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

Selem-Mojica Nelly, Martínez-Guerrero Christian , …, and Barona-Gómez Francisco

March, 2018

## Abstract

Microbial natural products has importance in human health and life. Due to the abundance of genomic and metagenomic data, new natural products research by genome mining is a growing field. Traditional genome mining approaches explored bacterial genomes localizing marks of previously knwon secondary metabolism enzymes organized on biosynthetic gene clusters (BGCs). Here we present EvoMining a downloadable visual genome mining tool that incorporates evolution theory into genome mining. On EvoMining databases are customizable, its based on enzyme expansions not on BGCs. The advantage of this method is that every expanded enzyme family is a candidate to explore recruitments, and all prokatyiotic genome, even the unexplored Archaea kingdom. On this study EvoMining was applied to several database such as Cyanobacteria, Actinobacteria, Pseudomonas and Archaea studying expansions for enzyme families such as PriA,TauD and other enzymes recently recruited onto secondary metabolism. This Evolutionary methods open the door to discover not previously knwon chemical compounds at private genome collections and prioritize them according to their genomic plasticity.

## Introduction

Natural products are synthesized by biosynthetical gene clusters (BGCs) codified on the genome of a wide range of microorganisms. Enzymes that belong to a BGC can either be mainly restricted to secondary metabolism, or be a recent recruitment acting as accesory enzymes.  
With the genomic era and 500,000 prokaryotic genomes available at NCBI, there has been a oom of development of specilized genome mining software. Traditional approaches are based on recognize marks of enzymes devoted to secondary metabolism (**???**), or domains (**???**) lattely Evolution (**???** nadine).  
On prokaryotic genomes enzyme families are expanded frequently either by duplication or by horizontal gene transfer and that this expansions are acting as evolutionary raw material being recruited into secondary metabolism to perform nobel chemical functionalities. A proof of concept of EvoMining idea was provided by the discovery of an arseno compound on Streptomyces coelicolor (Cruz-Morales et al. 2016), nevertheless.

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, on this work we free EvoMining as a downloadable stand alone tool implemented on a docker container. EvoMining is free and open to all users and there is no login requirement. Despite Actinobacteria are great natural product producers (**???**) other microrganisms can be explored.

Here we present the EvoMining expansions analysis using different genome-DB such as Actinobacteria, Cyanobacteria, Pseudomonas and Archaea. To enrich possibilities of central DB an example of what we called backward EvoMining was incorporated: BGCs from S coelicolor available at Mi-BIG were analyzed EvoMining backwards and all enzyme families expanded but not over represented were followed.

Finally to prioritize which clusters possess more metabolite variations, assuming a link between genomic and metabolite plasticity we introduce the idea of classifying the saturation of a pangenome as open/closed pangenome measuring BGCs as open / closed BGC.

## Results and Discussion

### Section 1 EvoMining 2.0 Updated model and pipeline

EvoMining is a visual genome mining tool that following evolutionary principles has the milestone of prioritize non standard secondary metabolite pathways. The algorithm follows enzyme families from central pathways on their recruitment as components of natural products biosynthetic gene clusters (BGCs) within a genomic database.

EvoMining inputs are a (1) a custom genomic database (genomic-DB), (2) a central pathways database (central-DB) and (3) a natural product database (natural-DB) composed of genes that belongs to experimentally tested BGCs. These three databases are provided and can be modified, replaced and expanded by the user. In this work genomic-DB are collection of up to date genomes in RAST format from taxonomically related organisms such as Actinobacteria, Cyanobacteria, Pseudomonas and Archaea. Selection of this taxa obeys to the possibility of comparing well known NPs producing organisms such as Actinobacteria and Cyanobacteria in contrast with Archaea that has been poorly investigated. The central-DB contains nine central pathways from Actinobacteria previously curated (Barona-Gómez, Cruz-Morales, and Noda-García 2012), plus an update of seed metabolic enzymes identified after manual curation congruent with the central EvoMining paradigm. The natural-DB currently comprises all sequences that belongs to some BGCs from The Minimum Information about a Biosynthetic Gene cluster (MIBiG) (Medema et al. 2015).

As output EvoMining identifies on the genomic-DB those expanded families from the central-DB that has at least a recruited member onto the natural-DB, proceeding then to the reconstruction of the evolutionary history of the enzyme family. Given an enzyme from the central-DB, the product of EvoMining analysis is a color coded tree of the expanded enzyme family that provides information about the metabolic fate. Specifically, enzymes from central metabolism are differentiated from known Natural Products enzymes and those expansions with potential activity into secondary metabolism are emphasised as putative novel recruitments. Further analysis of these hits allows visualization of the genomic vicinity guiding to the discovery of novel BGCs. In addition to the updates associated to the workflow of EvoMining, the version to be released will include the possibility of defining the dynamics of the gene content of any given BGC to explore the chemical plasticity related to EvoMining hits. This allows to prioritize which clusters possess more metabolite variations, therefore unmasking biosynthetic darkmatter (Medema and Fischbach 2015, Blin et al. 2017).

EvoMining code and components (blast, muscle, FastTree, newick utilities, Gblocks,apache and SVG perl module) are wrapped on the docker container nselem/newevomining downloadable at the Docker hub. Code is available at at github: nselem/EvoMining and manual at <https://github.com/nselem/EvoMining/wiki>. EvoMining tool will allow researchers to examine their own genomes and their own enzyme families in the search of expansions involved on nobel secondary metabolism.

EvoMining will identify those expanded families of the central-DB within the genomic-DB that has at least a recruited member onto the natural-DB, proceeding then to the reconstruction of the evolutionary history of the enzyme family. Given an enzyme from the central-DB, the product of EvoMining analysis is an interactive color coded tree of the enzyme expanded family where best bidirectional hits (BBH) of central-DB are differentiated from Natural Products members and those expansions close to a Natural Product sequence that are not BBH with central-DB enzymes are emphasised as putative nobel recruitments into secondary metabolism.

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 |  |

### Section 2 Update and insights from genomic databases

EvoMining discovery of arsenopoliketydes was set on a genomic database of Actinobacteria, a phylum known by being source of natural products biosynthetic gene clusters. Nevertheless, EvoMining genomic model i.e. that secondary metabolism BGCs often includes enzymes from enzyme families that have been expanded and recruited to perform new metabolic functions is not exclusive of Actinobacteria. As a consecuence EvoMining can be applied to other bacterial taxonomical groups beside Actinobacteria but potential genomic databases should at least posses expansions of enzyme families from central metabolism.

As candidates to new genomic databases were selected genus Pseudomonas, phylum Cyanobacteria and kingdom Archaea. Pseudomonas and Cyanobacteria were chosen because of their known presence of biosynthetic gene clusters, there are respectively 60 and 58 BGCs reported on MiBIG. Archaea is the opposite case, there is not archaeal BGC reported on MiBIG. Archaea resembled 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}. Archaeas do have introns, but they are mainly located on genes that encodes ribosomal and transfer RNA (**???**). 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, and as we will show some Archaea has central pathway expansions. This lead us to think that evoMining is suitable to analize Archaeal genomes, even more if evolutionary logic behave on Archaea as on bacteria, since EvoMining is an evolutionary oriented method, not entirely based on previous knowledge of BGC’s sequences, 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. Candidate genomic databases were explore to determine wether or not other taxa posses expansions of enzymatic families devoted to central metabolism. Some families in the core genome of a taxonomical group may not exists on another, because of that, before exploring enzyme family expansions besides Actinobacteria on new genomic databases, it is needed a previous identification of central families themselves on the other taxa. A common set of central enzyme families was determined for *Actinobacteria*, *Cyanobacteria*, *Pseudomonas* and *Archaeas* by finding best bidirectional hits (BBH) of original 339 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)Pseudomonas generus* Pseudomonas fluorescens pf0-1*,*Pseudomonas protegens Pf5*,*Pseudomonas syringae pv. actinidiae ICMP 18884 NZ\_CM0027 51.1*,*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. As an outcome of this process a central database with 42 enzyme families common to all taxonomical groups was assembled. Candidate genomic databases were integrated by 876 Archaea genomes, 416 Cyanobacteria genomes, and 219 Pseudomonas genomes. Actinobacteria database was expanded from 230 to 1244 Actinobacteria genomes. Finally, with the central and genomic databases as described, taxas *Actinobacteria*, *Cyanobacteria*, *Pseudomonas* and *Archaeas* were explored in search of central enzyme families expansions. (Figure 2A)

As it could be expected, central families expansions increases whith genome size, nevertheless at different velocities and not linearly, the growth rate is different on each genomic group. Archaea, Cyanobacteria, Actinobacteria and Pseudomonas expansions behave similar until a genome size of 5,000,000 bp, after that genome size Pseudomonas number of elements on the 42 expanded central families grows faster than Actinobacteria, which in turn surpasses Cyanobacteria and Archaea, but Archaea does not have until now many genomes bigger than 5000000 bp. *Streptomycetales* and *Nostocales* orders from Actinobacteria and Cyanobacteria respectively are among the orders with bigger genome size and also bigger Number of expansions. Some of the biggest Cyanobacteria genomes may be a contamination (Hassalya bisoydea).

Now that a common set of central enzymes are determined, and that it is shown that all genomic database posses expansions of central families, even at different rates, differences of the families across selected taxa can be explored.

#### 2b Diversity on central Families according to the lineage

Expansions and recruitments of the 42 common central families were quantified by EvoMining on *Actinobacteria*, *Cyanobacteria*, *Pseudomonas* and *Archaeas* (Figure 2B). While there are families like acetylornithine aminotransferase and acetolactate synthase expanded in all groups, specially on Pseudomonas Wether a family is expanded or not depends on the taxa selected as genomic-DB

This results showed that expansion of a central family is clade specific, a family can be expanded on Actinobacteria but not on Cyanobacteria, even more a family can be part of the core of Actinobacteria but not exist on Archaea. Expansions were find to be similar in these 4 taxonomical groups,

On cyanobacteria PriA an enzyme with no previously knwon secondary metabolism on Actinobacteria has been recruited into saxitoxin cluster.  
Para aumentar la base de datos tomamos retroevomining y obtuvimos 2 en comun.. falta considerar las super duplicadas… Finalmente tomamos una enzima del shell en lugar de el core. (Puede ser del shell de streptomyces)

### Section 3 Update and insights from central DB

core/ shell genome Expansions other central  
To limitate the search for enzymes of recent recruitment into natural products

As we have explored other genomic databases, central database customization is also implemented, on this case we have use TauD en enzyme from taurine assimilation system on enterobacteria, on the genomic database Pseudomonas. Enzymes does not necessary have to be central, in the sense that they belong to the core of its taxonomic group, but at least should have expansions.

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

On EvoMining 2.0 natural products database was improved from 200 manually annotated BGCs to the incorporation of MiBIG database, curated by the whole natural products community. This improvment provides a collection of genes annotated as part of a BGCs and consequently allows another strategy to find enzymes from the shell genome that are being recruited into natural products synthesis. With MiBIG incorporation EvoMining can be run backwards, that means using MiBIG enzymes as central DB enzymes, this backwards process guarantees that enzymes are recruited, and leaves pending that they belong to the shell genome, that they are not just devoted to secondary metanolism. To this purpose all reported BGCs from *S. coelicolor* and *S. lividans* were retrieved from MiBiG and as a result 235 enzymes were taken as central-DB to the retro EvoMining process. (Table 2)

This central database was tested on part of Actinobacteria genomic database, specifically on the XXX Streptomyces genomes. Backward EvoMining was also tun on the 106 enzyme families from central metabolism. Enzymes from central metabolism does not have mode bigger than XX differing from NRPS and PKs (Figura 4) To have an idea of the expansion number on the genomes it was stablished the expansion number by family. Tested family shikimate is ~.6, we divide enzymes on reduced, intermediated or overexpanded. Overexpaded enzymes are interesting because they appear on many reported BGCs, but they are also overrepsented on genomes, so t is posible that EvoMining method of a plylogenetic resconstruction is not the best for them. Intermediate enzymes may show the EvoMining ideal of some copies devoted to standar metabolism and another copies of an enzyme family being recruited into natural products

where Orgs stands for organisms with at least one copy and copies is the total number of copies on that family on the genomic-DB

Esto se hará solo sobre Streptomyces de los 1246

HERE TABLE WITH CLUSTERS

BBH gives us 2 enzymes trees are showed on supplementary material. 55 reduced  
56 intermediate  
124 Super

but maybe some of them are over recruited, and then analizing the family expansions comparing this families with enzymes from central pathways.

Backward Enzyme sequences from S. coelicolor and S. lividans biosynthetic gene clusters were retrieved from MiBiG to test how will behave as seeds from backward EvoMining  
EvoMining was run over enzymes with expansion number between .1 and .6

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 |

## Methodology

EvoMining algorithm and all its dependencies: blast, muscle, Gblocks, Quicktree and newick utilities were packaged on a docker container available at dockerhub. Every EvoMining container has the code inside which turns EvoMining in a totally open standalone genome mining tool, Dockerfile and code are available at github. EvoMining inputs are a genomic database (genomic-DB), a seed family enzymes database from central or shell metabolism (central-DB), and a set of sequences experimentally tested to be part of natural products biosynthetical gene clusters (np-DB), optionally a database containing antiSMASH analysis of the genomic DB may be integrated to the pipeline. To standardize protein annotation and take advantage of functional annotation genomes on genomic-DB were processed by the Rapid Annotation using Subsystem Technology (RAST) platform. Sequences of the central-DB and np-DB must be on fasta format. MIBiG repository was used and is included on EvoMining docker image as default np-DB, metadata such as producer organism and kind of compound are also integrated to the np-DB.

Once databases are set analysis of expansions-recruitment of enzyme families from central-DB within the context of the genomic and natural product databases is performed by EvoMining. The first step, initialized by the user trough the web interface, is to retrieve from the genomic-DB all those sequences that belong to the expanded families enzymes of the central-DB. To this end a blastp search (default:e-value 0.001, score 100) using as queries seed enzymes from the central-DB is run against the genome-DB. Next it is established which sequences are the most conserved respect to the central-DB, to this end best bidirectional hits between the expanded families and central-DB are identified and stored in a temporal database (conserved-db). Next a heat plot is displayed pinpointing by family which organisms posses an expansion on its copy number, on this work an expansion is understood as an organisms that posses a number of sequences on certain family above the average plus an 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.

At this point families that exhibits expansions and recruitments may be selected by the user to posterior phylogenetical analysis. Selected expanded-recruited enzyme families are then aligned with muscle, automatically curated by Gblocks, and phylogenetically reconstructed by Quicktree. 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.

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 reproducibilty 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. TauD, an enzyme that belongs to the shell genome on Pseudomonas was selected as seed of central-DB. Pseudomonas genome-DB was integrated as follows (KARINA como integraste la base de Pseudomonas, how where nps predicted). Genomes on genome-DB were RAST annotated and on this case no antiSMASH mining was conducted. EvoMining expansion-recruitment and phylogenetical analysis with default parameters was conducted for TauD enzyme family.

Backward Enzyme sequences from S. coelicolor and S. lividans biosynthetic gene clusters were retrieved from MiBiG to test how will behave as seeds from backward EvoMining  
Average copy number  
resulting 235 Enzymes from BGC clusters were tested 15 clusters S coelicolor and 3 S lividans

All databases are downloadable at Zenodo (Doi numbers genomic-DB: STREPTOMYCES, Actinobacteria, Cyanobacteria, Pseudomonas, Archaea, central-DB: Actinobacteria, Cyanobacteria, Pseudomonas, Archaea, lividans/coelicolor BGC ) #### Discussion During the decade between 1970 and 1980, Archaea was recognized as new life domain, a kingdom different from Bacteria and Eucarya 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 Eucarya and Archaeal celular walls are not composed from peptidoglycans as in Bacteria. Archaeal proteins may be higlhy 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

## Figures

## Acknowledgments

Secretaria de Inovacion, Conabio:Keri/Ernesto, Argonne cluster Cocina Tepeyac

[@dufresne\_algorithmique\_2016,@blin\_recent\_nodate,@kurtboke\_revisiting\_2017,@miller\_interpreting\_2017,@schniete\_expanding\_2017,@kim\_recent\_2017,@robertsen\_toward\_2017,@juarez-vazquez\_evolution\_nodate,@chavali\_bioinformatics\_nodate,@tracanna\_mining\_2017,@ren\_breaking\_2017,@choudhary\_current\_2017,@alanjary\_antibiotic\_2017,@chevrette\_sandpuma:\_2017,@wohlleben\_antibiotic\_2016,@weber\_secondary\_2016]

# References

Alanjary, Mohammad, Brent Kronmiller, Martina Adamek, Kai Blin, Tilmann Weber, Daniel Huson, Benjamin Philmus, and Nadine Ziemert. 2017. “The Antibiotic Resistant Target Seeker (ARTS), an Exploration Engine for Antibiotic Cluster Prioritization and Novel Drug Target Discovery.” *Nucleic Acids Research* 45 (W1): W42–W48. doi:[10.1093/nar/gkx360](https://doi.org/10.1093/nar/gkx360).

Barona-Gómez, Francisco, Pablo Cruz-Morales, and Lianet Noda-García. 2012. “What Can Genome-Scale Metabolic Network Reconstructions Do for Prokaryotic Systematics?” *Antonie van Leeuwenhoek* 101 (1): 35–43. doi:[10.1007/s10482-011-9655-1](https://doi.org/10.1007/s10482-011-9655-1).

Blin, Kai, Hyun Uk Kim, Marnix H. Medema, and Tilmann Weber. 2017. “Recent Development of antiSMASH and Other Computational Approaches to Mine Secondary Metabolite Biosynthetic Gene Clusters.” *Briefings in Bioinformatics*. Accessed January 16. doi:[10.1093/bib/bbx146](https://doi.org/10.1093/bib/bbx146).

Chavali, Arvind K., and Seung Y. Rhee. 2018. “Bioinformatics Tools for the Identification of Gene Clusters That Biosynthesize Specialized Metabolites.” *Briefings in Bioinformatics*. Accessed January 16. doi:[10.1093/bib/bbx020](https://doi.org/10.1093/bib/bbx020).

Chevrette, Marc G., Fabian Aicheler, Oliver Kohlbacher, Cameron R. Currie, and Marnix H. Medema. 2017. “SANDPUMA: Ensemble Predictions of Nonribosomal Peptide Chemistry Reveal Biosynthetic Diversity Across Actinobacteria.” *Bioinformatics* 33 (20): 3202–10. doi:[10.1093/bioinformatics/btx400](https://doi.org/10.1093/bioinformatics/btx400).

Choudhary, Alka, Lynn M. Naughton, Itxaso Montánchez, Alan D. W. Dobson, and Dilip K. Rai. 2017. “Current Status and Future Prospects of Marine Natural Products (MNPs) as Antimicrobials.” *Marine Drugs* 15 (9): 272. doi:[10.3390/md15090272](https://doi.org/10.3390/md15090272).

Cibrián-Jaramillo, Angélica, and Francisco Barona-Gómez. 2016. “Increasing Metagenomic Resolution of Microbiome Interactions Through Functional Phylogenomics and Bacterial Sub-Communities.” *Frontiers in Genetics* 7. doi:[10.3389/fgene.2016.00004](https://doi.org/10.3389/fgene.2016.00004).

Cruz-Morales, Pablo, Johannes Florian Kopp, Christian Martínez-Guerrero, Luis Alfonso Yáñez-Guerra, Nelly Selem-Mojica, Hilda Ramos-Aboites, Jörg Feldmann, and Francisco Barona-Gómez. 2016. “Phylogenomic Analysis of Natural Products Biosynthetic Gene Clusters Allows Discovery of Arseno-Organic Metabolites in Model Streptomycetes.” *Genome Biology and Evolution* 8 (6): 1906–16. doi:[10.1093/gbe/evw125](https://doi.org/10.1093/gbe/evw125).

Dufresne, Yoann. 2016. “Algorithmique Pour L’annotation Automatique de Peptides Non Ribosomiques.” PhD thesis, Lille1. <https://tel.archives-ouvertes.fr/tel-01563992/document>.

Juárez-Vázquez, Ana Lilia, Janaka N Edirisinghe, Ernesto A Verduzco-Castro, Karolina Michalska, Chenggang Wu, Lianet Noda-García, Gyorgy Babnigg, et al. 2017. “Evolution of Substrate Specificity in a Retained Enzyme Driven by Gene Loss.” *eLife* 6. Accessed January 16. doi:[10.7554/eLife.22679](https://doi.org/10.7554/eLife.22679).

Kim, Hyun Uk, Kai Blin, Sang Yup Lee, and Tilmann Weber. 2017. “Recent Development of Computational Resources for New Antibiotics Discovery.” *Current Opinion in Microbiology* 39 (October): 113–20. doi:[10.1016/j.mib.2017.10.027](https://doi.org/10.1016/j.mib.2017.10.027).

Kurtböke, İpek. 2017. “Revisiting Biodiscovery from Microbial Sources in the Light of Molecular Advances.” *Microbiology Australia* 38 (2): 58–61. doi:[10.1071/MA17028](https://doi.org/10.1071/MA17028).

Medema, Marnix H., and Michael A. Fischbach. 2015. “Computational Approaches to Natural Product Discovery.” *Nature Chemical Biology* 11 (9): 639–48. doi:[10.1038/nchembio.1884](https://doi.org/10.1038/nchembio.1884).

Medema, Marnix H., Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B. Biggins, Kai Blin, Irene de Bruijn, et al. 2015. “Minimum Information About a Biosynthetic Gene Cluster.” *Nature Chemical Biology* 11 (9): 625–31. doi:[10.1038/nchembio.1890](https://doi.org/10.1038/nchembio.1890).

Miller, Ian J., Marc G. Chevrette, and Jason C. Kwan. 2017. “Interpreting Microbial Biosynthesis in the Genomic Age: Biological and Practical Considerations.” *Marine Drugs* 15 (6): 165. doi:[10.3390/md15060165](https://doi.org/10.3390/md15060165).

Ren, Hengqian, Bin Wang, and Huimin Zhao. 2017. “Breaking the Silence: New Strategies for Discovering Novel Natural Products.” *Current Opinion in Biotechnology*, Chemical biotechnology • Pharmaceutical biotechnology, 48 (December): 21–27. doi:[10.1016/j.copbio.2017.02.008](https://doi.org/10.1016/j.copbio.2017.02.008).

Robertsen, Helene Lunde, Tilmann Weber, Hyun Uk Kim, and Sang Yup Lee. 2017. “Toward Systems Metabolic Engineering of Streptomycetes for Secondary Metabolites Production.” *Biotechnology Journal* 13 (1): n/a–n/a. doi:[10.1002/biot.201700465](https://doi.org/10.1002/biot.201700465).

Schniete, Jana K., Pablo Cruz-Morales, Nelly Selem, Lorena T. Fernandez-Martinez, Iain S. Hunter, Francisco Barona-Gomez, and Paul Hoskisson. 2017. “Expanding Gene Families Helps Generate the Metabolic Robustness Required for Antibiotic Biosynthesis.” *bioRxiv*, March, 119354. doi:[10.1101/119354](https://doi.org/10.1101/119354).

Tracanna, Vittorio, Anne de Jong, Marnix H. Medema, and Oscar P. Kuipers. 2017. “Mining Prokaryotes for Antimicrobial Compounds: From Diversity to Function.” *FEMS Microbiology Reviews* 41 (3): 417–29. doi:[10.1093/femsre/fux014](https://doi.org/10.1093/femsre/fux014).

Weber, Tilmann, and Hyun Uk Kim. 2016. “The Secondary Metabolite Bioinformatics Portal: Computational Tools to Facilitate Synthetic Biology of Secondary Metabolite Production.” *Synthetic and Systems Biotechnology*, Special Issue on “Bioinformatic tools and approaches for Synthetic Biology of natural products”, 1 (2): 69–79. doi:[10.1016/j.synbio.2015.12.002](https://doi.org/10.1016/j.synbio.2015.12.002).

Wohlleben, Wolfgang, Yvonne Mast, Evi Stegmann, and Nadine Ziemert. 2016. “Antibiotic Drug Discovery.” *Microbial Biotechnology* 9 (5): 541–48. doi:[10.1111/1751-7915.12388](https://doi.org/10.1111/1751-7915.12388).

Ziemert, Nadine, Mohammad Alanjary, and Tilmann Weber. 2016. “The Evolution of Genome Mining in Microbes – a Review.” *Natural Product Reports* 33 (8): 988–1005. doi:[10.1039/C6NP00025H](https://doi.org/10.1039/C6NP00025H).