EvoMining

Nelly Selem

December31, 2017

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). The assumption behind EvoMining is that 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 EvoMining software was not released, on this work we free EvoMining as a downloadable stand alone tool implemented on a docker container.  
EvoMining inputs are a custom genomic database (genomic-DB), a central pathways database (central-DB) and a natural product database (natural-DB) composed of genes that belongs to experimentally tested BGCs. The genomic-DB is a collection genomes in RAST format from taxonomically reated organisms. 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). Finally the also customizable central-DB used on this studio contains nine central pathways of Actinobacteria including amino acid biosynthesis, glycolysis, pentose phosphate pathway, and tricarboxylic acids cycle (Barona-Gómez, Cruz-Morales, and Noda-García 2012).

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 a 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 emphazised as putative nobel recruitments into secondary metabolism.

Here we present the EvoMining expansions analysis using different genome-DB such as Actinobacteria, Cyanobacteria, Pseudomonas and Archaea. Finally in order to complement our central database we incorporate an example of what we called *backward EvoMining*. *S coelicolor* BGCs available at MiBIG were analized *EvoMining backwards* and all enzymes not over represented were followed on an EvoMining analysis. Following the idea of measure the saturation of a peangenome also a cluster can be classified as open or closed.

EvoMining is wraped on a docker container downloadable at <https://hub.docker.com/r/nselem/newevomining/> with the code available at <https://github.com/nselem/EvoMining> and manual at <https://github.com/nselem/EvoMining/wiki> .

## Manual pages

github

table <- read.csv("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 |

Figure 1 EvoMining pipe-line  
Figure 2 Expansions on some databases  
Figure 3 Expansions on genomic dinamics (Bakward EvoMining)  
Coelicolor clusters Figure 4 Pan cluster Idea on closed Streptomyces  
Open /closed coelicolor Took 15 custers from Streptomyces coelicolor on MiBig Analize its open/close pancluster according to EvoMining backwards

# References

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).

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).

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).