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.  
 Mitochondrial 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, EvoMining, secondary metabolism, pipeline.

**Subject category:** Genomics and systems biology

**Word Count:** 5000

**Depositories:**  
Zenodo  
React  
**ABBREVIATIONS:**  
**GDH** Glutamate dehydrogenase ; **ASL** Acetolactate synthase large subunit ; **EF** Enzyme families; **NP-BGCs** Natural product biosynthetic gene cluster; **DB** database;

**DATA SUMMARY:** Databases are downloadable at Zenodo (Doi numbers Genomic-DB: Actinobacteria, Cyanobacteria, *Pseudomonas*, Archaea, central-DB: Actinobacteria, Cyanobacteria, *Pseudomonas*, Archaea ) (<Doi:10.5281/zenodo.1162336>). Optionally antiSMASH analysis of genomic DB may be integrated to the pipeline.

## 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. Standar 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 wrapped 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. 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 is a continuous need for developing specialized genome mining approaches with its corresponding softwares. Traditional approaches of NP-BGCs identification are based in sequence similarity with enzymes devoted to secondary metabolism and their domains [2],[3]. Though this platforms has been very successful they rely in previous secondary metabolism biosynthetic knowledge without addressing the problem of discover the involvement of new enzymes in NP-BGCs. Recently, evolutionary logic have been incorporated into genome mining [4],[5], 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 that has been further exploited in the context of antibiotic resistance (ARTS).

Accessory genes in NP-BGCs may be extra copies in expanded families recruited into secondary metabolism BGCs [6][7]. Expansions in enzyme families (EF) are extra copies due to gene duplication and horizontal gene transfer retained in genomes to perform novel chemical functionalities [8]. EvoMining logic follows enzyme families evolutive expansion and recruitment events as follows. First, given an enzyme database (Enzyme-DB) the algorithm identifies those expanded families in a given lineage (Genomic-DB) similar to a member in a natural products biosynthetic genes database (NP-DB). Next a phylogenetic reconstruction of the EF culminates differentiating from sequences that are best bidirectional hits of Enzyme-DB and are presumingly devoted to conserved metabolism from those expansions closest to the natural products database i.e. copies that are putative novel recruitments into secondary metabolism. In our previous work EvoMining principles succeed in the discovery, in *Streptomyces coelicolor*, of the recruitment of an expansion of an actinobacterial core enzyme in an arseno compound synthesis [4].

Despite EvoMining mindset has recently being present in the natural products community [9],[5],[10],[11] the corresponding software has not been released. Software containers such as conda and docker are transforming the way scientists exchange software, currently thousands of bioinformatic containers that includes the software with its dependencies and data with metadata facilitates the development of reproducible bioinformatic pipelines [12]. Prior to containers scientific software consists mainly in inhouse scripts difficult to instal which was translated into analysis hard to reproduce. Here we present EvoMining 2.0 a visual genome mining tool downloadable as a docker container with the milestone of prioritize non standard secondary metabolite pathways by considering evolutionary principles. Further analysis of EvoMining hits includes visualization of the genomic vicinity guiding to the discovery of novel BGCs and allowing to prioritize biosynthetic darkmatter [13], [9].

In our previous work EvoMining was applied into a genomic database of 230 Actinobacteria genomes. Nevertheless expansion and posterior recruitment is not an exclusive Actinobacteria feature, in consequence other microorganisms may also be evolutionary mined. Taking advantage of the new feature of customizable databases both genome-DB, Enzyme-DB and NP-DB databases were enriched and explored to bring about novel biological insights. Here is presented a systematic analysis of expansion and recruitment events in different genome-DBs including Actinobacteria, Cyanobacteria, *Pseudomonas* and Archaea. Scytonemin is a pigment exclusively produced by Cyanobacteria with medical and pharmaceutical importance [14]. The recruited housekeeping genes in scytonemin NP-BGC are examples of expansions devoted to secondary metabolism that are exclusively detected by EvoMining. Remarkably, glutamate dehydrogenase (GDH-2) and acetolactate synthase (ALS) are families within scytonemin BGC not shared by all Cyanobacteria that still possess expansions and recruitments. This observation transforms the main concept behind EvoMining expanding the enzyme universe from core enzymes driving attention instead into enzymes that belong to the so-called ‘shell genome’, shared by the majority but not all genomes in a taxonomic group. Overall, these analyses suggest that EF recruitments are taxa-dependent in each family and that non traditional approaches complements standard genome mining.

## METHODS

### Input databases

Previous EvoMining genome-DB comprises 230 genomes from Actinobacteria with 106 central families as the corresponding central-DB (table 1). In this work Actinobacteria Genomic-DB was expanded to 1245 genomes, additionally three new Genomic-DB were integrated by 416 Cyanobacteria, 219 *Pseudomonas* and 876 Archaea. Genomes in Genomic-DBs were functionally annotated by the Rapid Annotation using Subsystem Technology (RAST) [15] and mined by antiSMASH docker [16] (–cf\_threshold 0.7) storing the results in the optional antiSMASH-DB.

Actinobacteria Enzyme-DB in our previous work was composed by 339 Actinobacteria sequences of enzymes organized in 106 families involved in nine central 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*. 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 EF was identified in Actinobacteria, Cyanobacteria, *Pseudomonas* and Archaea.

To integrate Cyanobacteria Enzyme-DB best bidirectional hits of Actinobacteria 106 families in original Enzyme-DB were determined in Cyanobacteria using Metaphor [17] in selected genomes. The organisms selected were: *Cyanothece sp. ATCC 51142*, *Synechococcus sp. PCC 7002 CP000951.1*, *Synechocystis sp. PCC 6803 AP012205.1*. The same procedure was followed in genus Pseudomonas with *Pseudomonas fluorescens pf0-1*, *Pseudomonas protegens Pf5*, *Pseudomonas syringae CM002751.1*, *Pseudomonas fulva 12-X NC\_015556.1* and in Archaea kingdom with *Natronomonas pharaonis*, *Methanosarcina acetivorans*, *Sulfolobus solfataricus* and *Nanoarchaeum equitans Kin4-M*. Selected genomes were chosen to be in one contig to avoid missing BBH due to genome quality. The outcome of this process is a collection of 42 EF common to these taxonomic groups.

The last EvoMining input is the NP-DB, which previously included 226 manually curated NP-BGCs. This DB was improved by adopting MIBiG database [1], 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 in 2017.

### Expansion and recruitment analysis

The first EvoMining output is an expansion-recruitment analysis of EF. Expanded families are obtained by running a blastp search between EF in Enzyme-DB against the Genomic-DB (e-value 0.001, score 100). Next, the most conserved sequences in expanded families respect are filtered by best bidirectional against Enzyme-DB to posteriorly store them in a temporal database (conserved-db). This information is gather in a heatplot that pinpoints in each family those organisms that possess an expansion, understood here as the extra copies that are above the average plus a standard deviation of the copy number of the lineage. Finally, recruitments into secondary metabolism are determined with a new blastp search with query the expanded family against the NP-DB (e-value 0.001).

### Phylogenetic reconstruction and visualization

At this point EF that exhibits expansions and recruitments are aligned with muscle[18], automatically curated with 5 positions as minimum block length, 10 as maximum number of contiguous non-conserved positions and only considering positions with a gap in less than 50% of the sequences in the final alignment using Gblocks[19]. Curated alignment is then phylogenetically reconstructed by FastTree[20]. In the trees EvoMining gives evolutionary insights by coloring the expanded family differentially according to the metabolic fates of each copy as follows: in red are colored sequences stored in the conserved-DB, in cyan antiSMASH hits whenever antiSMASH-DB was provided, in purple those sequences that are at the same time antiSMASH hits and members of the conserved-DB, in blue recruitments from NP-DB, and in green EvoMining predictions. An EvoMining prediction are those sequences that are divergent enough from central metabolism and therefore were not red marked and at the same time are similar enough to secondary metabolism that are closer to a NP-DB sequence (blue) than to a conserved-db sequence (red). Leaves without color are kept black and are considered expansions with not known metabolic destination. Tree coloration is carry on by the Newick utilities [21]

EvoMining and its dependencies: blast, muscle, Gblocks, FastTree and Newick utilities are wrapped in the docker container nselem/newevomining downloadable at Dockerhub. Code is available at github: nselem/EvoMining and manual at <https://github.com/nselem/EvoMining/wiki>. Interactive notebook at github/nselem/phd.

## RESULTS

### 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 (i) Genomic-DB, (ii) Enzyme-DB and (iii) NP-DB are provided, but can be modified, replaced or expanded by the user (Table 1), Fig. 1(a). As in our previous proof-of-concept version 1.0[4], given an EF from the Enzyme-DB, EvoMining produces an interactive, color coded tree of the expanded enzyme family, which provides information about the metabolic fate of each EF member Fig. 1(b). 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-DBs in EvoMining 2.0 are collection of genomes of taxonomically related organisms, including the phyla Actinobacteria and Cyanobacteria, the genus Pseudomonas, and the kingdom Archaea. 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 [6]. To compare expansions in families from central metabolism in the new genomic databases the 106 families were filtered keeping only 42 families shared among Actinobacteria, Cyanobacteria, *Pseudomonas* and Archaea (see methods). The observation that some families were left aside leads to the conclusion that enzyme families from central metabolism and in the core genome of some taxa may belong to the shell genome of another. Shell enzymes, defined on this work as enzymes present in the majority of the genomes, may also be subject to an expansion process that can be detected by EvoMining. Candidate genomic databases were explored to determine whether or not enzymatic families devoted to central metabolism posses expansions on other taxa. In following sections, glutamate dehydrogenase (GDH-2) is presented as an example of how phylogenetic histories of shared enzyme families may be highly depend on the Genome-DB.

In synthesis, EvoMining tool 2.0 allows researchers to examine their own genomes and EF in search of expansions involved in novel 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 conserved enzymes and includes not only core genes like GDH-2 in Actinobacteria, but also shell genes as GDH-2 in Cyanobacteria.

### Update and insights from genomic databases

EvoMining can be applied to taxonomic lineages beyond Actinobacteria as long as there are annotated genomes available. As it was mentioned Cyanobacteria and *Pseudomonas* genome-DBs were included here because of this organisms biosynthetical potential. Archaea in contrast is an unexplored kingdom where EvoMining may succeed in mining because Archaea resemble Bacteria in the facts that i) Archaea uses horizontal gene transfer as a genic interchange [22],[23],[24], ii) Archaeal genomes contains operons that helps to identify functional roll [24], iii) in general there is introns absence which allows automatic genome annotation[24] and iv) Archaea produces natural products even tough there are not genes reported[25].

Some Archaea phyla has an open pangenome [26], and as we will show Archaea has expansions in EF in central pathways. Considering that EvoMining is an evolutionary oriented method not entirely based on previous knowledge of NP-BGCs an unbiased analysis may be performed and new NP-BGCs classes may be found. Ultimately discovering novel recruitments will depend on the number and fate of expansions in the EF that conforms the Enzyme-DB. Using enzyme and genomic databases as previously described Fig. 2(a), Actinobacteria, Cyanobacteria, *Pseudomonas* and Archaeas were explored in search of EF expansions Fig. 2(b).

As it could be expected Fig.3 shows that central enzyme family expansions correlates with genome size, however, the expansions rate increment is different in each genomic group and this increment is not linear. Expansions behave similar until a genome size of 5,000,000 bp in every lineage. After that threshold the total number of sequences in the 42 expanded EF grows faster in *Pseudomonas* than in Actinobacteria, which in turn surpasses Cyanobacteria and Archaea Fig. 2(b). Until now there are no Archaea genomes reported with size comparable to *Streptomycetales* or *Pseudomonas* [23]. Cyanobacteria despite their big genome size is the taxa with less expansions possibly because there are other families besides the 42 considered here the ones that are expanded. Orders with biggest genome size and also greatest number of expansions were *Streptomycetales* and *Nostocales* in Actinobacteria and Cyanobacteria respectively while Halobacteria is the most expanded class in Archaea.

After shown that selected genomic-DB actually posses expansions of the common set of EF, now differences across the expansion and recruitment patterns across taxa of the 42 families can be explored. In Fig, 3 true orthologues of the query enzymes are marked as conserved metabolism sequences (red), genes that belong to natural products BGCs are shown in blue when they are reported at MIBiG or in cyan whenever they have been bioinformatically inferred by antiSMASH. Sequences recognized as both an antiSMASH prediction and a member of conserved metabolism were marked in purple. Those sequences without a known metabolic fate were left black. While there are families like acetylornithine aminotransferase and acetolactate synthase strongly expanded in every group many others such as GDH-2 are differentially expanded across taxa. Another example of differential behavior is the fumarate reductase iron sulfur subunit, C3 coordinate in Fig. 3, this family is expanded in Actinobacteria but reduced in Cyanobacteria. *Pseudomonas* has on average more copies by genome than other taxa in 54.8% of the 42 EF explored, it is 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 expansion strength in the different EF, in general, maintains 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. GDH-2 is one of the 4 families that are more expanded in Archaea than in other taxa. In fact GDH-2 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-2 belongs to the shell genome of Actinobacteria, Cyanobacteria and *Pseudomonas* since it is present in more than 50% of their genomes, Fig. 3(b). Regarding secondary metabolism hits in GDH-2, it is found that antiSMASH predictions are present in Actinobacteria, Cyanobacteria and Archaea but not in *Pseudomonas*, idem is certain for the recruitments from the BGCs 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 Fig. 3(c).

Until now, it has been presented GDH-2 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, GDH-2 EvoMining trees will be shown, contrasting Archaea were GDH-2 is a core enzyme and Cyanobacteria where GDH-2 is part of the shell genome.

### Update and insights from central DB

A very interesting example of evolution of the metabolic diversity is exposed in the GDH-2 family which catalyzes the reversible oxidative deamination of glutamate to -ketoglutarate and ammonia. This class of enzyme is hexameric and utilizes either NAD+, NADP+ or both as a cofactor. GDH-2 is found in all domains of life and function mainly in ammonia assimilation [27],[28],[29],[30]. As it was mentioned, this diversified family was found common among our genomic groups and in general presents few expansions events. A detailed examination of this family shows that despite expansion events were not usual Fig. 4(a), Cyanobacteria and Archaea present some interesting expansions cases that have been recruited in secondary metabolic pathways.

The tree of the Archaea group, Fig. 4(a) 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 (SFig XX). 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 some antiSMASH infered terpenes BGCs and of enzymes related with glyceraldehyde 3-phosphate (fructose bisphosphate aldolase and fructose 1,6 bisphosphatase), a common precursor of the non-mevalonate pathway allow 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-2 expansions in Archaea, the resulting Cyanobacteria’s tree shows only a small branch with few expansions ditributed in 4 antiSMASH hits and 4 EvoMining predictions Fig. 4(b). 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 [14],[31]. The recruited gene in GDH tree is *scyB*, annotated as a leucine dehydrogenase with a domain of glutamate dehydrogenase [7]. Despite *Nostoc punctiforme PCC 73102* is the organism whose scytonemin BGC is reported at MiBIG its gdh sequence is not part of the expansion branch close to *scyB*. This result suggests that evohits in this branch does not belong to scytonemin route but perhaps to the pathway of another similar compound. Nevertheless, in the genomic vicinity of some evoHits are located sequences similar to *scyA* the first gene of scytonemin BGC.

Scytonemin gene *scyA* encodes for an acetolactate synthase large subunit (ALS), an enzyme family included in the 42 EF in the Enzyme-DB. Scrutinizing ALS EvoMining cyanobacterial tree it was found that *scyA* as a recruitment in this tree and that next to the recruitment is an ALS sequence from *N. punctiforme PCC 73102* indicated as EvoMining hit. In the same branch are distinguish in green more than 20 exclusively evoMining hits including organisms that are known to produce scytonemin [14],[7].

Additionally, next branch in ALS tree shows only antiSMASH hits. The following branch gathers ALS expansions marked either as EvoMining or antisMASH hits, interestingly organisms in this branch corresponds to genes in the vicinity of GDH-2 EvoMining in the *scyB* branch. This results shows a possible coevolution between expansions in *scyA* and *scyB* branches in ALS and GDH-2 tree respectively, even tough sequence similarity between GDH-2 family and homologues of *scyB* is not enough to recover a *scyB* branch as an expansion branch in GDH-2 EvoMining tree. Besides GDH-2 and ALS, in Fig. 4(b) are marked seven other genes in scytonemin BGC that are also present in the 42 EF in evoMining Enzyme-DB. From the this seven families, six EvoMining trees contains expansions i.e. extra copies beyond copies devoted to primary metabolism that are recognized and correctly marked as EvoMining hits. Expansion branches in this trees have the corresponding genes of scytonemin BGC as mark of secondary metabolism recruitment. This hits ilustrate how EvoMining can complement antiSMASH by identifying sequences that belongs to non traditional natural products BGCs such as scytonemin which lacks of PKS or NRPS enzymes. At the same time this example corroborates that expansions of primary metabolism are being recruited into secondary metabolism biosynthesis.

## DISCUSSION

A question often asked about EvoMining is how it is assembled an Enzyme-DB 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. For this reason original EvoMining Enzyme-DB conception was a database composed by enzymes that belong to central metabolism. Nevertheless, how central is central? The difference between core and shell genome depends on the taxa selected, whit this observation it is logical to broad the conception of central metabolism and take instead the conserved metabolism as Enzyme-DB i.e. considering both core and shell enzymes.

Another key point in EvoMining success is the improvement of the NP-DB due to the availability of a repository of genes that belongs to NP-BGCs so comprehensively and carefully curated as it is MIBiG. Previous EvoMinig version does not include cyanobacterial NP-BGCs and for that reason it would have fail to indicate EvoMining hits in the *scyA* and *scyB* branches of Cyanobacterial trees. Nevertheless, despite the metabolic fate of secondary metabolism would not be estabished, this sequences would have been marked as expansions not in central metabolism and this may be the case in Archaea. Thoug some sequences in GDH-2 tree are expansions that may be terpenes this expansions are not marked as EvoMining hits, this situation will be improved as long archeal NP-BGC are get to known, if they exists.

Summarizing in this work EvoMining was developed as a docker container, allowing customizable databases. EvoMining2.0 was applyed to selected Enzyme-DBs families common to Actinobacteria, Cyanobacteria, *Pseudomonas* and Archaea leading to the conclussion that expansion and recruitment are EF and genomic lineage dependent. Finally, starting with GDH-2 an enzyme more expanded in Archaea than in the other taxa it was shown that GDH-2, even with less than one copy by genome in Cyanobacteria and with only a few expansions, constitutes an example of a gene in a NP-BGCs exclusively detected by EvoMinig. This results suggests that EvoMining is a functional low-confidence/high-novelty bioinformatic genome mining algorithm that complements traditional genome mining.

## 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

# References

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)

2. **Medema MH, Blin K, Cimermancic P, Jager V de, Zakrzewski P, Fischbach MA, et al.** antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*. 2011;39[Web Server issue]: W339–W346. doi:[10.1093/nar/gkr466](https://doi.org/10.1093/nar/gkr466)

3. **Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al.** Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*. 2014;158[2]: 412–421. doi:[10.1016/j.cell.2014.06.034](https://doi.org/10.1016/j.cell.2014.06.034)

4. **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*. 2016;8[6]: 1906–1916. doi:[10.1093/gbe/evw125](https://doi.org/10.1093/gbe/evw125)

5. **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*. 2017;45[W1]: W42–W48. doi:[10.1093/nar/gkx360](https://doi.org/10.1093/nar/gkx360)

6. **Barona-Gómez F, Cruz-Morales P, Noda-García L**. What can genome-scale metabolic network reconstructions do for prokaryotic systematics? *Antonie van Leeuwenhoek*. 2012;101[1]: 35–43. doi:[10.1007/s10482-011-9655-1](https://doi.org/10.1007/s10482-011-9655-1)

7. **Soule T, Palmer K, Gao Q, Potrafka RM, Stout V, Garcia-Pichel F**. A comparative genomics approach to understanding the biosynthesis of the sunscreen scytonemin in cyanobacteria. *BMC Genomics*. 2009;10: 336. doi:[10.1186/1471-2164-10-336](https://doi.org/10.1186/1471-2164-10-336)

8. **Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mittenthal JE**. The origin and evolution of modern metabolism. *The International Journal of Biochemistry & Cell Biology*. 2009;41[2]: 285–297. doi:[10.1016/j.biocel.2008.08.022](https://doi.org/10.1016/j.biocel.2008.08.022)

9. **Blin K, Kim HU, Medema MH, Weber T**. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Briefings in Bioinformatics*. doi:[10.1093/bib/bbx146](https://doi.org/10.1093/bib/bbx146)

10. **Weber T, Kim HU**. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. *Synthetic and Systems Biotechnology*. 2016;1[2]: 69–79. doi:[10.1016/j.synbio.2015.12.002](https://doi.org/10.1016/j.synbio.2015.12.002)

11. **Miller IJ, Chevrette MG, Kwan JC**. Interpreting Microbial Biosynthesis in the Genomic Age: Biological and Practical Considerations. *Marine Drugs*. 2017;15[6]: 165. doi:[10.3390/md15060165](https://doi.org/10.3390/md15060165)

12. **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*. 2018;7: 742. doi:[10.12688/f1000research.15140.1](https://doi.org/10.12688/f1000research.15140.1)

13. **Medema MH, Fischbach MA**. Computational approaches to natural product discovery. *Nature Chemical Biology*. 2015;11[9]: 639–648. doi:[10.1038/nchembio.1884](https://doi.org/10.1038/nchembio.1884)

14. **Rastogi RP, Sonani RR, Madamwar D**. Cyanobacterial Sunscreen Scytonemin: Role in Photoprotection and Biomedical Research. *Applied Biochemistry and Biotechnology*. 2015;176[6]: 1551–1563. doi:[10.1007/s12010-015-1676-1](https://doi.org/10.1007/s12010-015-1676-1)

15. **Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al.** The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*. 2008;9: 75. doi:[10.1186/1471-2164-9-75](https://doi.org/10.1186/1471-2164-9-75)

16. **Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al.** antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Research*. 2017;45[Web Server issue]: W36–W41. doi:[10.1093/nar/gkx319](https://doi.org/10.1093/nar/gkx319)

17. **Veen BE van der, Harris HM, O´Toole PW, Claesson MJ**. Metaphor: Finding Bi-directional Best Hit homology relationships in (meta)genomic datasets. *Genomics*. 2014;104[6, Part B]: 459–463. doi:[10.1016/j.ygeno.2014.10.008](https://doi.org/10.1016/j.ygeno.2014.10.008)

18. **Edgar RC**. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*. 2004;32[5]: 1792–1797. doi:[10.1093/nar/gkh340](https://doi.org/10.1093/nar/gkh340)

19. **Castresana J**. Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*. 2000;17[4]: 540–552. doi:[10.1093/oxfordjournals.molbev.a026334](https://doi.org/10.1093/oxfordjournals.molbev.a026334)

20. **Price MN, Dehal PS, Arkin AP**. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*. 2010;5[3]: e9490. doi:[10.1371/journal.pone.0009490](https://doi.org/10.1371/journal.pone.0009490)

21. **Junier T, Zdobnov EM**. The Newick utilities: High-throughput phylogenetic tree processing in the Unix shell. *Bioinformatics*. 2010;26[13]: 1669–1670. doi:[10.1093/bioinformatics/btq243](https://doi.org/10.1093/bioinformatics/btq243)

22. **Treangen TJ, Rocha EPC**. Horizontal Transfer, Not Duplication, Drives the Expansion of Protein Families in Prokaryotes. *PLOS Genetics*. 2011;7[1]: e1001284. doi:[10.1371/journal.pgen.1001284](https://doi.org/10.1371/journal.pgen.1001284)

23. **Koonin EV**. The Turbulent Network Dynamics of Microbial Evolution and the Statistical Tree of Life. *Journal of Molecular Evolution*. 2015;80[5-6]: 244–250. doi:[10.1007/s00239-015-9679-7](https://doi.org/10.1007/s00239-015-9679-7)

24. **Koonin EV, Wolf YI**. Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. *Nucleic Acids Research*. 2008;36[21]: 6688–6719. doi:[10.1093/nar/gkn668](https://doi.org/10.1093/nar/gkn668)

25. **Charlesworth JC, Burns BP**. Untapped Resources: Biotechnological Potential of Peptides and Secondary Metabolites in Archaea. *Archaea*. 2015;2015: e282035. doi:[10.1155/2015/282035](https://doi.org/10.1155/2015/282035)

26. **Kislyuk AO, Haegeman B, Bergman NH, Weitz JS**. Genomic fluidity: An integrative view of gene diversity within microbial populations. *BMC Genomics*. 2011;12: 32. doi:[10.1186/1471-2164-12-32](https://doi.org/10.1186/1471-2164-12-32)

27. **Benachenhou-Lahfa N, Forterre P, Labedan B**. Evolution of glutamate dehydrogenase genes: Evidence for two paralogous protein families and unusual branching patterns of the archaebacteria in the universal tree of life. *Journal of Molecular Evolution*. 1993;36[4]: 335–346.

28. **Andersson JO, Roger AJ**. Evolution of glutamate dehydrogenase genes: Evidence for lateral gene transfer within and between prokaryotes and eukaryotes. *BMC evolutionary biology*. 2003;3: 14. doi:[10.1186/1471-2148-3-14](https://doi.org/10.1186/1471-2148-3-14)

29. **Brown JR, Doolittle WF**. Archaea and the prokaryote-to-eukaryote transition. *Microbiology and molecular biology reviews: MMBR*. 1997;61[4]: 456–502.

30. **Miñambres B, Olivera ER, Jensen RA, Luengo JM**. A new class of glutamate dehydrogenases (GDH). Biochemical and genetic characterization of the first member, the AMP-requiring NAD-specific GDH of Streptomyces clavuligerus. *The Journal of Biological Chemistry*. 2000;275[50]: 39529–39542. doi:[10.1074/jbc.M005136200](https://doi.org/10.1074/jbc.M005136200)

31. **Sorrels CM, Proteau PJ, Gerwick WH**. Organization, Evolution, and Expression Analysis of the Biosynthetic Gene Cluster for Scytonemin, a Cyanobacterial UV-Absorbing Pigment. *Applied and Environmental Microbiology*. 2009;75[14]: 4861–4869. doi:[10.1128/AEM.02508-08](https://doi.org/10.1128/AEM.02508-08)