Downloaded from orbit.dtu.dk on: Apr 19, 2024

DTU Library

antiSMASH 6.0: improving cluster detection and comparison capabilities

Blin, Kai; Shaw, Simon; Kloosterman, Alexander M; Charlop-Powers, Zach; van Wezel, Gilles P; Medema, Marnix H; Weber, Tilmann

Published in:
Nucleic Acids Research

Link to article, DOI:
10.1093/nar/gkab335

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Blin, K., Shaw, S., Kloosterman, A. M., Charlop-Powers, Z., van Wezel, G. P., Medema, M. H., & Weber, T. (2021). antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkab335

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Nucleic Acids Research, 2021 **1**
doi: 10.1093/nar/gkab335

antiSMASH 6.0: improving cluster detection and comparison capabilities

Kai Blin [1,*], Simon Shaw[1], Alexander M. Kloosterman[2], Zach Charlop-Powers [3], Gilles P. van Wezel [2,4], Marnix H. Medema [2,5,*] and Tilmann Weber [1,*]

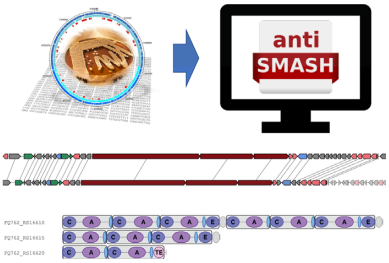
[1]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, Denmark, [2]Institute of Biology, Leiden University, Leiden, The Netherlands, [3]Bioinformatics, Lodo Therapeutics, New York, USA, [4]Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands and [5]Bioinformatics Group, Wageningen University, Wageningen, The Netherlands

Received February 22, 2021; Revised April 12, 2021; Editorial Decision April 18, 2021; Accepted April 19, 2021

ABSTRACT

Many microorganisms produce natural products that form the basis of antimicrobials, antivirals, and other drugs. Genome mining is routinely used to complement screening-based workflows to discover novel natural products. Since 2011, the "antibiotics and secondary metabolite analysis shell—antiSMASH" (https://antismash.secondarymetabolites.org/) has supported researchers in their microbial genome mining tasks, both as a free-to-use web server and as a standalone tool under an OSI-approved open-source license. It is currently the most widely used tool for detecting and characterising biosynthetic gene clusters (BGCs) in bacteria and fungi. Here, we present the updated version 6 of antiSMASH. antiSMASH 6 increases the number of supported cluster types from 58 to 71, displays the modular structure of multi-modular BGCs, adds a new BGC comparison algorithm, allows for the integration of results from other prediction tools, and more effectively detects tailoring enzymes in RiPP clusters.

GRAPHICAL ABSTRACT



INTRODUCTION

Natural compounds produced by microorganisms form the basis of many drugs (1). Traditionally, new compounds were discovered by extracting, chemically isolating, purifying, and then testing from natural sources. This approach can now be complemented by sequencing and subsequent mining of genome and metagenome data to identify natural product biosynthetic pathways (2). Software tools to assist researchers in their natural product genome mining tasks have existed for over a decade. Recently published or updated examples include BAGEL (3), PRISM (4), RiPPER (5) and TOUCAN (6). For more in-depth reviews on the topic, see (7–10).

Since its initial release in 2011, antiSMASH (11–15) has become the most widely used tool for mining microbial genomes for secondary/specialised metabolite (SM) biosynthetic gene clusters (BGCs) and is regarded as the gold standard. An ecosystem of independent tools incorporating or utilising antiSMASH results has developed over the years, such as the antibiotics resistance target seeker ARTS (16), the mass-spectronomy-guided peptide mining tool Pep2Path (17), the sgRNA design tool CRISPY-web (18), the BGC classification and clustering platform BiG-SCAPE (19), or its big data BGC clustering cousin BiG-SLiCE (20). antiSMASH is also used to annotate BGCs in many genomic as well as BGC-oriented databases, such as the Joint Genome Institute's Integrated Microbial Genomes database with its Atlas of Biosynthetic gene Clusters IMG-ABC (21), the MicroScope platform for microbial genome annotation and analysis (22), the MIBiG database of manually curated BGCs (23), the BGC family database BiG-FAM (24), and, of course, the antiSMASH database (25).

antiSMASH uses a rule-based approach to identify many different types of biosynthetic pathways involved in SM production. More in-depth analyses are performed for BGCs encoding non-ribosomal peptide synthetases

*To whom correspondence should be addressed. Tel: +45 24 89 61 32; Email: tiwe@biosustain.dtu.dk
Correspondence may also be addressed to Marnix. H. Medema. Email: marnix.medema@wur.nl
Correspondence may also be addressed to Kai Blin. Email: kblin@biossutain.dtu.dk

© The Author(s) 2021. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab335/6274535 by guest on 13 May 2021

(NRPSs), type I and type II polyketide synthases (PKSs), lanthipeptides, lasso peptides, sactipeptides, and thiopeptides, for which cluster-specific analyses can provide more information about the biosynthetic steps performed and thus also provide more detailed predictions on the compound(s) produced (Figure 1).

Here we present version 6 of antiSMASH. It extends and improves upon previous versions by adding and improving BGC detection rules, making the modular structure of multimodular enzymes more accessible, introducing an additional, more robust, cluster comparison tool, and improving the interoperability with other gene cluster annotation tools.

NEW FEATURES AND UPDATES

New cluster types

antiSMASH uses manually curated and validated "rules" that define which core biosynthetic functions need to exist in a genomic region in order to constitute a BGC. To identify these biosynthetic functions, antiSMASH uses profile hidden Markov models (pHMMs) from PFAM (26), TIGRFAMs (27), SMART (28), BAGEL (3), Yadav *et al.* (29), or custom models. antiSMASH 5 contained rules for 58 different BGC types (15), version 6 increases this number to 71. Especially in focus for this release were the rules for ribosomally synthesized and post-translationally modified peptides (RiPPs). The lanthipeptide rule was split up into individual rules for classes I through IV. Several RiPP families that were previously jointly designated as "bacteriocins" are now identified by specific rules. The term "bacteriocins" was therefore deprecated and replaced by "RiPP-like", which is defined by profiles that are frequently associated with RiPPs but are insufficient to detect RiPP clusters by themselves. The old "head_to_tail" rule was folded into the "sactipeptide" rule because they covered the same class of RiPPs. New rules were added for class V lanthipeptides, lanthidines, thioamitides, ranthipeptides, PQQ- and mycofactocin-like redox cofactors, epipeptides, cyclic lactone autoinducers, and spliceotides. Outside of RiPPs, antiSMASH added support for BGCs producing thioamide-containing non-ribosomal peptides, tropodithietic acid, *Serratia*-type prodigiosins, non-alpha poly-amino acids and pyrrolidines.

Module detection

Nonribosomal peptide synthetases, non-iterative type I polyketide synthases and trans-AT polyketide synthases are large, multimodular enzymes (for a review, see (30)). While antiSMASH has always detected the individual enzymatic domains in these megaenzymes, it now also detects and displays the modules explicitly (Figure 2A). Module detection allows for better prediction of modifications made to the respective monomers during biosynthesis. Hence, displaying the modules makes it easier for researchers to interpret the likely biosynthetic mechanisms of an enzymatic assembly line encoded by a BGC. In order to examine the protein domains making up each module, hovering the mouse cursor over a module will send the module lid to the background and will reveal the domains. For a more complete view of all protein domains, module monomer display can be turned off using the "show module domains" toggle button above the graph to hide all module lids (Figure 2B).

ClusterCompare

Since the release of antiSMASH 1, antiSMASH has provided a comparison of the identified region to similar clusters in other genomes via ClusterBlast (see (11) for a description of the algorithm). As the ClusterBlast algorithm is based on protein sequence comparisons by local alignments (initially using BLAST (31), now using DIAMOND (32)), it does not perform optimally on multimodular enzymes like NRPSes and PKSes. Of particular note, BGCs with very different module counts can score similarly if a single module has a good match. To address this issue, we have added a new ClusterCompare algorithm (Shaw *et al.*, manuscript in preparation) to antiSMASH 6. Like ClusterBlast, ClusterCompare builds on a local protein-alignment-based sequence comparison, but only uses that as one part of the comparison score. Additional parts of the score are the gene synteny and the presence/absence of biosynthetic components of the query and reference gene clusters, based on antiSMASH annotations of each. Biosynthetic components are the collection of gene products matching one of antiSMASH's BGC detection profiles, gene products with a functional annotation due to either their presence in an antiSMASH detection rule or based on the classification of their secondary metabolite clusters of orthologous groups (smCOG) class, and, if applicable, NRPS and PKS domains. All score parts are scaled to a 0 to 1 range, and the final score for a single comparison is calculated as the geometric mean of the parts. The top hits in the reference database are displayed in a table, along with the scoring information. Selecting a table row will also display a pairwise comparison of the query and reference clusters (Figure 3).

Sideloading

antiSMASH exists in an ecosystem of BGC prediction and analysis tools. While tools like ARTS (16) and BiG-SCAPE (19) sit downstream of antiSMASH and consume antiSMASH results, other tools like ClusterFinder (33) and DeepBGC (34) offer alternative, machine-learning based methods to predict gene clusters and thus function more in parallel to antiSMASH's rule based cluster detection. Depending on the research question and the desired sensitivity vs. specificity trade-off, there can be a value in preferring one cluster detection over the other (see (8,35) for a discussion). It is not feasible from a development resource perspective to include and maintain all different cluster detection approaches within the antiSMASH pipeline. Hence, to use analysis modules such as ClusterBlast, ClusterCompare, their respective MIBiG-based variants, or PFAM/TIGRFAMs analysis on the identified BGCs regardless of their mode of detection, we have created a JSON-based file format that can be used by external tools to annotate additional regions and clusters in the input and allow antiSMASH to analyse those areas exactly as per the natively predicted regions and clusters. An explanation of the external annotation file format and the currently

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab335/6274535 by guest on 13 May 2021

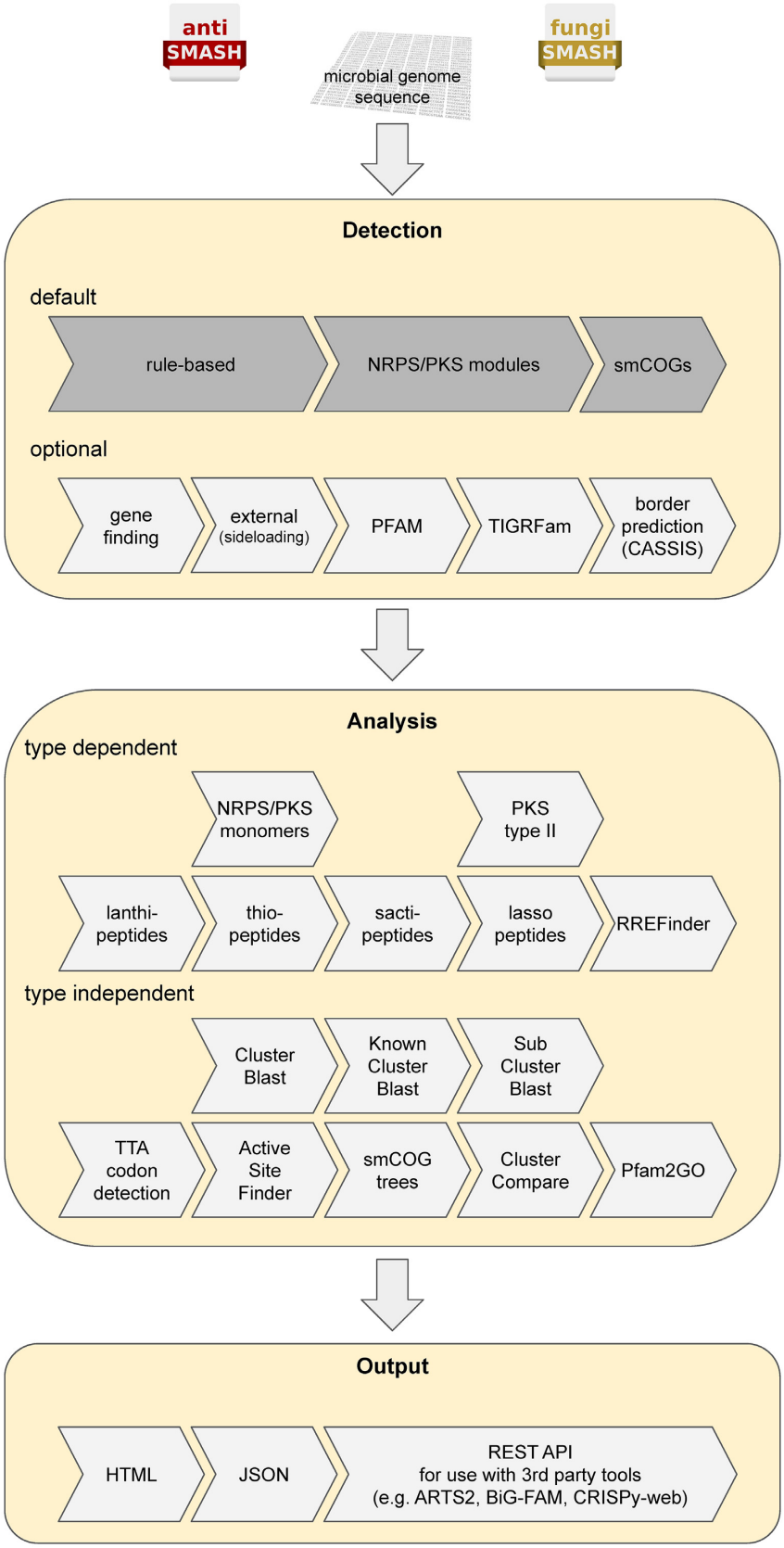


Figure 1. Schematic workflow of the antiSMASH secondary/specialized metabolite genome mining platform.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab335/6274535 by guest on 13 May 2021

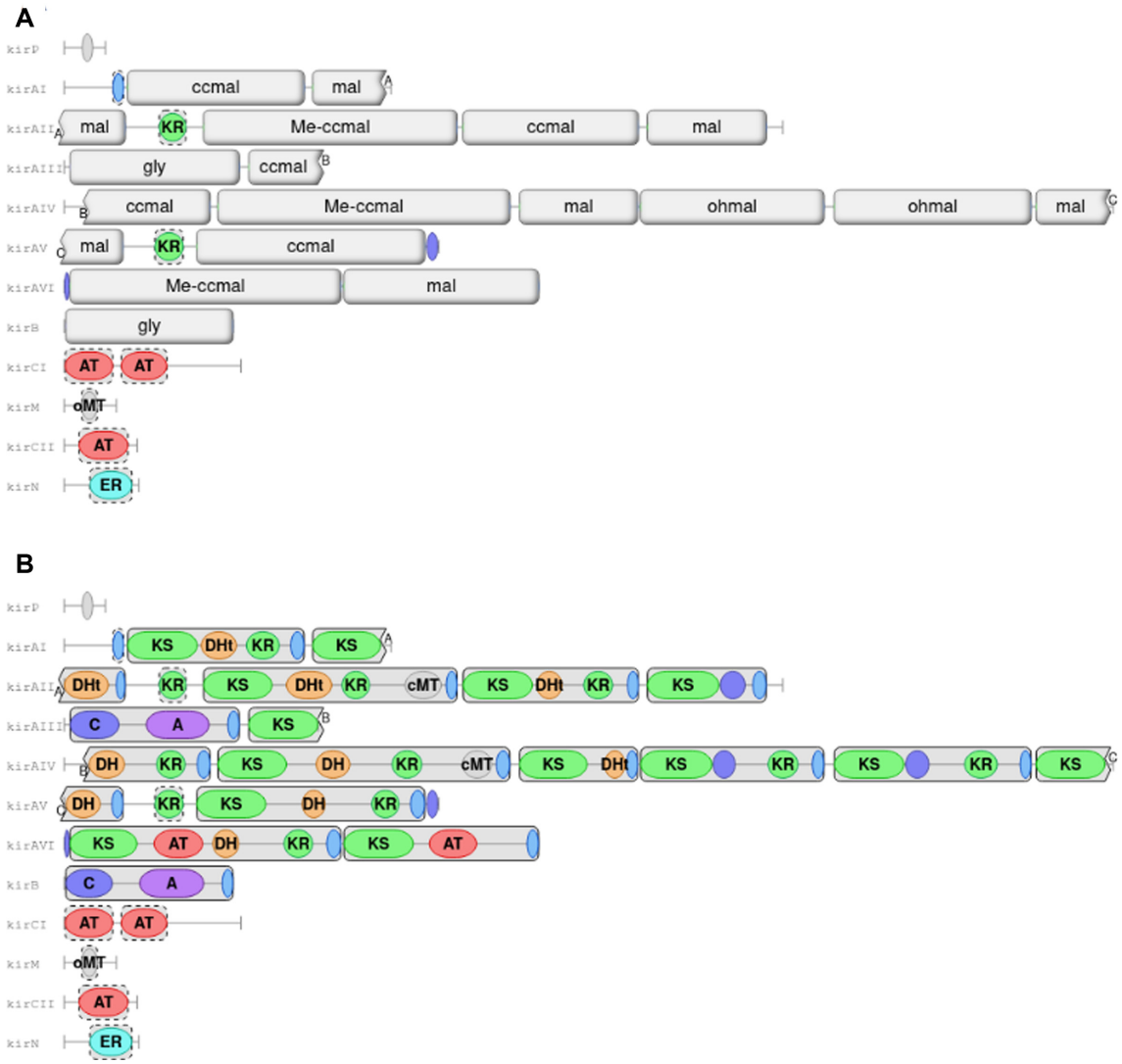


Figure 2. The NRPS/PKS domain view of the kirromycin biosynthetic gene cluster (NCBI ID: AM746336.1), consisting of trans-AT PKS, modular type I PKS and NRPS modules. **A** View with module lids, displaying the monomer predicted to be integrated into the final product. The jagged module edges on KirAI/AII/AIII/AIV show modules that are split across different protein-coding sequences, with the small lettering next to the edges indicating how the modules link up. **B** View with the module lids hidden, revealing the underlying protein domains.

implemented sideloading functionality is available at https://docs.antismash.secondarymetabolites.org/sideloading/.

Improved annotation for RiPP clusters

For lanthipeptide, lasso peptide, sactipeptide, and thiopeptide BGCs, antiSMASH 5 already provided more detailed product predictions by detecting the cluster's prepeptide and commonly occurring tailoring enzymes. Because some of the tailoring enzymes can match relatively generic functional profiles, it was not always trivial to determine whether a given enzyme was indeed interacting with the RiPP precursor peptide as a tailoring enzyme or whether it was just an unrelated enzyme that happened to be encoded in the vicinity. Often, RiPP tailoring enzymes will harbour RiPP recognition element (RRE) domains that can identify and bind the RiPP precursor peptide. RRE-Finder (36), which will annotate these RRE domains, has been integrated into antiSMASH 6, thus helping to more confidently identify tailoring enzymes in antiSMASH-detected RiPP clusters. Additionally, an "RRE-containing" detection rule using the RRE-Finder pHMMs was added to the "relaxed" strictness ruleset, allowing antiSMASH to identify potentially novel RiPP clusters for which antiSMASH does not have a specific detection rule set up, as long as that gene cluster contains an RRE.

Other optimizations

The sequence data used in antiSMASH's ClusterBlast are based on the records contained in the antiSMASH database. As the antiSMASH database was recently updated to version 3 (25), the ClusterBlast dataset was also refreshed to include 147 517 high quality BGC regions from

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab335/6274535 by guest on 13 May 2021

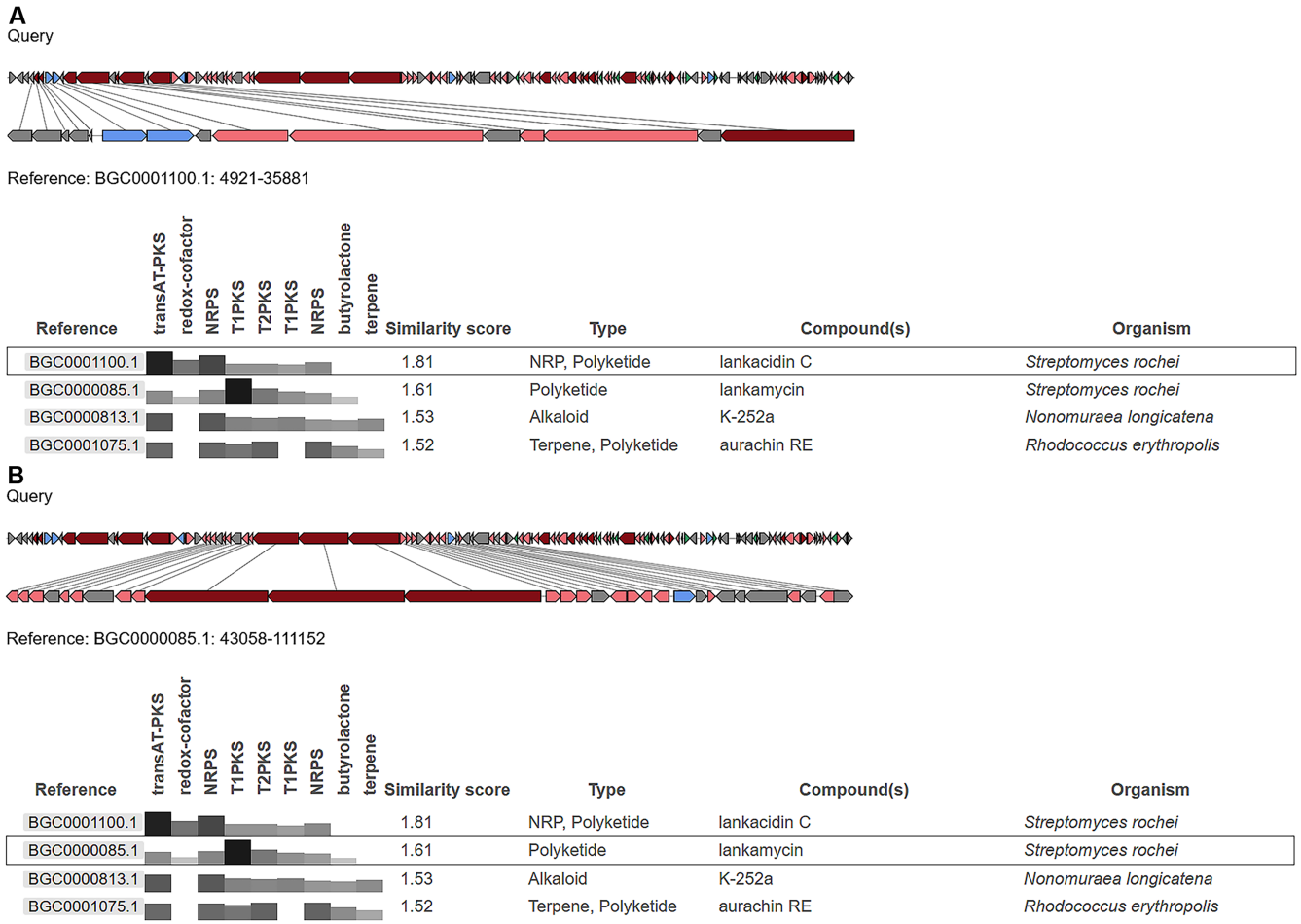


Figure 3. ClusterCompare output for the *Streptomyces rochei* large linear plasmid pSLA2-L (NCBI ID: NC_004808.2), which is densely packed with secondary metabolite biosynthetic genes (see (39)) with the MIBiG dataset in protocluster-to-region mode. Lines connect pairs of protein-coding genes with the highest similarity to make conserved functions more visible even at different scaling of query and reference. The comparisons show the similarity of (**A**) the left part of the region to the trans-AT PKS, NRPS hybrid cluster responsible for lankacidin C production (MIBiG ID: BGC0001100.1), as well as (**B**) the middle part of the region to the modular PKS type I cluster responsible for lankamycin biosynthesis (MIBiG ID: BGC0000085.1). The example illustrates how ClusterCompare can be used to distinguish between hybrid gene clusters and adjacent gene clusters that are part of the same region, based on comparison with individual reference BGCs.

388 archaeal, 25 236 bacterial and 177 fungal genomes. The antiSMASH database version 3 is the first version to also contain both archaeal and fungal sequences along with bacterial sequences, so ClusterBlast will now also give more relevant hits for users running antiSMASH on inputs originating from those taxa.

In addition to the region PFAM analysis added in version 5, antiSMASH 6 can now also scan regions using profiles from the TIGRFAMS database (27).

CONCLUSIONS AND FUTURE PERSPECTIVES

Genome mining with tools like antiSMASH has become an established part of many natural product discovery workflows. With the updates and additions to the feature set, antiSMASH is positioning itself to remain the go-to tool for microbial genome mining for natural products. By improving the interoperability with other tools, the open-source software antiSMASH integrates even better with the thriving ecosystem of computational tools in the natural products field. Future updates will further improve the predictions of the chemical structures of the compounds produced by the detected BGCs. This will help to connect gene clusters to molecules identified via metabolomics or other analytical chemistry approaches (37), and to link up with databases such as GNPS (38).

DATA AVAILABILITY

The bacteria and fungal versions of antiSMASH 6 can be freely accessed at https://antismash.secondarymetabolites.org and https://fungismash.secondarymetabolites.org, respectively.

The antiSMASH documentation is available at https://docs.antismash.secondarymetabolites.org/.

The antiSMASH source code, licensed under the GNU Affero General Public License (AGPL) v3.0, is available

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab335/6274535 by guest on 13 May 2021

at https://github.com/antismash/antismash. antiSMASH is also available via Docker.

FUNDING

Novo Nordisk Foundation [NNF20CC0035580 to T.W., NNF16OC0021746 to T.W.]; Center for Microbial Secondary Metabolites (CeMiSt), Danish National Research Foundation [DNRF137 to T.W.]; ERC Starting Grant [948770-DECIPHER to M.H.M.]; Netherlands Organization for Scientific Research (NWO) [731.014.206 to G.P.v.W.]. Funding for open access charge: The Novo Nordisk Foundation.

Conflict of interest statement. M.H.M. is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio. Z.C.-P. is employed by Lodo Therapeutics, New York, USA.

REFERENCES

1. Newman,D.J. and Cragg,G.M. (2020) Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *J. Nat. Prod.*, **83**, 770–803.
2. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.*, **33**, 988–1005.
3. van Heel,A.J., de Jong,A., Song,C., Viel,J.H., Kok,J. and Kuipers,O.P. (2018) BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. *Nucleic Acids Res.*, **46**, W278–W281.
4. Skinnider,M.A., Johnston,C.W., Gunabalasingam,M., Merwin,N.J., Kieliszek,A.M., MacLellan,R.J., Li,H., Ranieri,M.R.M., Webster,A.L.H., Cao,M.P.T. *et al.* (2020) Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.*, **11**, 6058.
5. Santos-Aberturas,J., Chandra,G., Frattaruolo,L., Lacret,R., Pham,T.H., Vior,N.M., Eyles,T.H. and Truman,A.W. (2019) Uncovering the unexplored diversity of thioamidated ribosomal peptides in Actinobacteria using the RiPPER genome mining tool. *Nucleic Acids Res.*, **47**, 4624–4637.
6. Almeida,H., Palys,S., Tsang,A. and Diallo,A.B. (2020) TOUCAN: a framework for fungal biosynthetic gene cluster discovery. *NAR Genom Bioinform*, **2**, lqaa098.
7. Weber,T. (2014) In silico tools for the analysis of antibiotic biosynthetic pathways. *Int. J. Med. Microbiol.*, **304**, 230–235.
8. Medema,M.H. and Fischbach,M.A. (2015) Computational approaches to natural product discovery. *Nat. Chem. Biol.*, **11**, 639–648.
9. Weber,T. and Kim,H.U. (2016) The secondary metabolite bioinformatics portal: computational tools to facilitate synthetic biology of secondary metabolite production. *Synth Syst Biotechnol*, **1**, 69–79.
10. Blin,K., Kim,H.U., Medema,M.H. and Weber,T. (2019) Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. *Brief. Bioinform.*, **20**, 1103–1113.
11. Medema,M.H., Blin,K., Cimermancic,P., de Jager,V., Zakrzewski,P., Fischbach,M.A., Weber,T., Takano,E. and Breitling,R. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.*, **39**, W339–W346.
12. Blin,K., Medema,M.H., Kazempour,D., Fischbach,M.A., Breitling,R., Takano,E. and Weber,T. (2013) antiSMASH 2.0–a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.*, **41**, W204–W12.
13. Weber,T., Blin,K., Duddela,S., Krug,D., Kim,H.U., Bruccoleri,R., Lee,S.Y., Fischbach,M.A., Müller,R., Wohlleben,W. *et al.* (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Res.*, **43**, W237–W2443.
14. Blin,K., Wolf,T., Chevrette,M.G., Lu,X., Schwalen,C.J., Kautsar,S.A., Suarez Duran,H.G., de Los Santos,E.L.C., Kim,H.U., Nave,M. *et al.* (2017) antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.*, **45**, W36–W41.
15. Blin,K., Shaw,S., Steinke,K., Villebro,R., Ziemert,N., Lee,S.Y., Medema,M.H. and Weber,T. (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.*, **47**, W81–W87.
16. Mungan,M.D., Alanjary,M., Blin,K., Weber,T., Medema,M.H. and Ziemert,N. (2020) ARTS 2.0: feature updates and expansion of the Antibiotic Resistant Target Seeker for comparative genome mining. *Nucleic Acids Res.*, **48**, W546–W552.
17. Medema,M.H., Paalvast,Y., Nguyen,D.D., Melnik,A., Dorrestein,P.C., Takano,E. and Breitling,R. (2014) Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. *PLoS Comput. Biol.*, **10**, e1003822.
18. Blin,K., Shaw,S., Tong,Y. and Weber,T. (2020) Designing sgRNAs for CRISPR-BEST base editing applications with CRISPy-web 2.0. *Synth Syst Biotechnol*, **5**, 99–102.
19. Navarro-Muñoz,J.C., Selem-Mojica,N., Mullowney,M.W., Kautsar,S.A., Tryon,J.H., Parkinson,E.I., De Los Santos,E.L.C., Yeong,M., Cruz-Morales,P., Abubucker,S. *et al.* (2020) A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.*, **16**, 60–68.
20. Kautsar,S.A., van der Hooft,J.J.J., de Ridder,D. and Medema,M.H. (2021) BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. *Gigascience*, **10**, doi:10.1093/gigascience/giaa154.
21. Palaniappan,K., Chen,I.-M.A., Chu,K., Ratner,A., Seshadri,R., Kyrpides,N.C., Ivanova,N.N. and Mouncey,N.J. (2020) IMG-ABC v.5.0: an update to the IMG/Atlas of Biosynthetic Gene Clusters Knowledgebase. *Nucleic Acids Res.*, **48**, D422–D430.
22. Vallenet,D., Calteau,A., Dubois,M., Amours,P., Bazin,A., Beuvin,M., Burlot,L., Bussell,X., Fouteau,S., Gautreau,G. *et al.* (2020) MicroScope: an integrated platform for the annotation and exploration of microbial gene functions through genomic, pangenomic and metabolic comparative analysis. *Nucleic Acids Res.*, **48**, D579–D589.
23. Kautsar,S.A., Blin,K., Shaw,S., Navarro-Muñoz,J.C., Terlouw,B.R., van der Hooft,J.J.J., van Santen,J.A., Tracanna,V., Suarez Duran,H.G., Pascal Andreu,V. *et al.* (2020) MIBiG 2.0: a repository for biosynthetic gene clusters of known function. *Nucleic Acids Res.*, **48**, D454–D458.
24. Kautsar,S.A., Blin,K., Shaw,S., Weber,T. and Medema,M.H. (2021) BiG-FAM: the biosynthetic gene cluster families database. *Nucleic Acids Res.*, **49**, D490–D497.
25. Blin,K., Shaw,S., Kautsar,S.A., Medema,M.H. and Weber,T. (2021) The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. *Nucleic Acids Res.*, **49**, D639–D643.
26. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
27. Haft,D.H., Selengut,J.D., Richter,R.A., Harkins,D., Basu,M.K. and Beck,E. (2013) TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.*, **41**, D387–D395.
28. Letunic,I., Khedkar,S. and Bork,P. (2021) SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.*, **49**, D458–D460.
29. Yadav,G., Gokhale,R.S. and Mohanty,D. (2009) Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS Comput. Biol.*, **5**, e1000351.
30. Weissman,K.J. (2015) The structural biology of biosynthetic megaenzymes. *Nat. Chem. Biol.*, **11**, 660–670.
31. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
32. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
33. Cimermancic,P., Medema,M.H., Claesen,J., Kurita,K., Wieland Brown,L.C., Mavrommatis,K., Pati,A., Godfrey,P.A., Koehrsen,M., Clardy,J. *et al.* (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, **158**, 412–421.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab335/6274535 by guest on 13 May 2021

34. Hannigan,G.D., Prihoda,D., Palicka,A., Soukup,J., Klempir,O., Rampula,L., Durcak,J., Wurst,M., Kotowski,J., Chang,D. *et al.* (2019) A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.*, **47**, e110.
35. Baltz,R.H. (2019) Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. *J. Ind. Microbiol. Biotechnol.*, **46**, 281–299.
36. Kloosterman,A.M., Shelton,K.E., van Wezel,G.P., Medema,M.H. and Mitchell,D.A. (2020) RRE-Finder: a genome-mining tool for class-independent RiPP discovery. *mSystems*, **5**, e00267-20.
37. van der Hooft,J.J.J., Mohimani,H., Bauermeister,A., Dorrestein,P.C., Duncan,K.R. and Medema,M.H. (2020) Linking genomics and metabolomics to chart specialized metabolic diversity. *Chem. Soc. Rev.*, **49**, 3297–3314.
38. Wang,M., Carver,J.J., Phelan,V.V., Sanchez,L.M., Garg,N., Peng,Y., Nguyen,D.D., Watrous,J., Kapono,C.A., Luzzatto-Knaan,T. *et al.* (2016) Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.*, **34**, 828–837.
39. Mochizuki,S., Hiratsu,K., Suwa,M., Ishii,T., Sugino,F., Yamada,K. and Kinashi,H. (2003) The large linear plasmid pSLA2-L of *Streptomyces rochei* has an unusually condensed gene organization for secondary metabolism. *Mol. Microbiol.*, **48**, 1501–1510.

Downloaded from https://academic.oup.com/nar/advance-article/doi/10.1093/nar/gkab335/6274535 by guest on 13 May 2021