EvoMining

Nelly Selem

December31, 2017

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

Selem-Mojica Nelly, Cruz-Morales Pablo, Martínez-Guerrero Christian , ..., and Barona-Gómez Francisco

## 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 Archea studying expansions for enzyme families such as TauD and other enzymes recently recruited onto secondary metabolism. Finally the genomic plasticity of Streptomyces coelicolor known BGCs i explored generlizind applying the open/Close pangenome approach to a BGCs. 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

### Figure 1 EvoMining pipe-line

EvoMining is a visual, evolutionary based genome mining tool with 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.

Pipeline

Pipeline

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.

### Figure 2 Expansions on some databases

Archaea Cyanobacteria, and Actinobacteria based on central metabolism from actinobacteria  
2.1 Expansions same central  
Expansions other central  
To acotate the search for enzymes of recent recruitment into natural products TauD

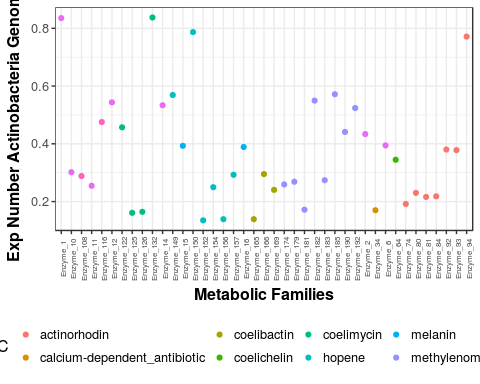
GenomicDatabases ### Figure 3.1 Expansions on genomic dinamics  
3.2(Bakward EvoMining)  
Coelicolor clusters  
Esto se hará solo sobre Streptomyces de los 1246

table <- read.csv("Figuras/CoelicolorMiBIG", row.names = 1,sep="\t")  
kable(table, caption = "Coelicolor\\label{tab:Coelicolor MiBig}",caption.short = "CoelicolorMiBig ")

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 |

# Expansions of enzime sequences from MiBIG from S coelicolor will be explored within the scope of the genomic database Streptomyces. The goal is to recover those enzymes that are not yet been considered as common on secondary metabolism.   
## Moda is the most common copy number on an organism, Organisms with an extra copy are the ones that may have this copy recrutied into secondary metabolism  
# This extra copy on at least 4 organismos  
## In addtiion the distribution of the enzyme is deseried present on at least half o the organisms (Not to exclusive)  
## Too exclusive means only belong to secondary metabolism, we are looking for switches   
## looking for an esay number between 0 and one that reflects too expanded, too exclusive  
# y el exp number, mas entre .2<= Exp <=.6 y analizar eso árboles.   
# One minus average organisms that contains one copy.  
# More copies than organisms this number tends to one ## too expanded  
# few copies on homogeneously on few organisms tends to cero ## too particular   
# two copies by organism .5 , that is not usually the case because there is some variance   
  
  
#### Functions   
OneOrMode <- function(x){ #max between 0 and one  
a = table(x) # x is a vector  
moda=a[which.max(a)]  
inte=max(1,as.integer(names(moda)))  
return(inte)  
}  
  
Mode <- function(x){ ## mode  
a = table(x) # x is a vector  
moda=a[which.max(a)]  
inte=as.integer(names(moda))  
return(inte)  
}  
  
OrganismsExtraCopy <- function(x){ ##how many organisms has an extra copy than the mode  
 a = table(x) # x is a vector   
 moda=a[which.max(a)]  
 inte=as.integer(names(moda)) #the moda  
 subx<-as.integer(a[which(as.integer(names(a))>inte)]) ## vector of organisms with extra copies  
 suma<-sum(subx) ## how many  
return(suma)  
}  
  
OrgAtLeastOneCopy <- function(x){ ##how many organisms has an extra copy than the mode  
 a = table(x) # x is a vector   
 subx<-as.integer(a[which(as.integer(names(a))>0)]) ## vector of organisms with extra copies  
 suma<-sum(subx) ## how many  
return(suma)  
}  
  
  
Copies <- function(x){ ##how many organisms has an extra copy than the mode  
 suma<-sum(x) ## how many  
return(suma)  
}  
  
######## Reading and sorting data   
## Read EvoMining tables  
tableExp <- read.csv("Figuras/ExpansionBlast.data", header=TRUE, sep="\t")   
tableDistribution <- read.csv("Figuras/Enzymes.Distribution", header=TRUE, sep="\t")   
  
#necesito poner el valor de names de moda en el renglon con el mismo valor en enzima  
#number of organisms greater than mode at least tenpercent of the genome  
##Reducing tableExp to those I have distribution  
tableExp=tableExp[tableExp$Enzyme %in% names(tableDistribution),]  
tableExp <- tableExp[order(tableExp$Enzyme),]   
tableDistribution <- tableDistribution[,order(names(tableDistribution))]   
  
  
################# Extra copy present at least in three organism  
  
  
modaOrOne=apply(tableDistribution,2, OneOrMode)  
#modaOrOne  
  
moda=apply(tableDistribution,2, Mode)  
#moda  
  
ExtraCopy=apply(tableDistribution,2, OrganismsExtraCopy)  
#ExtraCopy  
  
OneCopy=apply(tableDistribution,2, OrgAtLeastOneCopy)  
#OneCopy  
  
CopiesEvo=apply(tableDistribution,2, Copies)  
#CopiesEvo  
  
#names(moda)  
tableExp$Moda=moda  
tableExp$ExtraCopy=ExtraCopy  
tableExp$OneCopy=OneCopy  
tableExp$CopiesEvo=CopiesEvo  
  
  
#One minus average organisms that contains one copy.  
#More copies than organisms this number tends to one ## too expanded  
# few copies on homogeneously on few organisms tends to cero ## too particular   
# two copies by organism .5 , that is not usually the case because there is some variance   
tableExp$ExpNum=(modaOrOne-tableExp$Organisms/tableExp$Copies)/(modaOrOne)   
  
tableExp2 <- tableExp[order(tableExp$BGC),]   
  
ggplot(tableExp2,aes(x=tableExp2$Enzyme, y=tableExp2$ExpNum, color=tableExp2$BGC))+ geom\_point() + labs(x = "Metabolic Families", y = "Exp Number Actinobacteria Genomes",text = element\_text(size=12)) + theme\_bw()+theme(plot.title = element\_text(size = 14, face = "bold"), text = element\_text(size = 12), axis.title = element\_text(face="bold"), axis.text.x=element\_text(angle = 90,size = 6), legend.position = "bottom")



#kable(tableExp, caption = "CoelicolorExpansions\\label{tab:Coelicolor Expansions}",caption.short = "CoelicolorExpansions")

Presence Absence EvoMining was run over enzymes with expansion number between .1 and .6 [Backwars](Figuras/Backward.svg) ### Figure 4 Pan cluster Idea on closed Streptomyces

Cluster visualization

Cluster visualization

Open /closed coelicolor How spread is the cluster How to describe the cluster  
Conservation  
Enzymes that appear x%

Variability How variable is the region  
derivative of rarefaction curve

Took 15 clusters from Streptomyces coelicolor on MiBig Analize its open/close pancluster according to EvoMining backwards  
O sea 15 corasones, no necesito escoger las query enzyme, al menos 3 por cluster... y que no sean NRPS o PKS

## MEthodology

[@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).