of moonlighting proteins. Frontiers in Genetics [Internet]. 2015 Jul [cited 2017 Feb 8];6. Available from: http://www.ncbi.nlm.nih.gov/pmc/ articles/PMC4493404/

36. Sanchez-Ruiz JM. On promiscuity, changing environments and the possibility of replaying the molecular tape of life. Biochemical Journal [Internet]. 2012 Jul [cited 161 2017 Feb 9];445(1):e1–3. Available from: http://www.biochemj.org/content/ 445/1/e1

37. Martínez-Núñez MA, Rodríguez-Vázquez K, Pérez-Rueda E. The lifestyle of prokaryotic organisms influences the repertoire of promiscuous enzymes. Proteins: Structure, Function, and Bioinformatics [Internet]. 2015 Sep [cited 2017 Feb 8];83(9):1625–31. Available from: http://onlinelibrary.wiley.com/doi/ 10.1002/prot.24847/abstract

38. Notebaart RA, Szappanos B, Kintses B, Pál F, Györkei Á, Bogos B, et al. Networklevel architecture and the evolutionary potential of underground metabolism. Proceedings of the National Academy of Sciences [Internet]. 2014 Aug [cited 2017 Feb 9];111(32):11762–7. Available from: http://www.pnas.org/content/111/32/ 11762

39. Linster CL, Van Schaftingen E, Hanson AD. Metabolite damage and its repair or pre-emption. Nature Chemical Biology [Internet]. 2013 Feb [cited 2017 Feb 8];9(2):72–80. Available from: http://www.nature.com/nchembio/journal/v9/ n2/full/nchembio.1141.html

40. Khanal A, Yu McLoughlin S, Kershner JP, Copley SD. Differential Effects of a Mutation on the Normal and Promiscuous Activities of Orthologs: Implications for Natural and Directed Evolution. Molecular Biology and Evolution [Internet]. 2015 Jan [cited 2017 Feb 8];32(1):100–8. Available from: https://academic.oup.com/mbe/article/32/1/100/2925554/DifferentialEffects-of-a-Mutation-on-the-Normal

41. Ma H-M, Zhou Q, Tang Y-M, Zhang Z, Chen Y-S, He H-Y, et al. Unconventional Origin and Hybrid System for Construction of Pyrrolopyrrole Moiety in Kosinostatin Biosynthesis. Chemistry & Biology [Internet]. 2013 Jun [cited 2017 Feb 8];20(6):796– 805. Available from: https://www.sciencedirect.com/science/article/pii/ S1074552113001701

42. Adams NE, Thiaville JJ, Proestos J, Juárez-Vázquez AL, McCoy AJ, BaronaGómez F, et al. Promiscuous and Adaptable Enzymes Fill “Holes” in the Tetrahydrofolate Pathway in Chlamydia Species. mBio [Internet]. 2014 Jul [cited 2017 Jan 31];5(4). Available from: http://www.ncbi.nlm.nih.gov/pmc/ articles/PMC4161248/

43. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions. Nature Reviews Genetics [Internet]. 2010 Aug [cited 2017 Feb 9];11(8):572–82. Available from: http://www.nature.com/nrg/journal/v11/n8/full/nrg2808. html

44. Fondi M, Emiliani G, Liò P, Gribaldo S, Fani R. The evolution of histidine biosynthesis in archaea: Insights into the his genes structure and organization in LUCA. Journal of Molecular Evolution. 2009 Nov;69(5):512–26.

45. Merino E, Jensen RA, Yanofsky C. Evolution of bacterial trp operons and their 162 Apéndice D. References regulation. Current opinion in microbiology [Internet]. 2008 Apr [cited 2017 Feb 10];11(2):78–86. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/ PMC2387123/

46. Verduzco-Castro EA, Michalska K, Endres M, Juárez-Vazquez AL, Noda-García L, Chang C, et al. Co-occurrence of analogous enzymes determines evolution of a novel βα8-isomerase sub-family after non-conserved mutations in flexible loop. Biochemical Journal [Internet]. 2016 May;473(9):1141–52. Available from: http://www.biochemj.org/content/473/9/1141

47. Juárez-Vázquez AL, Edirisinghe JN, Verduzco-Castro EA, Michalska K, Wu C, Noda-García L, et al. Evolution of substrate specificity in a retained enzyme driven by gene loss. eLife [Internet]. [cited 2018 Jan 16];6. Available from: https: //www.ncbi.nlm.nih.gov/pmc/articles/PMC5404923/

48. Halachev MR, Loman NJ, Pallen MJ. Calculating Orthologs in Bacteria and Archaea: A Divide and Conquer Approach. PLOS ONE [Internet]. 2011 Dec [cited 2016 Sep 16];6(12):e28388. Available from: http://journals.plos.org/plosone/ article?id=10.1371/journal.pone.0028388

49. Kislyuk AO, Haegeman B, Bergman NH, Weitz JS. Genomic fluidity: An integrative view of gene diversity within microbial populations. BMC Genomics [Internet]. 2011 [cited 2017 Jan 26];12:32. Available from: http://dx.doi.org/10.1186/1471- 2164-12-32

50. Noda-García L, Juárez-Vázquez AL, Ávila-Arcos MC, Verduzco-Castro EA, Montero-Morán G, Gaytán P, et al. Insights into the evolution of enzyme substrate promiscuity after the discovery of βα8 isomerase evolutionary intermediates from a diverse metagenome. BMC Evolutionary Biology [Internet]. 2015 Jun [cited 2017 Jan 31];15. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/ PMC4462073/

51. Garcia-Seisdedos H, Ibarra-Molero B, Sanchez-Ruiz JM. Probing the Mutational Interplay between Primary and Promiscuous Protein Functions: A ComputationalExperimental Approach. PLOS Computational Biology [Internet]. 2012 Jun [cited 2017 Feb 8];8(6):e1002558. Available from: http://journals.plos.org/ ploscompbiol/article?id=10.1371/journal.pcbi.1002558

52. Nesvizhskii AI, Vitek O, Aebersold R. Analysis and validation of proteomic data generated by tandem mass spectrometry. Nature Methods [Internet]. 2007 Oct [cited 2017 Feb 9];4(10):787–97. Available from: http://www.nature.com/nmeth/ journal/v4/n10/abs/nmeth1088.html

53. Medema MH, Fischbach MA. Computational approaches to natural product discovery. Nature Chemical Biology [Internet]. 2015 Sep [cited 2017 Jan 24];11(9):639– 48. Available from: http://www.nature.com/nchembio/journal/v11/n9/full/ nchembio.1884.html

54. Campbell I. Biophysical Techniques - Paperback - Iain D. Campbell - Oxford 163 University Press. 2012.

55. Yang JY, Sanchez LM, Rath CM, Liu X, Boudreau PD, Bruns N, et al. Molecular Networking as a Dereplication Strategy. Journal of Natural Products [Internet]. 2013 Sep [cited 2017 Feb 9];76(9):1686–99. Available from: http://dx.doi.org/ 10.1021/np400413s

56. Köcher T, Superti-Furga G. Mass spectrometry–based functional proteomics: From molecular machines to protein networks. Nature Methods [Internet]. 2007 Oct [cited 2017 Feb 10];4(10):807–15. Available from: http://www.nature.com/nmeth/ journal/v4/n10/full/nmeth1093.html

57. James LC, Tawfik DS. Conformational diversity and protein evolution – a 60-year-old hypothesis revisited. Trends in Biochemical Sciences [Internet]. 2003 Jul;28(7):361–8. Available from: https://www.sciencedirect.com/science/ article/pii/S096800040300135X

58. Parisi G, Zea DJ, Monzon AM, Marino-Buslje C. Conformational diversity and the emergence of sequence signatures during evolution. Current Opinion in Structural Biology [Internet]. 2015 Jun [cited 2017 Feb 9];32:58–65. Available from: https: //www.sciencedirect.com/science/article/pii/S0959440X15000147

59. Javier Zea D, Miguel Monzon A, Fornasari MS, Marino-Buslje C, Parisi G. Protein Conformational Diversity Correlates with Evolutionary Rate. Molecular Biology and Evolution [Internet]. 2013 Jul;30(7):1500–3. Available from: https://academic.oup.com/mbe/article/30/7/1500/972515/ProteinConformational-Diversity-Correlates-with

60. Gatti-Lafranconi P, Hollfelder F. Flexibility and Reactivity in Promiscuous Enzymes. ChemBioChem [Internet]. 2013 Feb [cited 2017 Feb 8];14(3):285–92. Available from: http://onlinelibrary.wiley.com/doi/10.1002/cbic.201200628/ abstract

61. Cruz-Morales P, Kopp JF, Martínez-Guerrero C, Yáñez-Guerra LA, Selem-Mojica N, Ramos-Aboites H, et al. Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes. Genome Biology and Evolution [Internet]. 2016 Jun [cited 2017 Jan 24];8(6):1906–16. Available from: http://gbe.oxfordjournals.org/content/8/ 6/1906

62. Li L, Stoeckert CJ, Roos DS. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research [Internet]. 2003 Sep [cited 2017 Feb 8];13(9):2178–89. Available from: http://genome.cshlp.org/content/13/9/ 2178

63. Waterhouse RM, Tegenfeldt F, Li J, Zdobnov EM, Kriventseva EV. OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. Nucleic Acids Research [Internet]. 2013 Jan [cited 2017 Feb 9];41(D1):D358–65. Available from: https://academic.oup.com/nar/article/41/D1/D358/1060216/OrthoDB-a- 164 Apéndice D. References hierarchical-catalog-of-animal-fungal

64. Gao B, Gupta RS. Phylogenetic Framework and Molecular Signatures for the Main Clades of the Phylum Actinobacteria. Microbiology and Molecular Biology Reviews : MMBR [Internet]. 2012 Mar [cited 2017 Feb 8];76(1):66–112. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3294427/

65. Sen A, Daubin V, Abrouk D, Gifford I, Berry AM, Normand P. Phylogeny of the class Actinobacteria revisited in the light of complete genomes. The orders “Frankiales” and Micrococcales should be split into coherent entities: Proposal of Frankiales ord. nov., Geodermatophilales ord. nov., Acidothermales ord. nov. and Nakamurellales ord. nov. International Journal of Systematic and Evolutionary Microbiology [Internet]. 2014 [cited 2017 Feb 9];64(11):3821–32. Available from: http://ijs.microbiologyresearch.org/content/journal/ijsem/ 10.1099/ijs.0.063966-0

66. Zhou Z, Gu J, Li Y-Q, Wang Y. Genome plasticity and systems evolution in Streptomyces. BMC Bioinformatics [Internet]. 2012 [cited 2017 Feb 9];13(10):S8. Available from: http://dx.doi.org/10.1186/1471-2105-13-S10-S8

67. Kim J-N, Kim Y, Jeong Y, Roe J-H, Kim B-G, Cho B-K. Comparative Genomics Reveals the Core and Accessory Genomes of Streptomyces Species. Journal of Microbiology and Biotechnology. 2015 Oct;25(10):1599–605.

68. Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, et al. Discovery of new enzymes and metabolic pathways by using structure and genome context. Nature [Internet]. 2013 Oct [cited 2017 Feb 9];502(7473):698–702. Available from: http:// www.nature.com/nature/journal/v502/n7473/full/nature12576.html

69. Hughes AL. The Evolution of Functionally Novel Proteins after Gene Duplication. Proceedings of the Royal Society of London B: Biological Sciences [Internet]. 1994 May;256(1346):119–24. Available from: http://rspb.royalsocietypublishing. org/content/256/1346/119

70. Divergent Evolution of Enzymatic Function: Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies. Annual Review of Biochemistry [Internet]. 2001 [cited 2017 Feb 8];70(1):209–46. Available from: http://dx.doi.org/10. 1146/annurev.biochem.70.1.209

71. Huang R, Hippauf F, Rohrbeck D, Haustein M, Wenke K, Feike J, et al. Enzyme functional evolution through improved catalysis of ancestrally nonpreferred substrates. Proceedings of the National Academy of Sciences [Internet]. 2012 Feb;109(8):2966– 71. Available from: http://www.pnas.org/content/109/8/2966

72. Schulz WL, Durant TJS, Siddon AJ, Torres R. Use of application containers and workflows for genomic data analysis. Journal of Pathology Informatics [Internet]. 2016 Dec [cited 2018 Oct 9];7. Available from: https://www.ncbi.nlm.nih.gov/ 165 pmc/articles/PMC5248400/

73. Gruening B, Sallou O, Moreno P, Veiga Leprevost F da, Ménager H, Søndergaard D, et al. Recommendations for the packaging and containerizing of bioinformatics software. F1000Research [Internet]. 2018 Jun [cited 2018 Jun 27];7:742. Available from: https://f1000research.com/articles/7-742/v1

74. Lamble HJ, Heyer NI, Bull SD, Hough DW, Danson MJ. Metabolic Pathway Promiscuity in the Archaeon Sulfolobus solfataricus Revealed by Studies on Glucose Dehydrogenase and 2-Keto-3-deoxygluconate Aldolase. Journal of Biological Chemistry [Internet]. 2003 Sep [cited 2019 Jan 25];278(36):34066–72. Available from: http://www.jbc.org/content/278/36/34066

75. Noda-Garcia L. Estudio de la evolución molecular de la función enzimática susando como modelo una enzima con características ancestrales [PhD thesis]. [Irapuato, GTO]: Langebio, CINVESTAV; 2012.

76. Zhao S, Sakai A, Zhang X, Vetting MW, Kumar R, Hillerich B, et al. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. eLife [Internet]. 2014 Jun [cited 2017 Feb 9];3:e03275. Available from: https://elifesciences.org/content/3/e03275v2

77. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: Protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research [Internet]. 2015 Jan [cited 2017 Feb 9];43(D1):D447– 52. Available from: https://academic.oup.com/nar/article/43/D1/D447/ 2435295/STRING-v10-protein-protein-interaction-networks

78. Segata N, Börnigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nature Communications [Internet]. 2013 Aug [cited 2019 Jan 24];4:2304. Available from: https://www.nature.com/articles/ncomms3304

79. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, et al. Insights from 20 years of bacterial genome sequencing. Functional & Integrative Genomics [Internet]. 2015 [cited 2017 Jan 28];15(2):141–61. Available from: http://www. ncbi.nlm.nih.gov/pmc/articles/PMC4361730/

80. Koonin EV. The Turbulent Network Dynamics of Microbial Evolution and the Statistical Tree of Life. Journal of Molecular Evolution [Internet]. 2015 [cited 2017 Jan 28];80(5-6):244–50. Available from: http://www.ncbi.nlm.nih.gov/pmc/ articles/PMC4472940/

81. Veen BE van der, Harris HM, O´Toole PW, Claesson MJ. Metaphor: Finding Bidirectional Best Hit homology relationships in (meta)genomic datasets. Genomics [Internet]. 2014 Dec [cited 2018 Jul 3];104(6, Part B):459–63. Available from: http: //www.sciencedirect.com/science/article/pii/S0888754314002092

82. Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mit- 166 Apéndice D. References tenthal JE. The origin and evolution of modern metabolism. The International Journal of Biochemistry & Cell Biology [Internet]. 2009 Feb [cited 2018 Jul 3];41(2):285–97. Available from: http://www.sciencedirect.com/science/ article/pii/S1357272508003373

83. Schniete JK, Cruz-Morales P, Selem-Mojica N, Fernández-Martínez LT, Hunter IS, Barona-Gómez F, et al. Expanding Primary Metabolism Helps Generate the Metabolic Robustness To Facilitate Antibiotic Biosynthesis in Streptomyces. mBio [Internet]. 2018 Mar [cited 2018 Aug 10];9(1):e02283–17. Available from: http: //mbio.asm.org/content/9/1/e02283-17

84. Alanjary M, Kronmiller B, Adamek M, Blin K, Weber T, Huson D, et al. The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. Nucleic Acids Research [Internet]. 2017 Jul [cited 2018 Jan 16];45(W1):W42–8. Available from: https: //academic.oup.com/nar/article/45/W1/W42/3787867

85. Martínez-Núñez MA, Rodríguez-Escamilla Z, Rodríguez-Vázquez K, Pérez-Rueda E. Tracing the Repertoire of Promiscuous Enzymes along the Metabolic Pathways in Archaeal Organisms. Life [Internet]. 2017 Jul [cited 2019 Jan 24];7(3). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5617955/

86. Charlesworth JC, Burns BP. Untapped Resources: Biotechnological Potential of Peptides and Secondary Metabolites in Archaea. Archaea [Internet]. 2015 Oct [cited 2016 Sep 27];2015:e282035. Available from: http://www.hindawi.com/journals/ archaea/2015/282035/abs/

87. Pearson H. Prehistoric proteins: Raising the dead. Nature News [Internet]. 2012 Mar [cited 2017 Feb 9];483(7390):390. Available from: http://www.nature.com/ news/prehistoric-proteins-raising-the-dead-1.10261

88. Treangen TJ, Rocha EPC. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. PLOS Genetics [Internet]. 2011 Jan [cited 2017 Feb 9];7(1):e1001284. Available from: http://journals.plos.org/ plosgenetics/article?id=10.1371/journal.pgen.1001284

89. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences [Internet]. 1999 Mar [cited 2017 Feb 9];96(6):2896–901. Available from: http://www.pnas.org/content/96/6/2896

90. Snel B, Lehmann G, Bork P, Huynen MA. STRING: A web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Research [Internet]. 2000 Sep [cited 2017 Feb 9];28(18):3442–4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC110752/

91. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: Rapid Annotations using Subsystems Technology. BMC Genomics [Internet]. 2008 Feb [cited 2017 Feb 7];9:75. Available from: http://www.ncbi.nlm.nih.gov/ 167 pmc/articles/PMC2265698/

92. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research [Internet]. 2014 Jan [cited 2017 Feb 7];42(Database issue):D206–14. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/ PMC3965101/

93. Harvey AL, Edrada-Ebel R, Quinn RJ. The re-emergence of natural products for drug discovery in the genomics era. Nature Reviews Drug Discovery [Internet]. 2015 Feb [cited 2016 Sep 16];14(2):111–29. Available from: http://www.nature. com/nrd/journal/v14/n2/full/nrd4510.html

94. Petrenko R, Meller J. Molecular Dynamics. In: eLS [Internet]. John Wiley & Sons, Ltd; 2001. Available from: http://onlinelibrary.wiley.com/doi/10. 1002/9780470015902.a0003048.pub2/abstract

95. Molecular Modeling of Proteins Andreas Kukol Springer [Internet]. [cited 2017 Feb 8]. Available from: http://www.springer.com/us/book/9781588298645

96. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. Journal of The Royal Society Interface [Internet]. 2014 Nov [cited 2017 Feb 9];11(100):20140419. Available from: http://rsif.royalsocietypublishing. org/content/11/100/20140419

97. Edgar RC. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research [Internet]. 2004 [cited 2018 May 27];32(5):1792–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC390337/

98. Castresana J. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. Molecular Biology and Evolution [Internet]. 2000 Apr [cited 2018 Jul 3];17(4):540–52. Available from: https://academic.oup.com/ mbe/article/17/4/540/1127654

99. Junier T, Zdobnov EM. The Newick utilities: High-throughput phylogenetic tree processing in the Unix shell. Bioinformatics [Internet]. 2010 Jul [cited 2018 Jul 3];26(13):1669–70. Available from: https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC2887050/

100. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. GROMACS: Fast, flexible, and free. Journal of Computational Chemistry [Internet]. 2005 Dec [cited 2017 Feb 9];26(16):1701–18. Available from: http: //onlinelibrary.wiley.com/doi/10.1002/jcc.20291/abstract

101. Odokonyero D, Sakai A, Patskovsky Y, Malashkevich VN, Fedorov AA, Bonanno JB, et al. Loss of quaternary structure is associated with rapid sequence divergence in the OSBS family. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 2014 Jun [cited 2017 Feb 9];111(23):8535–40. Available 168 Apéndice D. References from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4060685/

102. Rodríguez-Orduña L. Microbioma del jitomate: Determinante en la infección por Clavibacter michiganensis subsp. michiganensis [PhD thesis]. [Irapuato, GTO]: Langebio, CINVESTAV; 2016.

103. Gutiérrez-García K, Bustos-Díaz ED, Corona-Gómez JA, Ramos-Aboites HE, Sélem-Mojica N, Cruz-Morales P, et al. Cycad Coralloid Roots Contain Bacterial Communities Including Cianobacteria and Caulobacter spp. That Encode NicheSpecific Biosynthetic Gene Clusters. Genome Biology and Evolution [Internet]. 2019 Jan;11(1):319–34. Available from: https://academic.oup.com/gbe/article/11/ 1/319/5238076

104. Delgado-Suárez EJ, Selem-Mojica N, Ortiz-López R, Gebreyes WA, Allard MW, Barona-Gómez F, et al. Whole genome sequencing reveals widespread distribution of typhoidal toxin genes and VirB/D4 plasmids in bovine-associated nontyphoidal Salmonella. Scientific Reports [Internet]. 2018 Jun;8(1):9864. Available from: https: //doi.org/10.1038/s41598-018-28169-4

105. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 1977 Nov [cited 2017 Jan 23];74(11):5088–90. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC432104/

106. Labeda DP, Dunlap CA, Rong X, Huang Y, Doroghazi JR, Ju K-S, et al. Phylogenetic relationships in the family Streptomycetaceae using multi-locus sequence analysis. Antonie van Leeuwenhoek [Internet]. 2017 Apr [cited 2019 Mar 13];110(4):563–83. Available from: https://doi.org/10.1007/s10482-016-0824-0

107. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Research [Internet]. 2014 Jan;42(Database issue):D581–91. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965095/

108. Contreras-Moreira B, Vinuesa P. GET_HOMOLOGUES, a Versatile Software Package for Scalable and Robust Microbial Pangenome Analysis. Applied and Environmental Microbiology [Internet]. 2013 Dec;79(24):7696–701. Available from: https://aem.asm.org/content/79/24/7696

109. Chaudhari NM, Gupta VK, Dutta C. BPGA- an ultra-fast pan-genome analysis pipeline. Scientific Reports [Internet]. 2016 Apr [cited 2016 Dec 7];6:24373. Available from: http://www.nature.com/srep/2016/160413/srep24373/full/ srep24373.html

110. Navarro-Muñoz J, Selem-Mojica N, Mullowney M, Kautsar S, Tryon J, Parkinson E, et al. A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. bioRxiv [Internet]. 2018 Oct;445270. Available 169 from: https://www.biorxiv.org/content/early/2018/10/17/445270

111. Yasuhara-Bell J, Marrero G, Alvarez AM. Genes clvA, clvF and clvG are unique to Clavibacter michiganensis subsp. michiganensis and highly conserved. European Journal of Plant Pathology [Internet]. 2014 Dec [cited 2019 Mar 13];140(4):655–64. Available from: https://doi.org/10.1007/s10658-014-0495-5

112. Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. Nature Chemical Biology [Internet]. 2015 Sep [cited 2017 Feb 12];11(9):625–31. Available from: http://www. nature.com/nchembio/journal/v11/n9/full/nchembio.1890.html

113. Cruz-Morales P. Genómica funcional y evolutiva del metabolismo de *Streptomyces* [PhD thesis]. [Irapuato, GTO]: Langebio, CINVESTAV; 2013.

114. Garcia Pichel F, Sherry ND, Castenholz RW. Evidence for an ultraviolet sunscreen role of the extra cellular pigment scytonemin in the terrestrial cianobacterium Chiorogloeopsis sp. Photochemistry and Photobiology [Internet]. 1992 Jul;56(1):17–23. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751- 1097.1992.tb09596.x

115. Balskus EP, Walsh CT. Investigating the Initial Steps in the Biosynthesis of Cianobacterial Sunscreen Scytonemin. Journal of the American Chemical Society [Internet]. 2008 Nov;130(46):15260–1. Available from: https://doi.org/10.1021/ ja807192u

116. Soule T, Palmer K, Gao Q, Potrafka RM, Stout V, Garcia-Pichel F. A comparative genomics approach to understanding the biosynthesis of the sunscreen scytonemin in cianobacteria. BMC Genomics [Internet]. 2009 Jul [cited 2018 Jun 27];10:336. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2726228/

117. Engel PC. Glutamate Dehydrogenases: The Why and How of Coenzyme Specificity. Neurochemical Research [Internet]. 2014 Mar;39(3):426–32. Available from: https: //doi.org/10.1007/s11064-013-1089-x

118. Liu Y, Li Y, Wang X. Acetohydroxyacid synthases: Evolution, structure, and function. Applied Microbiology and Biotechnology [Internet]. 2016 Oct;100(20):8633– 49. Available from: https://doi.org/10.1007/s00253-016-7809-9

119. Weber T, Blin K, Duddela S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research [Internet]. 2015 Jul [cited 2017 Feb 7];43(W1):W237– 43. Available from: https://academic.oup.com/nar/article/43/W1/W237/ 2467910/antiSMASH-3-0-a-comprehensive-resource-for-the

120. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE [Internet]. 2010 Mar [cited 2018 Jul 3];5(3):e9490. Available from: http://journals.plos.org/plosone/article?id= 170 Apéndice D. References 10.1371/journal.pone.0009490

121. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, et al. Microreact: Visualizing and sharing data for genomic epidemiology and phylogeography. Microbial Genomics. 2016 Nov;2(11).

122. Charlesworth JC, Burns BP. Untapped Resources: Biotechnological Potential of Peptides and Secondary Metabolites in Archaea. Archaea [Internet]. 2015 Oct [cited 2016 Sep 27];2015:e282035. Available from: http://www.hindawi.com/journals/ archaea/2015/282035/abs/

123. Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV. Lineage-Specific Gene Expansions in Bacterial and Archaeal Genomes. Genome Research [Internet]. 2001 Apr;11(4):555–65. Available from: https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC311027/

124. Koonin EV, Wolf YI. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. Nucleic Acids Research [Internet]. 2008 Dec [cited 2017 Jan 26];36(21):6688–719. Available from: https://academic.oup.com/nar/article/36/21/6688/2410005/Genomicsof-bacteria-and-archaea-the-emerging

125. Kudo F, Kasama Y, Hirayama T, Eguchi T. Cloning of the Pactamycin Biosynthetic Gene Cluster and Characterization of a Crucial Glycosyltransferase Prior to a Unique Cyclopentane Ring Formation. The Journal of Antibiotics [Internet]. 2007 Aug;60(8):492–503. Available from: https://www.nature.com/articles/ ja200763

126. Andersson JO, Roger AJ. Evolution of glutamate dehydrogenase genes: Evidence for lateral gene transfer within and between prokaryotes and eukaryotes. BMC evolutionary biology. 2003 Jun;3:14.

127. Lilley KS, Baker PJ, Linda Britton K, Stillman TJ, Brown PE, Moir AJG, et al. The partial amino acid sequence of the NAD -dependent glutamate dehydrogenase of Clostridium symbiosum: Implications for the evolution and structural basis of coenzyme specificity. Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology [Internet]. 1991 Nov;1080(3):191–7. Available from: http: //www.sciencedirect.com/science/article/pii/016748389190001G

128. Consalvi V, Chiaraluce R, Politi L, Gambacorta A, Rosa MD, Scandurra R. Glutamate dehydrogenase from the thermoacidophilic archaebacterium Sulfolobus solfataricus. European Journal of Biochemistry [Internet]. 1991 Mar;196(2):459–67. Available from: https://febs.onlinelibrary.wiley.com/doi/abs/10.1111/j. 1432-1033.1991.tb15837.x

129. Ferrer J, Pérez-Pomares F, Bonete MJ. NADP-glutamate dehydrogenase from the halophilic archaeon Haloferax mediterranei: Enzyme purification, N-terminal 171 sequence and stability. FEMS microbiology letters. 1996 Jul;141(1):59–63.

130. Tholl D. Terpene synthases and the regulation, diversity and biological roles of terpene metabolism. Current Opinion in Plant Biology [Internet]. 2006 Jun;9(3):297–304. Available from: http://www.sciencedirect.com/science/ article/pii/S1369526606000537

131. Balskus EP, Walsh CT. The genetic and molecular basis for sunscreen biosynthesis in cianobacteria. Science (New York, NY) [Internet]. 2010 Sep;329(5999):1653–6. Available from: https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC3116657/

132. Hillwig ML, Fuhrman HA, Ittiamornkul K, Sevco TJ, Kwak DH, Liu X. Identification and Characterization of A Welwitindolinone Alkaloid Biosynthetic Gene Cluster in Stigonematalean Cianobacterium Hapalosiphon welwitschii. Chembiochem : a European journal of chemical biology [Internet]. 2014 Mar;15(5):665–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4382313/

133. Li S, Lowell AN, Yu F, Raveh A, Newmister SA, Bair N, et al. Hapalindole/Ambiguine Biogenesis Is Mediated by a Cope Rearrangement, C–C Bond-Forming Cascade. Journal of the American Chemical Society [Internet]. 2015 Dec;137(49):15366–9. Available from: https://www.ncbi.nlm.nih.gov/pmc/ articles/PMC4681624/

134. Li S, Lowell AN, Newmister SA, Yu F, Williams RM, Sherman DH. Decoding cyclase-dependent assembly of hapalindole and fischerindole alkaloids. Nature chemical biology [Internet]. 2017 May;13(5):467–9. Available from: https://www. ncbi.nlm.nih.gov/pmc/articles/PMC5391265/

135. Grant CS, Louda JW. Scytonemin-imine, a mahogany-colored UV/Vis sunscreen of cianobacteria exposed to intense solar radiation. Organic Geochemistry [Internet]. 2013 Dec;65:29–36. Available from: http://www.sciencedirect.com/science/ article/pii/S0146638013002179

136. Cruz-Morales P, Ramos-Aboites HE, Licona-Cassani C, Selem-Mójica N, MejíaPonce PM, Souza-Saldívar V, et al. Actinobacteria phylogenomics, selective isolation from an iron oligotrophic environment and siderophore functional characterization, unveil new desferrioxamine traits. FEMS Microbiology Ecology [Internet]. 2017 Sep;93(9). Available from: https://academic.oup.com/femsec/article/93/9/ fix086/3934648

137. McClure RA, Goering AW, Ju K-S, Baccile JA, Schroeder FC, Metcalf WW, et al. Elucidating the Rimosamide-Detoxin Natural Product Families and Their Biosynthesis Using Metabolite/Gene Cluster Correlations. ACS Chemical Biology [Internet]. 2016 Dec;11(12):3452–60. Available from: http://dx.doi.org/10. 1021/acschembio.6b00779

138. Pasek S, Risler J-L, Brézellec P. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. Bioinformatics [Inter- 172 Apéndice D. References net]. 2006 Jun;22(12):1418–23. Available from: https://academic.oup.com/ bioinformatics/article/22/12/1418/207642

139. Fani R, Brilli M, Liò P. The Origin and Evolution of Operons: The Piecewise Building of the Proteobacterial Histidine Operon. Journal of Molecular Evolution [Internet]. 2005 Mar;60(3):378–90. Available from: https://doi.org/10.1007/ s00239-004-0198-1

140. Fani R, Liò P, Chiarelli I, Bazzicalupo M. The evolution of the histidine biosynthetic genes in prokaryotes: A common ancestor for the hisA and hisF genes. Journal of Molecular Evolution [Internet]. 1994 May;38(5):489–95. Available from: https://doi.org/10.1007/BF00178849

141. Fani R, Tamburini E, Mori E, Lazcano A, Liò P, Barberio C, et al. Paralogous histidine biosynthetic genes: Evolutionary analysis of the Saccharomyces cerevisiae HIS6 and HIS7 genes. Gene [Internet]. 1997 Sep;197(1):9–17. Available from: http: //www.sciencedirect.com/science/article/pii/S0378111997001467

142. Merkl R, Sterner R. Reconstruction of ancestral enzymes. Perspectives in Science [Internet]. 2016 Dec;9:17–23. Available from: http://www.sciencedirect.com/ science/article/pii/S2213020916302336

143. Moustafa A, Loram JE, Hackett JD, Anderson DM, Plumley FG, Bhattacharya D. Origin of Saxitoxin Biosynthetic Genes in Cianobacteria. PLOS ONE [Internet]. 2009 Jun [cited 2016 Sep 16];4(6):e5758. Available from: http://journals.plos. org/plosone/article?id=10.1371/journal.pone.0005758

144. Gao Y, Xu G, Wu P, Liu J, Cai Y-s, Deng Z, et al. Biosynthesis of 20- Chloropentostatin and 20-Amino-20-Deoxyadenosine Highlights a Single Gene Cluster Responsible for Two Independent Pathways in *Actinomadura* sp. Strain ATCC 39365. Applied and Environmental Microbiology [Internet]. 2017 May;83(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5411499/

145. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLOS ONE [Internet]. 2011 Dec;6(12):e28766. Available from: http://journals.plos. org/plosone/article?id=10.1371/journal.pone.0028766

146. Kufareva I, Abagyan R. Methods of protein structure comparison. Methods in molecular biology (Clifton, NJ) [Internet]. 2012;857:231–57. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4321859/

147. Hopf TA, Green AG, Schubert B, Mersmann S, Schärfe CPI, Ingraham JB, et al. The EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics [Internet]. 2019 May;35(9):1582–4. Available from: https: //academic.oup.com/bioinformatics/article/35/9/1582/5124274

148. Kuper J, Doenges C, Wilmanns M. Two fold repeated (βα)4 half barrels may provide a molecular tool for dual substrate specificity. EMBO reports [Internet]. 173 2005 Feb;6(2):134–9. Available from: https://www.embopress.org/doi/abs/10. 1038/sj.embor.7400330

149. Due AV, Kuper J, Geerlof A, Kries JP von, Wilmanns M. Bisubstrate specificity in histidine/tryptophan biosynthesis isomerase from Mycobacterium tuberculosis by active site metamorphosis. Proceedings of the National Academy of Sciences of the United States of America [Internet]. 2011 Mar [cited 2017 Jan 31];108(9):3554–9. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3048130/

150. Reisinger B, Kuzmanovic N, Löffler P, Merkl R, König B, Sterner R. Exploiting Protein Symmetry To Design Light-Controllable Enzyme Inhibitors. Angewandte Chemie International Edition [Internet]. 2014 [cited 2017 Feb 3];53(2):595–8. Available from: http://onlinelibrary.wiley.com/doi/10.1002/anie.201307207/ abstract

151. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, et al. MINEs: Open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. Journal of Cheminformatics [Internet]. 2015 Dec [cited 2016 Sep 16];7(1). Available from: http://www.jcheminf. com/content/7/1/44