Nucleic Acids Research, 2023, 1–4
https://doi.org/10.1093/nar/gkad984
Database issue

The antiSMASH database version 4: additional genomes and BGCs, new sequence-based searches and more

Kai Blin [1,*], Simon Shaw [1], Marnix H. Medema [2] and Tilmann Weber [1,*]

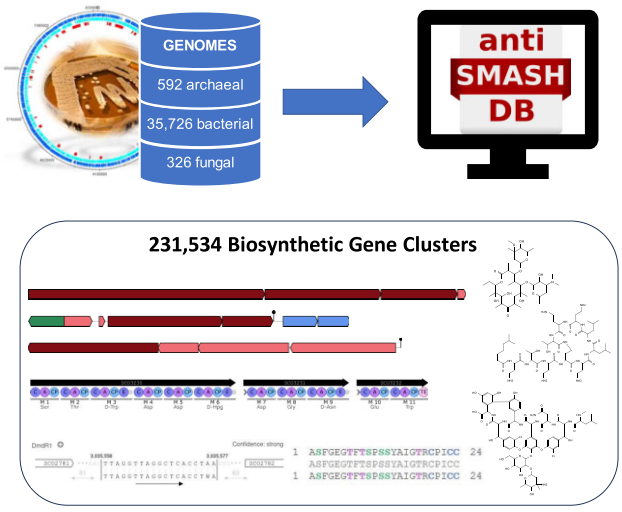
[1]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Lyngby 2800, Denmark
[2]Bioinformatics Group, Wageningen University, Wageningen, 6708PB, The Netherlands

*To whom correspondence should be addressed. Tel: +45 93511306; Email: kblin@biosustain.dtu.dk
Correspondence may also be addressed to Tilmann Weber. Tel: +45 24 89 61 32; Email: tiwe@biosustain.dtu.dk

Abstract

Many microorganisms produce natural products that are frequently used in the development of medicines and crop protection agents. Genome mining has evolved into a prominent method to access this potential. antiSMASH is the most popular tool for this task. Here we present version 4 of the antiSMASH database, providing biosynthetic gene clusters detected by antiSMASH 7.1 in publicly available, dereplicated, high-quality microbial genomes via an interactive graphical user interface. In version 4, the database contains 231 534 high quality BGC regions from 592 archaeal, 35 726 bacterial and 236 fungal genomes and is available at https://antismash-db.secondarymetabolites.org/.

Graphical abstract



Introduction

Secondary metabolites produced by microorganisms are the main source of bioactive compounds that are in use as antimicrobial and anticancer drugs (1), as well as fungicides, herbicides, pesticides and other crop protection agents (2). Classically, these compounds were discovered by making extracts out of samples from natural sources, followed by chemical isolation, purification and activity screening. The sequencing boom of the last decade has made microbial genome data readily available, making it possible to complement this traditional approach with genome mining technologies (3). Software tools for natural product genome mining have existed for over a decade (as discussed in various reviews (4–8)). Only a few databases made such data available, starting with the now-defunct ClusterMine360 (9) in 2013.

Since its initial release in 2011, antiSMASH (10–16) has become the most widely used tool for genome mining for secondary/specialised metabolites and is generally regarded as the gold standard. antiSMASH uses a rule-based approach to detect genome regions containing biosynthetic gene clusters based on conserved biosynthetic enzymes from (currently) 88 different biosynthetic pathway types. In addition to cluster-specific analyses for many of the better-understood pathways, antiSMASH also compares identified regions to the MIBiG database (17) of known BGCs, as well as a dataset of antiSMASH results predicted from publicly available genomes.

antiSMASH is a genome mining tool by design, meaning that it analyses and annotates individual microbial genomes, one at a time. To help with research questions that can better

Received: September 15, 2023. **Revised:** October 13, 2023. **Editorial Decision:** October 13, 2023. **Accepted:** October 17, 2023

be answered with cross-genome datasets, we developed the antiSMASH database (18–20). With its easy to use query builder, it allows researchers to quickly run even complex queries against BGCs identified across tens of thousands of genomes. Additionally the database is used as the basis for antiSMASH's ClusterBlast functionality, with ClusterBlast hits linking to the database. antiSMASH results are cross-referenced to similar results in the database, as well as to similar clusters from the MIBiG database.

Here we present the fourth version of this database, covering 592 archaeal, 35 726 bacterial and 236 fungal genomes.

Materials and methods

Selection of included genomes

Archaeal, bacterial and fungal genomes were downloaded from the NCBI RefSeq database on 4–5 April 2023, using the ncbi-genome-download tool (21) using the 'complete', 'chromosome' and 'scaffold' assembly levels. To avoid issues with badly fragmented assemblies negatively affecting BGC prediction quality, assemblies with >100 (for archaea and bacteria)/150 (fungi) contigs were discarded. Redundancy filtering was performed as described previously (20), with the exception that Mash (22) was now used for all taxa, including fungi. After redundancy filtering, 641 archaeal, 38 991 bacterial and 257 fungal sequences remained.

antiSMASH annotations and data import

On this filtered dataset, we used GNU parallel (23) to run antiSMASH 7 with the options '--cb-knownclusters --cb-subclusters --cc-mibig --clusterhmmer --tigrfam --pfam2go --rre --asf --tfbs'. antiSMASH successfully processed 640 archaeal, 38 940 bacterial and 255 fungal assemblies, the others were skipped due to the lack of gene annotations or other annotation-related errors. From this first run, we extracted all predicted ribosomally synthesised and posttranslationally modified peptide (RiPP) precursors and regions to build new CompaRiPPson and ClusterBlast datasets, respectively. After creating the updated datasets, antiSMASH 7 was re-run on the first round's results using the options '--cb-general --reuse', also updating CompaRiPPson results with the new dataset.

The SQL schema for the database (https://github.com/antismash/db-schema/) and importer (https://github.com/antismash/db-import/) were updated to support antiSMASH 7 results. During the import process, assemblies without any antiSMASH predictions were dropped, resulting in the final count of 592 archaeal, 35 726 bacterial and 236 fungal assemblies being represented in the database.

Results and discussion

The NCBI RefSeq database contains a wealth of microbial genomes. However, the database does contain a lot of redundancies caused by tens of thousands of sequences of common pathogens like *Escherichia coli*, *Salmonella enterica* or *Staphylococcus aureus*. Additionally, many of the genome assemblies are draft assemblies from short reads that leave the genome in hundreds or even thousands of tiny contigs, which has massive impacts on the quality of BGC detection (24). In order to have a good representation of BGCs across the whole sequenced microbial tree of life without overly biassing for the frequently sequenced species, and to ensure that clusters are as complete as possible without being spread over many contigs, the antiSMASH database applies rigorous quality filtering and sequence-similarity based filtering. After filtering and processing, the fourth version of the antiSMASH database contains 231 534 high-quality BGCs from 592 archaeal, 35 726 bacterial and 236 fungal high-quality, representative, genomes. Annotations were performed using antiSMASH 7.1, which has additional rules on top of those in the 7.0 release (16): isocyanides, NRP-related isocyanides, highly-reducing PKS type IIs, darobactins, triceptides, archaeal RiPPs and hydrogen cyanides. This results in a total of 88 different supported pathways.

Version 4 of the database makes all of the antiSMASH predictions available using its query functionality. The NRPS/PKS module search has been integrated into the regular query builder to allow for combined queries like 'find regions containing NRPS modules with *N*-methyltransferase domains in the genus *Streptomyces*' or any other filters on top of the module selection. During user testing of the simple text query, we noticed confusion about which search categories were covered, as well as frustration about the absence of type completion hints. As the current version of the database brings the number of supported search categories up to 39, we decided to remove the simple text query. This means that all database queries now run via the query builder interface.

For users who want to search the database using their own protein sequence data or known RiPP precursor sequences, two newly added search features are available. The RiPP precursor search (Figure 1A) uses NCBI blast+ blastp (25) to compare a user-provided sequence with all predicted RiPP precursors in the database (Figure 1B). Similarly, the protein sequence search (Figure 1C) uses DIAMOND (26) to compare user-provided sequences to all protein sequences from predicted BGCs (Figure 1D).

The development work needed to allow running these kinds of searches in the background has also been used to make CSV and FASTA downloads more reliable. Previously, queries returning CSV or FASTA results were generally slow and needed strict pagination to return data before the browser connection timed out. This in turn made these queries cumbersome to use, especially from automated scripts. In version 4 of the database, we instead create background processes to collect all of the requested data. Once collected, data is available from the antiSMASH database servers for a week before being cleaned up automatically. This should make it drastically easier to download larger slices of the antiSMASH database.

Additionally, the complete dataset for the whole database is available for download on our download server in various formats for users wishing to integrate any or all of the data into their own in-house tools, see the data availability statement for details. Examples on how to use the antiSMASH database search and download functions can be found on the database's 'Help' page.

Compared to version 3's 25 802 assemblies, version 4 contains 36 554, roughly a 42% increase. At the same time, the number of high-quality BGCs increased, increasing the number of BGCs that did not run into a contig edge from 147 517 to 231 534, almost 57%. This increase is likely due to the increased number of BGC types supported by the latest version of antiSMASH.

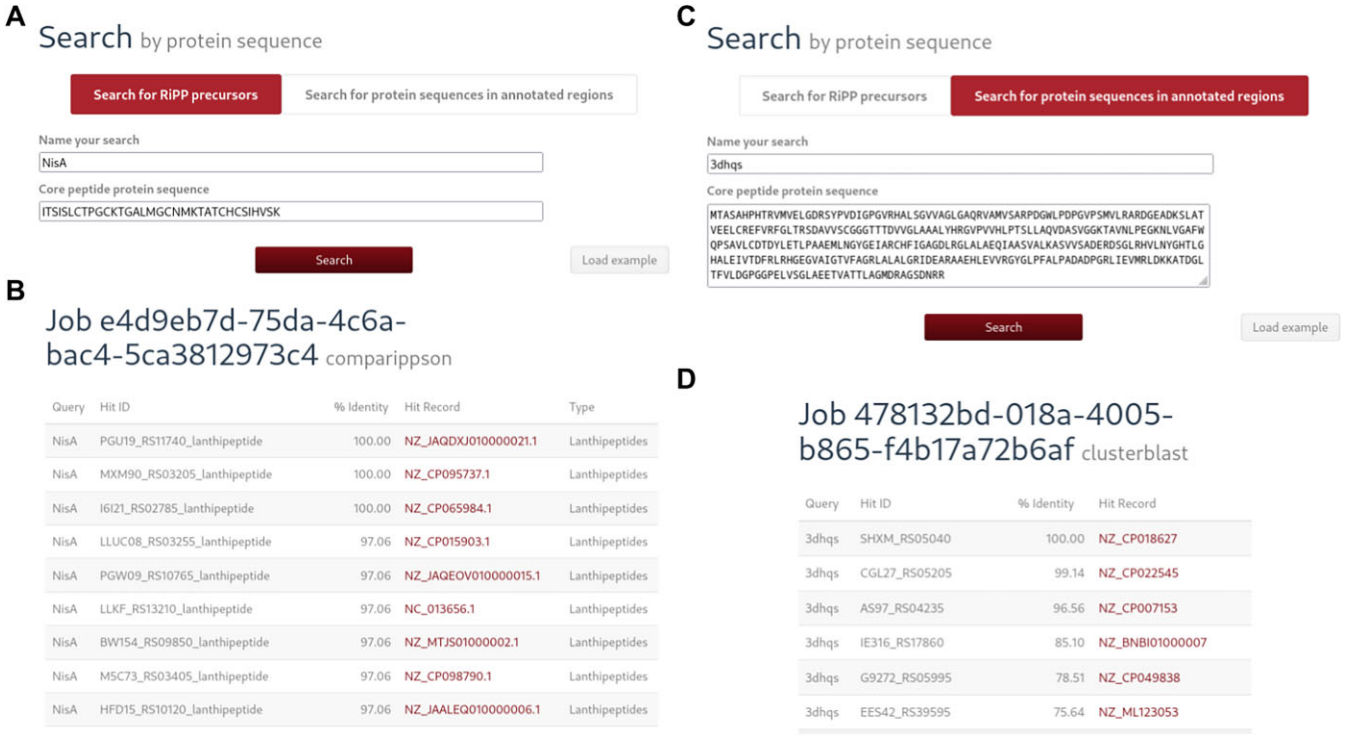


Figure 1. (**A**) The RiPP precursor search loaded with the lanthipeptide nisin A's core sequence. (**B**) The CompaRiPPson results for the above search (**C**) The protein sequence search using a 3-dehydroquinate synthase. (**D**) The protein sequence search results

Conclusions

Genome mining continues to be an invaluable technique for assessing microbial biosynthetic potential. Since 2011, antiSMASH has aided with these efforts. The antiSMASH database helps to compare identified clusters across genomes and allows for more complex searches to contextualise and cross-reference findings via a user-friendly web interface.

With a selection 231 534 BGC regions from archaea, bacteria and fungi, the antiSMASH database version 4 is a comprehensive collection of secondary/specialised metabolite biosynthetic gene clusters with up-to-date, high quality antiSMASH-based annotations available to the natural product research community.

Data availability

The antiSMASH database is available at https://antismash-db.secondarymetabolites.org/. There are no access restrictions for academic or commercial use of the web server. The source code components and SQL schema for the antiSMASH database are available on GitHub (https://github.com/antismash) under an OSI-approved Open Source license. The complete set of antiSMASH results, the antiSMASH JSON files, and an SQL dump of the database can be downloaded from the antiSMASH download server (https://dl.secondarymetabolites.org/database/4.0/).

Funding

Novo Nordisk Foundation [NNF20CC0035580 to T.W., K.B., S.S., NNF16OC0021746 to T.W.]; Danish National Research Foundation [DNRF137 to T.W.]; ERC Starting Grant [948770-DECIPHER to M.H.M.]. Funding for open access charge: 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.

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. Sparks,T.C. and Bryant,R.J. (2022) Impact of natural products on discovery of, and innovation in, crop protection compounds. *Pest Manag. Sci.*, **78**, 399–408.
3. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes – a review. *Nat. Prod. Rep.*, **33**, 988–1005.
4. Weber,T. (2014) In silico tools for the analysis of antibiotic biosynthetic pathways. *Int. J. Med. Microbiol.*, **304**, 230–235.
5. Medema,M.H. and Fischbach,M.A. (2015) Computational approaches to natural product discovery. *Nat. Chem. Biol.*, **11**, 639–648.
6. 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.
7. Baltz,R.H. (2019) Natural product drug discovery in the genomic era: realities, conjectures, misconceptions, and opportunities. *J. Ind. Microbiol. Biotechnol.*, **46**, 281–299.
8. 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.

9. Conway,K.R. and Boddy,C.N. (2013) ClusterMine360: a database of microbial PKS/NRPS biosynthesis. *Nucleic Acids Res.*, **41**, D402–D407.
10. 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.
11. 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–W212.
12. 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–W243.
13. 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.
14. 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.
15. Blin,K., Shaw,S., Kloosterman,A.M., Charlop-Powers,Z., van Wezel,G.P., Medema,M.H. and Weber,T. (2021) antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.*, **49**, W29–W35.
16. Blin,K., Shaw,S., Augustijn,H.E., Reitz,Z.L., Biermann,F., Alanjary,M., Fetter,A., Terlouw,B.R., Metcalf,W.W., Helfrich,E.J.N., *et al.* (2023) antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.*, **51**, W46–W50.
17. Terlouw,B.R., Blin,K., Navarro-Muñoz,J.C., Avalon,N.E., Chevrette,M.G., Egbert,S., Lee,S., Meijer,D., Recchia,M.J.J., Reitz,Z.L., *et al.* (2023) MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.*, **51**, D603–D610.
18. Blin,K., Medema,M.H., Kottmann,R., Lee,S.Y. and Weber,T. (2017) The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.*, **45**, D555–D559.
19. Blin,K., Pascal Andreu,V., de los Santos,E.L.C., Del Carratore,F., Lee,S.Y., Medema,M.H. and Weber,T. (2019) The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.*, **47**, D625–D630.
20. 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.
21. Blin,K. (2023) ncbi-genome-download. https://doi.org/10.5281/zenodo.8192486.
22. Ondov,B.D., Treangen,T.J., Melsted,P., Mallonee,A.B., Bergman,N.H., Koren,S. and Phillippy,A.M. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, **17**, 132.
23. Tange,O. (2022) GNU Parallel 20221122 ('Χερσόη'). https://doi.org/10.5281/zenodo.7347980.
24. Sánchez-Navarro,R., Nuhamunada,M., Mohite,O.S., Wasmund,K., Albertsen,M., Gram,L., Nielsen,P.H., Weber,T. and Singleton,C.M. (2022) Long-read metagenome-assembled genomes improve identification of novel complete biosynthetic gene clusters in a complex microbial activated sludge ecosystem. *Msystems*, **7**, e00632-22.
25. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinf.*, **10**, 421.
26. Buchfink,B., Reuter,K. and Drost,H.-G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods*, **18**, 366–368.

Received: September 15, 2023. **Revised:** October 13, 2023. **Editorial Decision:** October 13, 2023. **Accepted:** October 17, 2023