EvoMining 2.0: A customizable computational pipeline for evolutionary reconstructions during genome mining

Nelly Sélem-Mojica, Christian Martínez-Guerrero, César Aguilar, and Francisco Barona-Gómez

Evolution of Metabolic Diversity Laboratory, Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN, Irapuato, México.  
 Mitocondrial DNA and paleogenomics Laboratory, Unidad de Genómica Avanzada (Langebio), Cinvestav-IPN, Irapuato, México.

\*Corresponding author  
Email: [francisco.barona@cinvestav.mx](mailto:francisco.barona@cinvestav.mx) +52 473 652 02 46

**Keywords:** Genome mining, evolution, secondary metabolism, pipeline.

**Subject category:** Genomics and systems biology

**Word Count:** 5000

**Depositories:**  
Zenodo  
React  
**ABBREVIATIONS:**  
**GDH** Glutamate dehydrogenase  
**ASL** Acetolactate synthase large subunit

**DATA SUMMARY:**

## ABSTRACT

Genome mining has become a growing field due to genomic data abundance and to the importance of microbial natural products in human health and life. Traditional genome mining approaches explore genomes to recognize canonical domains of previously known secondary metabolism genes challenging to identify new enzymes involved in natural products biosynthesis. In 2016 we introduce EvoMining, a mindset that incorporates evolutionary theory into genome mining allowing to the discovery of enzymes from central metabolism repurposed into natural product biosynthesis. Software containers has recently been an answer to the challenge of develop and deploy analysis, allowing to convert inhouse scripts into a software package. EvoMining2.0 is a downloadable visual genome mining tool wraped in a container that incorporates the evolutive genome mining paradigm. This new version transforms EvoMining from a consulting site into a research tool by allowing customizable databases. One advantage is that every enzyme family is candidate to explore its expansion and recruitment events in the context of every prokaryotic genome lineage. In this study EvoMining was applied to several genomic databases such as Actinobacteria, Cyanobacteria, *Pseudomonas* and Archaea showing expansions and recruitments of 42 enzyme families conserved among this taxa. As a result it is presented glutamate dehydrogenase and acetolactate synthase, this enzymes does not belong to the core of Cyanobacteria but have been recruited into its secondary metabolism and some of this recruitments are exclusively detected by EvoMining. Evolutive genome mining methods complement traditional genome mining algorithms and open the door to discover not previously known chemical compounds at private genome collections.

## INTRODUCTION

Natural products are synthesized by biosynthetic gene clusters (NP-BGCs) distributed in the genome of a wide range of microorganisms. More than a thousand NP-BGCs are available at the community-driven repository called The Minimum Information about a Biosynthetic Gene cluster (MIBiG) [1]. With around 500,000 prokaryotic genomes available at public databases there has been a need for developing specialized genome mining approaches and related softwares. Traditional approaches are based in sequence similarity with enzymes devoted to secondary metabolism and their domains ([antiSMASH, cimermacic, clustrerFinder). previous biosynthetic knowledge. Recently, evolutionary process have been also exploited to mine genomes for novel BGCs (Cruz-Morales 2012, 2016, nadine). In these latter approaches, enzymes that belong to a BGC can either be mainly restricted to secondary metabolism, or be a recent recruitment acting as accessory enzymes, a concept has been further exploited in the context of antibiotic resistance (ARTS) .

Gene duplication and horizontal gene transfer constantly expand enzyme families. Extra copies in expanded families are recruited into secondary metabolism BGCs to perform novel chemical functionalities [REF scytonemin, moore]. EvoMining captures evolutive histories of enzyme families expansion and recruitment events. First, given an enzyme database (enzyme-DB), the algorithm identifies those expanded families in a given lineage (genomic-DB) that has at least a member in a natural products database. proceeding then to the reconstruction of the evolutionary history of the enzyme family. During this process enzymes from conserved metabolism (best bidirectional hits of enzyme database) are differentiated from expansions closest to the natural products database, i.e. putative novel recruitments into secondary metabolism. Further analysis of these hits includes visualization of the genomic vicinity guiding to the discovery of novel BGCs and allowing to prioritize biosynthetic darkmatter [Medema and Fischbach 2015, Blin et al. 2017].

Nevertheless, expansion and posterior recruitment is not an exclusive Actinobacteria feature, other microorganisms may also be explored evolutionary mined. Despite EvoMining analysis has recently being present on the natural products field (Blin et al. 2017,Alanjary et al. 2017,Ziemert, Alanjary, and Weber 2016, Miller, Chevrette, and Kwan 2017) EvoMining software has not been released.

EvoMining is a visual genome mining tool with the milestone of prioritize non standard secondary metabolite pathways by considering evolutionary principles. Software containers are transforming the way scientist exchange software. Prior to containers software analysis were difficult to instal and hard to reproduce. A proof of concept of EvoMining principles was provided by the discovery of an arseno compound on Streptomyces coelicolor (Cruz-Morales et al. 2016) by apply EvoMining into a genomic database of 230 Actinobacteria genomes.

Building up from EvoMining 1.0, which was a consulting web server with fixed databases, here we report EvoMining 2.0, a stand alone downloadable tool wrapped in docker. Taking advantage of the new feature of customizable databases, both the genomic and enzyme databases were enriched and explore to bring about novel biological insights. A systematic analysis of expansion and recruitment events in different genome-DBs, including Actinobacteria, Cyanobacteria, Pseudomonas and Archaea, is performed, with the finding that predictions are taxa-dependent and enzyme family specific. We also expanded on the main concept behind EvoMining, first, by analyzing as proof-of-concept the EvoMining profile of enzymes known to be part of BGCs in the model strains S. coelicolor and S. lividans, and second, by expanding the enzyme universe and introducing enzymes that belong to the so-called ‘shell genome’. Overall, these analyses suggest that….

## METHODS

### Databases and input options

To take advantage of functional annotation genome were processed by the Rapid Annotation using Subsystem Technology (RAST) platform [REF]. MIBiG repository is included in EvoMining docker image as default np-DB, metadata such as producer organism and kind of compound are also integrated to the np-DB. Sequences of the central-DB and np-DB must be on fasta format. To study the nature of expansions on other taxonomic groups beside Actinobacteria, three genomic databases were integrated by 1245 Actinobacteria, 416 Cyanobacteria, 876 Archaea, all genomes were annotated by RAST and mined by AntiSMASH stand alone docker parameters XXXXX. For this work Actinobacteria central-DB is composed by 295 Actinobacteria sequences of central enzymes organized in ninety-three families involved in nine pathways, including amino acid biosynthesis, glycolysis, pentose phosphate pathway, and tricarboxylic acids cycle. Seed sequences were taken mainly from 1) Streptomyces coelicolor A3, 2) Mycobacterium tuberculosis H37Rv and 3) Corynebacterium glutamicum. To integrate Cyanobacteria central-DB best bidirectional hits of Actinobacteria central-DB were selected among three organisms distributed on Cyanobacteria phylum 1) Cyanothece sp. ATCC 51142, 2) Synechococcus sp. PCC 7002 CP000951.1, 3) Synechocystis sp. PCC 6803 AP012205.1. Same procedure was taken to assemble Archaea central-DB, selecting organisms from diverse sections of Archaea kingdom, on this case 1) Natronomonas pharaonis, 2)Methanosarcina acetivorans, 3)Sulfolobus solfataricus and 4)Nanoarchaeum equitans Kin4-M. Expansions of enzyme families shared between this three taxonomic groups were calculated with EvoMining default parameters. Graphics for the data analysis on this work were generated by r, to guarantee reproducibility a rmarkdown document is available at: github/nselem/phd. Finally, EvoMining evolutionary analysis was conducted on PriA enzyme family on the three genomic database with default EvoMining parameters.

To amplify the concept of central-DB to shell-DB allowing to explore expansions beyond just enzymes belonging to the core genome of a taxonomic group a shell enzyme was explored. As am example of shell sequence tauD was selected as seed of central-DB. Pseudomonas genome-DB was integrated as follows (). Genomes on genome-DB were RAST annotated and on this case no antiSMASH mining was conducted. EvoMining expansion-recruitment and phylogenetic analysis with default parameters was conducted for TauD enzyme family.  
All databases are downloadable at Zenodo (Doi numbers genomic-DB: STREPTOMYCES, Actinobacteria, Cyanobacteria, Pseudomonas, Archaea, central-DB: Actinobacteria, Cyanobacteria, Pseudomonas, Archaea, lividans/coelicolor BGC )

### Expansion and recruitment analysis

Once databases are set, analysis of expansions-recruitment of enzyme families from enzyme-DB within the context of the genomic and natural product databases is performed by EvoMining. The first step retrieves from the genomic-DB those sequences similar to seed sequences from enzyme-DB. To this end a blastp search (default: e-value 0.001, score 100). Next it is established which sequences are the most conserved respect to the enzyme-DB, to this end best bidirectional hits between expanded families and enzyme-DB seeds are identified and stored in a temporal database (conserved-db). Next a heatplot pinpoints in each family those organisms that possess an expansion on its copy number. In this work an expansion is understood as those copies in an organism that possess a number of sequences above the average plus and standard deviation. To identify expanded enzyme families that may also have been recruited into secondary metabolism, i.e. which enzyme families has some members into the natural product database, a new blastp search (e-value 0.001) is run onto the np-DB using as queries all sequences on expanded families.

### Phylogenetic reconstruction and visualization

At this point families that exhibits expansions and recruitments may be selected by the user to posterior phylogenetic analysis. Selected expanded-recruited enzyme families are then aligned with muscle, automatically curated by Gblocks, and phylogenetically reconstructed by FastTree. Once the tree is available EvoMining gives evolutionary insights by coloring the expanded family differentiating its history as follows: in red enzyme sequences stored on the conserved-DB, in blue sequences from np-db, on cyan antiSMASH known hits if antiSMASH-DB was provided, and in green EvoMining predictions. An EvoMining prediction are all those enzymes divergent enough from central metabolism that were not red marked and similar enough to secondary metabolism that are closer to a np-DB sequence (blue) than to a conserved-db sequence (red). All enzymes not marked on red or cyan close to a blue enzyme and that are not on sister leaves with a red enzyme are green colored. This algorithm leaves black leaves on many cases.

EvoMining and all its dependencies: blast, muscle, Gblocks, FastTree and newick utilities are wrapped on the docker container nselem/newevomining downloadable at Dockerhub. Code is available at github: nselem/EvoMining and manual at <https://github.com/nselem/EvoMining/wiki>. Genomic, enzyme and natural products databases are downloadable at Zenodo (<Doi:10.5281/zenodo.1162336>). Optionally antiSMASH analysis of genomic DB may be integrated to the pipeline.

## RESULTS

### Section 1 EvoMining 2.0 Updated model and pipeline

In order to develop EvoMining 1.0 into a genome mining tool that allows visual inspection of data and analysis of large biological inputs, we aimed at making its databases customizable (Table 1). Technically, we opted to adopt a software container approach provided by Docker (see Methods). As a result, the three EvoMining inputs are (i) a genomic database (Genomic-DB), (ii) an enzyme database (Enzyme-DB), and (iii) a database of genes related to natural products pathways, or BGCs (NP-DB). These DBs are provided, but can be modified, replaced or expanded by the user (Table 1, Figure 1A). As in the previous proof-of-concept version 1.0, given an enzyme family from the central-DB, EvoMining produces an interactive, color coded tree of the expanded enzyme family, which provides information about the metabolic fate of each enzyme family member (Figure 1B). Building up from this, the newly incorporated changes permit to explore genome context, compare different taxa, and analyse more enzymes as queries.

Specifically, the provided genomic database in EvoMining 2.0 is a collection of genomes from taxonomically related organisms, including the phyla Actinobacteria and Cyanobacteria, the genus Pseudomonas, and the kingdom Archaea. The Actinobacteria database was expanded from 230 to 1244 genomes, which was complemented by integrating 876 Archaea genomes, 416 Cyanobacteria genomes and 219 Pseudomonas genomes. Selection of these taxa obeys to the possibility of analyzing both well known and recently recognised NPs producing organisms, such as in Actinobacteria (602 MIBiG BGCs), Cyanobacteria (60 MIBiG BGCs) and Pseudomonas (53 MIBiG BGCs); but also to compare these taxa with Archaea (0 MIBiG BGCs), which is not renowned because of it is ability to produce NPs, probably because it has been poorly investigated or the current tools and knowledges impose biases hampering our ability to discover specialized metabolites in this ancient lineage.

EvoMining original enzyme-DB contained 106 enzyme families from nine Actinobacteria central pathways previously curated (Barona-Gómez et al 20 12). To compare expansions of families from central metabolism on the new genomic databases the 106 families were filtered keeping only 42 families, those that are shared among Actinobacteria, Cyanobacteria, Pseudomonas and Archaea. The observation that some families were left aside leads to the conclusion that enzyme families from central metabolism may be on the core genome of some taxa but on the shell genome of another. Shell enzymes, defined on this work as enzymes present on the majority of the genomes, may also be subject to an expansion process that can be detected by EvoMining. In following sections, glutamate dehydrogenase is presented as an example of how phylogenetic histories of shared enzyme families may be highly depend of the genome-DB.

The last EvoMining input is the NP-DB, which previously included 226 manually curated NP-BGCs. This DB was improved by adopting The Minimum Information about a Biosynthetic Gene cluster (MIBiG) database [Medema et al. 2015], which is the result of a community systematic effort of linking a natural product with its BGC. Thus, default EvoMining NP-DB comprises all 22,276 sequences that belongs to some of the 1,296 reported NP-BGCs at MIBiG. However, it should be noticed that update of this DB still has to be done manually.

In synthesis, EvoMining tool 2.0 allow researchers to examine their own genomes and enzyme families in search of expansions involved on novel secondary or specialized metabolism. Novel insights described on table 1 were obtained after databases processing, and this insights will be detailed in the following sections: first, the genome-DB section inspects whether expansions and recruitments of enzyme families are lineage dependent. Second, in the enzyme-DB section our previous concept of central enzymes is extended to include not only core genes, like glutamate dehydrogenase, but also shell genes. TauD, a Pseudomonas and Streptomyces shell enzyme is showcased here as an example of this type of expansions and recruitments. Finally, on the section of NP-DB, EvoMining is applied retrospectively taking advantage of MIBiG existence by using the knowledge of which families are already known to be involved in NP biosynthesis.

### Section 2 Update and insights from genomic databases

EvoMining can be applied to bacterial taxonomic groups beyond Actinobacteria, but discovering novel recruitments due to the implementation of new genome-DBs will depend on the number and fate of expansions of the selected enzyme families. 2a (Figure 2A) y 2b (FIGURE 2B) Expansions of central Families depends on the lineage of the genomic-DB. As candidates to new genomic databases were selected phylum Cyanobacteria, genus Pseudomonas and kingdom Archaea. Pseudomonas and Cyanobacteria were chosen because of their known biosynthetic gene clusters content. Archaea is the opposite case, there is not archaeal BGC reported on MIBiG. Archaea resemble Bacteria in the facts that i) Archaea uses horizontal gene transfer as a genic interchange mecanism, ii) Archaeal genomes contains operons (???) and iii) in general there is introns absence{Reference to Computational Methods for Understanding Bacterial and Archaeal Genomes}. General lack of introns allows automatic genome annotation just like Bacteria, operon gene organization permits functional inference to a certain degree and HGT contribute to expansions on Archaeal genomes. Some phylum on Archaea has an open pangenome [REF fluidity], and as we will show some Archaea has central pathway expansions. This lead us to think that if evolutionary logic behave on Archaea as on bacteria evoMining is suitable to analyze Archaeal genomes. Considering that EvoMining is an evolutionary oriented method, not entirely based on previous knowledge of BGC’s sequences an unbiased analysis may be performed and new BGC’s classes may be found on Archaea.

Previous EvoMining genome-DB comprises 230 genomes from Actinobacteria with 106 central families as the corresponding central-DB (table 1). Candidate genomic databases were explored to determine whether or not enzymatic families devoted to central metabolism posses expansions on other taxa. Since some families in the core genome of a taxonomic group may not exists on another taxa, a comparative framework was needed. For this reason a common set of central enzyme families was identified on Actinobacteria, Cyanobacteria, Pseudomonas and Archaea. Homologues were set by finding best bidirectional hits (BBH) of original 339 enzyme-DB seeds belonging to the 106 central Actinobacteria enzyme families on the following genomes: 1) Cyanobacteria phylum: Cyanothece sp. ATCC 51142, Synechococcus sp. PCC 7002 CP000951.1, Synechocystis sp. PCC 6803 AP012205.1, 2) genus Pseudomonas: Pseudomonas fluorescens pf0-1, Pseudomonas protegens Pf5, Pseudomonas syringae CM002751.1, Pseudomonas fulva 12-X NC\_015556.1 3) Archaea kingdom Natronomonas pharaonis, Methanosarcina acetivorans, Sulfolobus solfataricus and Nanoarchaeum equitans Kin4-M. To avoid missing BBH due to genome quality selected genomes were chosen to be on one contig. The outcome of this process is a collection of 42 enzyme families common to all taxonomic groups. Foreach taxonomic group a central database with seeds within that group was assembled.

Using central and genomic databases as previously described, Actinobacteria, Cyanobacteria, Pseudomonas and Archaeas were explored in search of central enzyme families expansions. (Figure 2A). As it could be expected central enzyme family expansions correlates with genome size. However, the expansions rate increases in each genomic group and the increment is not linear. Expansions behave similar until a genome size of 5,000,000 bp on every testes taza, after that threshold the total number of sequences on the 42 expanded central families grows faster on Pseudomonas than on Actinobacteria, which in turn surpasses Cyanobacteria and Archaea (Figure 2). Cyanobacteria despite their bigger size is the taxa with less central expansions, which may be because their size is due to HGT or that are other families may be on shell genome the ones that are expanded. Nevertheless, some of the biggest Cyanobacterial genomes may be contaminated, for example, expansions of HisA on Hassallia bissoidea VB512170 when searched by blast on NCBI non redundant species database were closer to bacteroidetes than to any Cyanobacterial sequences (Supp Figure XX). Orders with bigger genome size and also bigger number of expansions were Streptomycetales and Nostocales corresponding to Actinobacteria and Cyanobacteria respectively. Archaea does not have reported until now genomes with size comparable with Streptomycetales or Pseudomonas but class Halobacteria is the one that shows more expansions.

Once determined a common set of central enzymes and shown that selected genomic databases actually posses expansions on this families, differences of the families across selected taxa can be explored.

Expansions and recruitments of the 42 common central families were quantified by EvoMining in Actinobacteria, Cyanobacteria, Pseudomonas and Archaea. True orthologues of the query enzymes are recognized by best bidirectional hits and marked as central metabolism sequences (red). Genes that belong to natural products BGCs are marked as blue when they are known and reported at MIBiG or cyan when they have been bioinformatically inferred by antiSMASH. Sequences recognized as both as an antiSMASH prediction and as a participant of central metabolism were marked in purple. Those sequences without a known metabolic fate were left black (Figure 2B).

While there are families like acetylornithine aminotransferase and acetolactate synthase strongly expanded in every group many others such as glutamate dehydrogenase are differentially expanded across taxa. Another example of differential behavior is the fumarate reductase iron sulfur subunit (C3 in figure XX) this family is expanded in Actinobacteria but reduced in Cyanobacteria. Pseudomonas has on average more copies than the other taxa in 54.8% of the 42 families explored, followed by Actinobacteria (26.2%) and lastly by Archaea and Cyanobacteria that ties in being the most expanded lineages only 9.5% of the time (SI table XX). The expansions strength in the different enzyme families, in general, maintain the proportion being Pseudomonas the most expanded group while Archaea stands as the least expanded, however, there are some families which do not accomplish this feature.

Glutamate dehydrogenase (GDH) is one of the 4 families that are most expanded in Archaea than in other taxa. In fact GDH has less than one copy by genome in the other taxa, because of that, this family is not part of the core of these lineages. Instead, GDH belongs to the shell genome of Actinobacteria, Cyanobacteria and Pseudomonas since is present in more than 50% of their genomes (Figure XX b). Regarding secondary metabolism hits in GDH, it is found that antiSMASH predictions are present in Actinobacteria, Cyanobacteria and Archaea but not in Pseudomonas, idem is certain for the recruitments scytonemin and pactamycin. This results shown that the appearance and evolution of the secondary metabolism is lineage dependent and that shell enzymes, just as core enzyme families, possess the potential to be expanded and recruited into natural products BGCs (Figure XX c).

Until now, it has been presented GDH EvoMining expansion/recruitment analysis, but the most attracting EvoMining feature consists in the phylogenetic visualisation. EvoMining trees provides EvoMining hits, those expansions closer to a MIBiG recruitment than to a central metabolism hit. Next will be shown GDH EvoMining trees contrasting Archaea were GDH is a core enzyme and Cyanobacteria where GDH is part of the shell genome.

## Section 3 Update and insights from central DB

A very interesting example of the evolution of the metabolic diversity is exposed in the glutamate dehydrogenase (GDH-2) family. GDH-2 catalyzes the reversible oxidative deamination of glutamate to -ketoglutarate and ammonia. This class of enzyme is hexameric, utilize either NAD+, NADP+ or both as a cofactor. GDH-2 is found in all domains of life and function mainly in ammonia assimilation. As it was mentioned, this diversified family was found common among our groups and in general presents few expansions events. A detailed examination of this family in each group shows that despite expansion events were not usual, Cyanobacteria and Archaea present some interesting expansions cases which has been recruited in secondary metabolic pathways.

The tree of the Archaea group was rooted with the query sequence of Sulfolobus islandicus, which is a NAD(P)+ utilizing enzyme (E.C. 1.4.1.3). Therefore, in the internal branches of the tree are located mainly dual cofactor acceptor related enzymes. More distant in the tree is possible to visualize small branches where are located NADP specific GDHs (E.C. 1.4.1.4), followed by NAD specific GDHs (E.C. 1.4.1.2). Those three isoforms of the enzymes alternate along the phylogenetic reconstruction, where become obvious a big mixed middle branch composed by NAD, and NAD(P) specific enzymes. Most of the best bidirectional hits, considered as central enzymes are distributed along the before branches independently of the cofactor they use. More divergent in the phylogeny, a large clade composed almost exclusively of NAD(P) specific enzymes is located. Only two enzymes recognized as central by EvoMining are in this clade. According to the gene context analysis, several enzymes appear in a genetic environment which points towards a potential recruitment by secondary metabolism. Most of these possible recruitments were identified in organisms from Haladaptatus, Haloterrigena, Natrialba, Natrinema, Natrialbaceae, and Natronococcus genus, and seems to be involved in the synthesis of terpenes. The functional annotation of those clusters shows many gene products as hypothetical proteins; however, the presence of enzymes related with glyceraldehyde 3-phosphate (fructose bisphosphate aldolase and fructose 1,6 bisphosphatase), a common precursor of the non-mevalonate pathway allows propose them as potential terpene BGCs.

Despite its higher divergency degree, the following two major branches in the tree correspond to enzymes identified as central by EvoMining. In the penultimate branch, EvoMining throws a potential hit, corresponding to a very divergent enzyme annotated as a leucine dehydrogenase from Candidate division, however, after the gene context analysis we do not find elements to ensure this gene as part of a potential natural product BGC.

In contrast with the overwhelmed GDH expansions in Archaea, the resulting Cyanobacteria’s tree shows only a small branch with few expansions: 4 antiSMASH hits and 4 EvoMining predictions. Interestingly this branch shows a recruitment from a cyanobacterial BGC that encodes for scytonemin. This metabolite is a yellow sunscreen pigment produced exclusively by some Cyanobacteria that protects them against UV-A radiation. The recruited gene in GDH tree is scyB, annotated as a leucine dehydrogenase with a domain of glutamate dehydrogenase [REF] (Figure XX b). Despite Nostoc punctiforme PCC 73102 is the organism whose scytonemin BGC is reported at MiBIG its gdh sequence is not part of the expansions close to scyB. This result suggests that evohits does not belong to scytonemin pathway but perhaps for another similar compound. Nevertheless, other genes are shared between scytonemin and the genomic vicinity of some evoHits.

#### Section 4 Update and insights from Natural products DB

The first gene in the synthesis of scytonemin is scyA that encodes for an acetolactate synthase large subunit (ALS), an enzyme family included in the 42 family enzymes explored in this work. Scrutinizing ALS EvoMining cyanobacterial tree it was found that scyA is a recruitment in this tree and that next to the recruitment is N. punctiforme PCC 73102 EvoMining hit. In the same branch are distinguish in green more than 20 exclusively evoMining hits including organisms that are known to produce scytonemin [REF]. Additionally, in ALS tree there is a branch that gathers ALS expansions marked as EvoMining or antisMASH hits that corresponds to genes in the vicinity of EvoMining hits in GDH tree in the scyB branch. This results shows a possible coevolution between expansions in scyA and scyB branches in ALS and GDH tree respectively. Besides GDH and ALS seven other genes in scytonemin BGC were also present in the 42 families in evoMining enzymeDB (Figure XX b). EvoMining trees were produced for this families and six of them contains expansions i.e. extra copies beyond copies devoted to primary metabolism that are recognized and correctly marked as EvoMining hits. The expansion branches in this trees have as mark of secondary metabolism recruitment the corresponding genes of scytonemin BGC. This hits shown how EvoMining can complement antiSMASH by identifying sequences that belongs to non traditional natural products BGCs. At the same time this example reinforces that expansions of primary metabolism are being recruited to secondary metabolism biosynthesis.

In Actinobacteria shows few expansions tree posses a well differentiated branch, Cyanobacteria also has a well differentiated branch, but members of that branch are the unique copy of their genomes, and are marked as central metabolism. On Pseudomonas glutamate dehydrogenase has on average less than one copy for genome, so its not part of the genus core, while in Archaea same family has expansions and recruitments.  
Recruitments are also taxa dependent, gdhA Pseudomonas no, other yes.  
This results showed that expansion and recruitment of a central family are clade specific.  
Databases used on this work are available at zenodo with doi number zenodo.1162336. After exploring genomic databases it is left to answer the existence of other enzymes besides core metabolism that may be under a detectable process of expansion - recruitment. Considering that some central enzymes for Actinobacteria where not part of the core on the other taxa the following section will explore other shell enzymes suitable as enzyme-DB.

A question often asked about EvoMining is how can be an enzyme-DB assembled for a certain taxonomic group. The purpose of the enzyme-DB is to propose a set of enzyme families were expansions maybe detected. However if a family is present only on a small percentage of genomes it will be hard to say that the family is being expanded. The property of being expanded, as we have seen in previous section depends on the choice of the genomic-DB. For this reason original EvoMining enzyme-DB conception was a database composed by enzymes that belong to central metabolism. Central metabolism is by definition metabolism shared among certain taxa, consequently sequences that codify enzymes participating on central metabolic pathways are shared among genomes.

Many enzymes of original EvoMining enzyme-DB were enzymes taken from metabolic reconstructions that happened to belong to Streptomyces core genome, nevertheless, once we expanded genomic-DB to all Actinobacteria, enzymes that were on the Streptomyces core now are located at Actinobacteria shell genome. Once noticed that the difference between core and shell genome depends on the taxa selected, it is logical to broad the conception of central metabolism and take instead conserved metabolism, considering both core and shell enzymes.

To obtain an estimation of the Streptomyces shell genome, genetic content of 41 Streptomyces closed genomes were partitioned into families. Those XXX families whose sequences were present on at least on 51% of organisms (21 genomes) were considered shell genomes. After EvoMining expansion - recruitment analysis XXX of them has expansions on at least 5% of the organisms (2 genomes). TauD, an enzyme known on taurine assimilation system on Enterobacteria [REF], was chosen as an example of EvoMining pipeline on Streptomyces shell enzymes. TauD was analyzed on Actinobacteria, Archaea, Cyanobacteria and Pseudomonas (Figura XX supp info). EvoMining tree with genomic-DB Pseudomonas shows in addition to the main branches related to taurine metabolism, a branch where genomic contexts of TauD includes NRPS and PKS, the enzymes that are classical marks of secondary metabolism BGCs. (Figure 3) Predicted structures without consider TauD function are shown [KARINA] Summarizing on this section it was successfully proposed calculate the shell genome as an algorithm to construct an interesting enzyme-DB given a taxonomic group of interest. Shell genome is not the only course to take when constructing an enzyme-DB, in fact CONECTION TO NEXT PART  
## DISCUSSION

During the decade between 1970 and 1980, Archaea was recognized as new life domain, a kingdom different from Bacteria and Eukarya in an exciting first great application of 16S phylogeny(???,(???)) . Main differences between this kingdoms are that Archaeal DNA is not arranged in a nucleus as in Eukarya and Archaea celular walls are not composed from peptidoglycans as in Bacteria. Archaeal proteins may be highly valuable to biotechnology industry for their great stability due to extreme temperature, PH and salt content conditions on Archeal habitats. Despite no Archaeal Natural products biosynthetic gene clusters (BGC’s) has been reported on MIBiG, Archaea do have BGC’s, some of them seems to be acquired by horizontal gene transfer (HGT) like methano nrps {search reference}. Other Archeal natural products known are archaeosins, Diketopiperazines, Acyl Homoserine Lactones, Exopolysaccharides, Carotenoids, Biosurfactants, Phenazines and Organic Solutes but this knowledge is not comparable to Bacterial BGC’s knowledge(???).

Natural products biosynthetic gene clusters search is actually performed using either high-confidence/low-novelty or low-confidence/high-novelty bioinformatic approaches (Medema and Fischbach 2015). High confidence methods compares query sequences with previously known BGC’s such as nrps or PKS, examples of this algorithms are antiSMASH and clusterfinder (????). EvoMining searches on expansions from central metabolic pathways enzyme families, it has been classified as low confidence/high novelty method. EvoMining has proved useful on Actinobacteria phylum where its use lead to Arseno-compounds discovery (Cruz-Morales et al. 2016). Also on Actinobacteria antiSMASH analysis on 1245 genomes found 774 different classes of natural products, the same analysis on 876 Archaeal genomes, a full kingdom, identifies only 35 BGC’s classes. So either Archaea does not have natural products BGC’s or this are not yet known. Next paragraph deals with a possible approach about how natural products BGC’s can be find.

EvoMining is a trade off between conserved known central metabolic function and enough expansions divergence on sequence and on clusters to divergence

## FUNDING INFORMATION

The research reported in this publication was supported by CONACyT (NSM, PhD scholarship 204482) and Innovation Secretary of Guanajuato through the program “Apoyo a Emprendedores de Programas de Formación de Capital Humano 2017”

## ACKNOWLEDGMENTS

Some analyses were carried out using CONABIO’s computing cluster, which was partially funded by Secretaría de Medio Ambiente y Recursos Naturales (SEMARNAT) through the grant “Contribución de la Biodiversidad para el Cambio Climático” to Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO). We express our gratitude to José Daniel Hernández for their technical support and to Pablo Cruz Morales for his academical advices. Argonne cluster, Cocina Tepeyac, Strain Biotech, DNAbits

## CONFLICTS OF INTEREST

There are no conflicts to declare.

# References

## Figures and tables

Genomic DB

|  |  |  |  |
| --- | --- | --- | --- |
| Feature | EvoMining 1 | EvoMining 2 | Novel insights |
| Code |  |  |  |
|  | Consulting web site | Standalone docker tool |  |
| Databases |  |  |  |
|  | Fixed | Customizable |  |
| Genomic-DB |  |  |  |
|  | Actinobacteria 230 genomes | Actinobacteria 1245 genomes | EvoMining predictions are lineage dependnet |
|  |  | Cyanobacteria 416 genomes |  |
|  |  | Archaea 876 genomes |  |
| Central-DB |  |  |  |
|  | Actinobacteria 106 families | Actinobacteria 106 families |  |
|  |  | Cyanobacteria 81 families |  |
|  |  | Archaea 80 families |  |
|  |  | RetroEvoMining XX num |  |
| Natural-DB |  |  |  |
|  | Manual curation | MiBIG |  |

Coelicolor

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Full…partial | Main.product | Biosynthetic.class | Organism | X..Backward.EvoMining.Hits | Open.closed |
| BGC0000038 | Full | coelimycin | Polyketide | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000194 | Full | actinorhodin | Polyketide | Streptomyces coelicolor A3(2) | NA | NA |
| GC0000315 | Full | calcium-dependent antibiotic | NRP | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000551 | Full | sapB | RiPP | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000595 | Full | SCO-2138 | RiPP | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000849 | Full | gamma-butyrolactone | Other | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000940 | Full | desferrioxamine B | Other | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000324 | Partial | coelibactin | NRP | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000325 | Partial | coelichelin | NRP | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000660 | Partial | albaflavenone | Terpene | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000663 | Partial | hopene | Terpene | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000910 | Partial | melanin | Other | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0000914 | Partial | methylenomycin | Other | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0001063 | Partial | undecylprodigiosin | NRP / Polyketide | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0001181 | Partial | geosmin | Terpene | Streptomyces coelicolor A3(2) | NA | NA |
| BGC0001168 | Full | livipeptin | NRP | Streptomyces lividans 1326 | NA | NA |
| BGC0001283 | Full | arsenopolyketides | Other | Streptomyces lividans 1326 | NA | NA |
| BGC0000596 | Full | SLI-2138 | RiPP | Streptomyces lividans TK24 | NA | NA |

1. **Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al.** Minimum Information about a Biosynthetic Gene cluster. *Nature Chemical Biology*. 2015;11[9]: 625–631. doi:[10.1038/nchembio.1890](https://doi.org/10.1038/nchembio.1890)