Genome analysis

Palantir: a springboard for the analysis of secondary metabolite gene clusters in large-scale genome mining projects

Loïc Meunier[1,2*], Pierre Tocquin[3,4], Luc Cornet[5], Damien Sirjacobs[1], Valérie Leclère[6], Maude Pupin[7,8], Philippe Jacques[2], and Denis Baurain[1,4*]

[1]InBioS-PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, B-4000 Liège, Belgium, [2]TERRA Teaching and Research Centre, Microbial Processes and Interactions, Gembloux Agro-Bio Tech, University of Liège, B-5030 Gembloux, Belgium, [3]InBioS-PhytoSYSTEMS, Plant Physiology, University of Liège, B-4000 Liège, Belgium, [4]Hedera-22 SCRL, B-4130 Tilff, Belgium, [5]GIGA institute, Medical Genomics-Unit of Animal Genomics, University of Liège, B-4000 Liège, Belgium, [6]Univ. Lille, INRA, ISA, Univ. Artois, Univ. Littoral Côte d'Opale, EA 7394-ICV-Institut Charles Viollette, F-59000 Lille, France, [7]UMR 9189- CRIStAL- Centre de Recherche en Informatique Signal et Automatique de Lille, University of Lille, CNRS, Centrale Lille, F-59000 Lille, France, [8]Bonsai Team, Inria-Lille Nord Europe, F-59655 Villeneuve d'Ascq Cedex, France

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: To support small and large-scale genome mining projects, we present Palantir (Post-processing Analysis tooLbox for ANTIsmash Reports), a dedicated software suite for handling and refining secondary metabolite biosynthetic gene cluster (BGC) data annotated with the popular antiSMASH pipeline. Palantir provides new functionalities building on NRPS/PKS predictions from antiSMASH, such as improved BGC annotation, module delineation and easy access to sub-sequences at different levels (cluster, gene, module, domain). Moreover, it can parse user-provided antiSMASH reports and reformat them for direct use or storage in a relational database.

Availability: Palantir is released both as a Perl API available on CPAN (https://metacpan.org/release/Bio-Palantir) and as a web application (http://palantir.uliege.be). As a practical use case, the web interface also features a database built from the mining of 1616 cyanobacterial genomes, of which 1488 were predicted to encode at least one BGC.

Contact: denis.baurain@uliege.be, lmeunier@uliege.be.

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Natural products, synthesized by diverse microorganisms and plants, are a precious source of bioactive compounds for different industrial fields, especially the pharmaceutical industry (Harvey, 2008; Pham et al., 2019). Affordable and fast, genome mining has become a method of choice to support the discovery of such molecules (Niu 2018; Trivella and Felicio, 2018). Hence, numerous bioactive compounds have been unveiled over the last decade thanks to the active development of dedicated tools (Chen et al., 2013; Kersten et al., 2013; Kačar et al., 2019). An extensive list of currently available tools for genome mining is provided on The Secondary Metabolite Bioinformatics Portal (Weber et al., 2016) and many of these have been described in several reviews (Ziemert et al., 2016; Chavali et al., 2017). Here, we present Palantir (Post-processing Analysis tooLbox for ANTIsmash Reports), a tool to facilitate large-scale genome mining analyses of hundreds or thousands of genomes with antiSMASH, one of the most comprehensive and up-to-date software packages dedicated to secondary metabolism pathways (Blin et al., 2017a; Blin et al., 2019). Besides, we devised additional functionalities (complementary to antiSMASH methods) to improve the annotation of NonRibosomal Peptide Synthases (NRPSs) and PolyKetide Synthases (PKSs), two major classes of secondary metabolism pathways. These two enzymatic systems, encoded as large biosynthetic gene clusters (BGCs), are similarly based on a multimodular enzymatic mechanism, where the role of each module is to add, through the action of independent catalytic sites (*i.e.,* domains), a monomer (amino acids or beta-keto functional groups for NRPS and PKS, respectively) to the final product. In particular, Palantir features unique visualization tools for the comparison and selection among competing annotations for the architecture of the same BGC.

2 Input data

Palantir accepts report files from antiSMASH versions 3.x and 4.x (biosynML.xml) (Medema et al., 2015; Weber et al., 2015; Blin et al., 2017b), and from the newer version 5.x (regions.js). The regions.js file can be extracted from the results downloaded through the current antiSMASH web server (https://antismash.secondarymetabolites.org) or obtained locally with the standalone version of antiSMASH.

3 Functionalities

3.1 Automated parsing of antiSMASH reports

The annotation of BGCs is characterized by multidimensional and rich data structures, which can be complex to manage. Regarding NRPS and PKS, not only their genetic pathway annotation can be approached at different levels (*i.e.*, whole gene cluster, gene, module, domain or even motif), but also each level can be further annotated (*e.g.*, substrate selection by adenylation and acyl-transferase domains, stereochemistry produced by condensation, ketosynthase or tailoring PKS domains). Moreover, if screening a few genomes to assess their biosynthetic potential is trivial, considering the availability of GUI software, large-scale projects involving the analysis of many genomes require automation of the parsing, storage and querying of such a large output.

In regard to this unusual type of biological data, Palantir API offers functionalities (Table 1) to handle both small and large-scale projects: (1) FASTA sequence extraction at any BGC level, (2) customizable PDF/Word reporting and (3) relational (SQL) database generation for more advanced data analysis. Since functionalities 2 and 3 mostly find their use when applied to hundreds or thousands of antiSMASH reports, they are only accessible through the Perl distribution.

3.2 Refinement of NRPS and PKS BGC annotation

The annotation of NRPS and PKS BGCs is commonly performed by the identification of biosynthetic core genes containing domain protein signatures, also referred to as pHMMs (profile Hidden Markov Models), which are probabilistic models capturing the versatile information contained at different positions of a multiple sequence alignment (*e.g.*, amino acid composition, indels) (Eddy, 1998; Eddy, 2011). Despite the generally good accuracy of BGC screening achieved by antiSMASH and similar software packages, exceptions in detection rules still occur, due to the impossibility of establishing a universal threshold to distinguish between true and spurious pHMM matches. Missed domains (gaps) are especially problematic for modular enzymes, which are composed of multiple subunits, often leaving BGC architectures incomplete.

An additional issue is that protein signatures used by antiSMASH only partially cover the sequences of experimentally characterized domains (*e.g.*, condensation domain signature is 300 amino acid long, while characterized domains have a length ranging between 450 and 500 amino acids). Domain sequences returned by antiSMASH thus lack information that would be useful in downstream analyses, such as phylogenetic inference (see Supplementary information).

Palantir API (and web application) brings three methods to tackle these issues: (Table1, feature 4) module delineation, (5) gap-filling for completing BGC annotation, and (6) dynamic elongation of their core domain sequences. Furthermore, (7) a visualization functionality allows the user to easily check the refinements applied to the BGC domain architecture and compare these with the antiSMASH version. Finally, (8) an "exploratory mode" devised to interpret the architecture from scratch, *i.e.*, without any bias from predefined BGC and gene construction rules, is also provided. This additional mode displays all protein signature matches, allowing the user to design her/his preferred BGC annotation based on her/his expertise and interpretation.

The usefulness of these methods was assessed by re-analyzing data from a case study: the published dataset of cyanobacterial NRPS BGCs from Calteau et al. (2014) (Supplementary information). A first test showed that Palantir dynamic elongation of core domain sequences significantly improves the phylogenetic signal, and thus the resolution of evolutionary trees (*i.e.*, statistical support and biological plausibility of BGC phylogenetic analyses), whereas a second test demonstrated that Palantir exploratory mode led to the identification of a previously missed putative alpha/beta hydrolase, a domain type known to be implied in Claisen cyclase domains in aflatoxin biosynthesis (Korman et al., 2010).

4 Application example: An online database summarizing the large-scale BGC mining of available cyanobacterial genomes

To illustrate and make these functionalities easier to use, we released a web interface featuring both the Palantir application and the first database dedicated to cyanobacterial secondary metabolites (http://palantir.uliege.be). This SQL database is constituted of a curated collection of 1616 cyanobacterial (and related lineages) genomes with links to both antiSMASH and Palantir annotations. Of these, 1488 genomes were predicted to encode at least one BGC. Furthermore, to support informed comparisons between cyanobacterial species or strains, we assessed for each genome its putative contamination level by running CheckM (Parks et al., 2015). The web interface allows users to browse through the database results and visualize antiSMASH and Palantir annotations for each genome. Additionally, users can import their own antiSMASH results into the web application and compare them to Palantir gap-filled and exploratory annotations (dynamic BGC visualization and CSV file download).

Table 1. Overview of Palantir functionalities

#	Functionalities	Description	Applications
1	Extraction of sequences	Export FASTA files from BGC datasets at different levels: cluster, gene, module, domain	Data formatting for downstream analyses
2	Generation of PDF/Word reports	Export customizable reports of refined BGC data	End-user consumption of numerous BGC datasets
3	Generation of SQL tables	Export SQL tables containing detailed BGC data	Large-scale or more complex queries and statistics
4	Module delineation	Apply biological rules to group domains into modules	Analyses at module level
5	Dynamic elongation of the coordinates of core domains	Enrich the information contained in the sequences	Similarity searches and evolutionary analyses
6	Filling gaps in BGC architectures	Recover missed domains due to exceptions in detection rules	Resolution of ambiguous or incoherent BGC architectures
7	BGC visualization	Visualize and compare antiSMASH and Palantir annotations	Comparison of the refinements applied to the antiSMASH annotation

8	“Exploratory mode” visualization	Visualize and design the domain architecture consensus from a raw view of all detected signatures	Manual curation of the domain architecture consensus

Description and possible applications of Palantir functionalities.

Acknowledgements

Romain Lamoot and Brahim Djarallah are acknowledged for their work on a preliminary version of the web interface. We are grateful to Catherine Colson for her help with the graphical formatting of phylogenetic trees.

Funding

This work was supported by the F.R.S-FNRS. Computational resources were provided through two grants to DB (University of Liège "Crédit de démarrage 2012" SFRD-12/04; F.R.S.- FNRS "Crédit de recherche 2014" CDR J.0080.15).

Conflict of Interest: none declared.

References

Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y., & Weber, T. (2017a). The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. Nucleic Acids Research, 45(D1), D555–D559.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, Y., … Weber, T. (2019). antiSMASH 5 . 0 : updates to the secondary metabolite genome mining pipeline. Nucleic Acids Research, 47(W1), W81–W87.

Blin, K., Wolf, T., Chevrette, M. G., Lu, X., Schwalen, C. J., Kautsar, S. A., … Medema, M. H. (2017b). antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Research, 45(W1), W36–W41.

Calteau, A., Fewer, D. P., Latifi, A., Coursin, T., Laurent, T., Jokela, J., … Gugger, M. (2014). Phylum-wide comparative genomics unravel the diversity of secondary metabolism in Cyanobacteria. BMC Genomics, 15, 1–14.

Chavali, A. K., & Rhee, S. Y. (2017). Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. Briefings in Bioinformatics, (January), 1–13.

Chen, L., Yue, Q., Zhang, X., Xiang, M., Wang, C., Li, S., … An, Z. (2013). Genomics-driven discovery of the pneumocandin biosynthetic gene cluster in the fungus Glarea lozoyensis. BMC Genomics, 14(1), 1–18.

Eddy, S. R. (2011). Accelerated profile HMM searches. PLoS Computational Biology, 7(10).

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics (Vol. 14).

Harvey, A. L. (2008). Natural products in drug discovery. Drug Discovery Today, 13(19/20), 894–901.

Kačar, D., Schleissner, C., Cañedo, L. M., Rodríguez, P., de la Calle, F., Galán, B., & García, J. L. (2019). Genome of Labrenzia sp. PHM005 Reveals a Complete and Active Trans-AT PKS Gene Cluster for the Biosynthesis of Labrenzin. Frontiers in Microbiology, 10(November), 1–14.

Kesten, R. D., Lane, A. L., Nett, M., Richter, T. K. S., Duggan, B. M., Dorrestein, P. C., & Moore, B. S. (2013). Bioactivity-guided genome mining reveals the lomaiviticin biosynthetic gene cluster in Salinispora tropica. ChemBioChem, 14(8), 955–962.

Korman, T. P., Crawford, J. M., Labonte, J. W., Newman, A. G., Wong, J., Townsend, C. A., & Tsai, S. C. (2010). Structure and function of an iterative polyketide synthase thioesterase domain catalyzing Claisen cyclization in aflatoxin biosynthesis. Proceedings of the National Academy of Sciences of the United States of America, 107(14), 6246–6251.

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., … Glöckner, F. O. (2015). The Minimum Information about a Biosynthetic Gene cluster (MIBiG) specification. Nature Chemical Biology, 11(9), 625–631.

Niu, G. (2018). Genomics-Driven Natural Product Discovery in Actinomycetes. Trends in Biotechnology, 36(3), 238–241.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Research, 25(7), 1043–1055.

Pham, J. V., Yilma, M. A., Feliz, A., Majid, M. T., Maffetone, N., Walker, J. R., … Yoon, Y. J. (2019). A review of the microbial production of bioactive natural products and biologics. Frontiers in Microbiology, 10(JUN), 1–27.

Trivella, D. B. B., & De Felicio, R. (2018). The Tripod for Bacterial Natural Product Discovery : Genome. MSystems, 3(2), 1–5.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., … Medema, M. H. (2015). AntiSMASH 3.0-A comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research, 43(W1), W237–W243.

Weber, T., & Uk, H. (2016). The secondary metabolite bioinformatics portal : Computational tools to facilitate synthetic biology of secondary metabolite production. Synthetic and Systems Biotechnology, 1(2), 69–79.

Ziemert, N., Alanjary, M., & Weber, T. (2016). Natural Product Reports The evolution of genome mining in microbes – a review. Natural Product Reports, 33, 988–1005.