Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx319

antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification

Kai Blin[1], Thomas Wolf[2], Marc G. Chevrette[3], Xiaowen Lu[4], Christopher J. Schwalen[5], Satria A. Kautsar[4], Hernando G. Suarez Duran[4], Emmanuel L. C. de los Santos[6], Hyun Uk Kim[1,7], Mariana Nave[8], Jeroen S. Dickschat[9], Douglas A. Mitchell[5,10], Ekaterina Shelest[2], Rainer Breitling[11], Eriko Takano[11], Sang Yup Lee[1,7], Tilmann Weber[1,*] and Marnix H. Medema[4,*]

[1]Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark, [2]Leibniz Institute for Natural Product Research and Infection Biology—Hans-Knöll-Institute, 07745 Jena, Germany, [3]Laboratory of Genetics, University of Wisconsin—Madison, Madison, WI 53706, USA, [4]Bioinformatics Group, Wageningen University, 6708PB Wageningen, Netherlands, [5]Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA, [6]Warwick Integrative Synthetic Biology Centre, University of Warwick, Coventry CV4 7AL, UK, [7]Department of Chemical and Biomolecular Engineering & BioInformatics Research Center, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea, [8]Faculty of Sciences, University of Lisbon, 1749-016 Lisbon, Portugal, [9]Kekulé-Institute of Organic Chemistry and Biochemistry, University of Bonn, 53121 Bonn, Germany, [10]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA and [11]Manchester Synthetic Biology Research Centre (SYNBIOCHEM), Manchester Institute of Biotechnology, University of Manchester, Manchester M1 7DN, UK

Received February 25, 2017; Revised April 07, 2017; Editorial Decision April 12, 2017; Accepted April 13, 2017

ABSTRACT

Many antibiotics, chemotherapeutics, crop protection agents and food preservatives originate from molecules produced by bacteria, fungi or plants. In recent years, genome mining methodologies have been widely adopted to identify and characterize the biosynthetic gene clusters encoding the production of such compounds. Since 2011, the 'antibiotics and secondary metabolite analysis shell—antiSMASH' has assisted researchers in efficiently performing this, both as a web server and a standalone tool. Here, we present the thoroughly updated antiSMASH version 4, which adds several novel features, including prediction of gene cluster boundaries using the ClusterFinder method or the newly integrated CASSIS algorithm, improved substrate specificity prediction for non-ribosomal peptide synthetase adenylation domains based on the new SANDPUMA algorithm, improved predictions for terpene and ribosomally synthesized and post-translationally modified peptides cluster products, reporting of sequence similarity to proteins encoded in experimentally characterized gene clusters on a per-protein basis and a domain-level alignment tool for comparative analysis of *trans*-AT polyketide synthase assembly line architectures. Additionally, several usability features have been updated and improved. Together, these improvements make antiSMASH up-to-date with the latest developments in natural product research and will further facilitate computational genome mining for the discovery of novel bioactive molecules.

INTRODUCTION

Natural products, also referred to as secondary or specialized metabolites, are the basis of many drugs and are also important molecules for agricultural and nutritional applications; moreover, they play key roles in scientific research as chemical probes to study many aspects of molecular and cellular biology. The observation that the genomes of many microorganisms contain multiple biosynthetic gene clusters (BGCs) that code for the production of such molecules has led to a paradigm shift in natural products research: within the last 10 years, genome mining has been established as an important technology complementing the bioassay- and chemistry-driven classical natural products discovery process (1). This fundamental change was supported by the development and public availability of various genome min-

*To whom correspondence should be addressed. Tel: +31 317484706; Email: marnix.medema@wur.nl
Correspondence may also be addressed to Tilmann Weber. Tel: +45 24896132; Email: tiwe@biosustain.dtu.dk

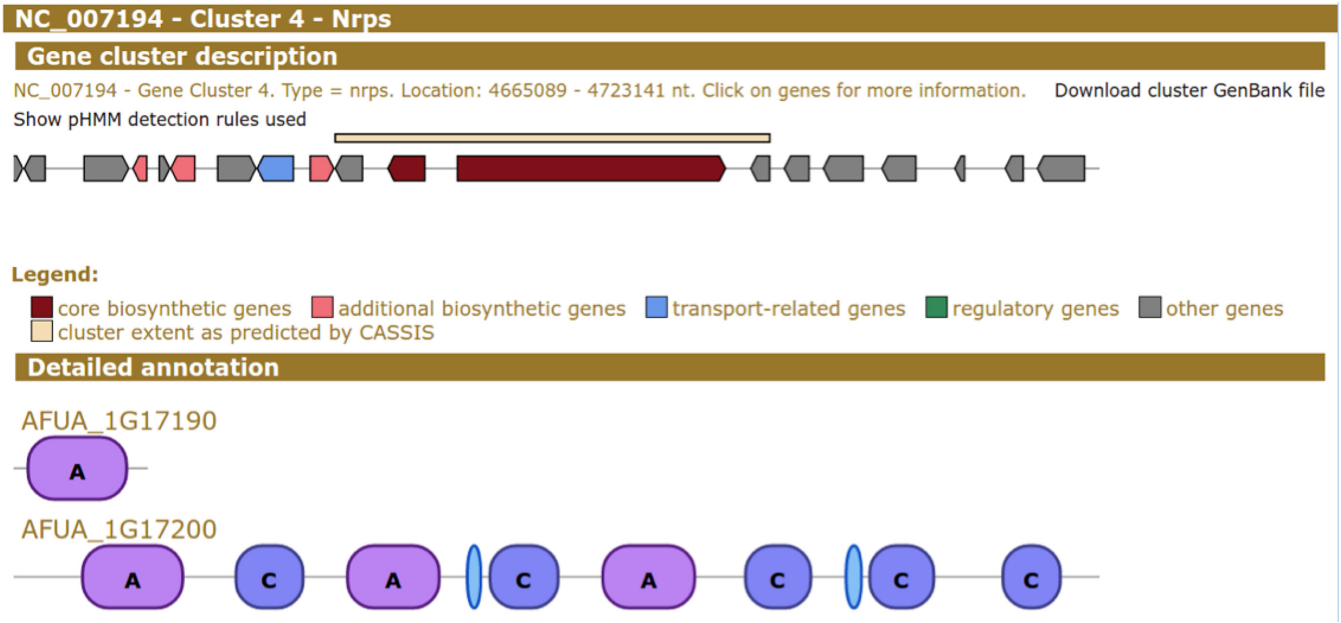

Figure 1. Gene cluster border prediction by the Cluster Assignment by Islands of Sites (CASSIS) algorithm. The fourth cluster on chromosome 1 of *Aspergillus nidulans* is shown. The cream-colored bar above the gene arrows spans the genes predicted to be clustered by CASSIS. Further genes in the surrounding are displayed for additional context. Similar functionality is available when using ClusterFinder to predict gene cluster borders.

ing software tools that are usable by wet-lab microbiologists and chemists (as reviewed in (2–4)), such as NP.searcher (5), antiSMASH (6–8), NaPDoS (9) and recently PRISM/GNP (10,11).

The comprehensive open-source BGC mining platform antiSMASH (6–8) was first released in 2011 and has been regularly updated with extended functionality. antiSMASH facilitates the mining of bacterial and fungal genomes and is tightly interconnected with plantiSMASH, a new variant for BGC mining in plants (12), the antiSMASH database (13) and the Minimum Information on Biosynthetic Gene Cluster (MIBiG) repository of experimentally characterized BGCs (14).

Here, we report version 4 of antiSMASH, which includes several major extensions, such as gene cluster boundary prediction for fungal BGCs, improved chemistry predictions for terpene, ribosomal peptide and non-ribosomal peptide BGCs, comparative alignment of *trans*-AT polyketide synthase (PKS) assembly lines and TTA codon annotation. Moreover, an improved user interface was introduced, along with several other usability and efficiency improvements. The public antiSMASH web server is freely accessible at http://antismash.secondarymetabolites.org.

NEW FEATURES AND UPDATES

Improved prediction of gene cluster boundaries. Estimating the boundaries of BGCs is a continuing challenge for genome mining tools. Traditionally, antiSMASH has opted for a 'greedy' approach by design, in order to ensure a greater likelihood of including all pertinent biosynthetic genes. The rationale behind this was that expert users would be better at estimating cluster boundaries than automated algorithms would. However, for certain purposes, it is still highly beneficial for users to review a computer-assisted estimate of where a BGC may start and end. For this reason, antiSMASH has now added two methods to predict the boundaries of BGCs. For fungal genomes, the Cluster Assignment by Islands of Sites (CASSIS) algorithm (15) is used for this purpose, which identifies genes within the BGC that share a common pathway-specific regulatory motif (Figure 1). Additionally, for both bacterial and fungal genomes, the user can now choose to use the ClusterFinder algorithm (16) to estimate cluster boundaries based on frequencies of locally encoded protein domains detected by Pfam (17) (based on these being either more or less BGC-like). If the user selects one of the BGC boundary prediction options (ClusterFinder for bacteria and fungi, CASSIS for fungi only), the extents of the predicted cluster region are displayed as bars above the BGC and also annotated in the GenBank files that can be downloaded.

New algorithms for non-ribosomal peptide and terpene chemistry prediction. Since the first version of antiSMASH, three algorithms have been used within the pipeline to predict the substrate specificities of non-ribosomal peptide synthetase (NRPS) adenylation (A) domains: the support-vector machine (SVM) and active-site motif (ASM) prediction methods from NRPSPredictor2 (18) and the profile HMM (pHMM)-based method from Minowa *et al.* Since then, several new algorithms have been published to predict A-domain specificity (19–21). More recently, Chevrette *et al.* (manuscript in review) substantially expanded the training sets for these algorithms, introduced an additional (phylogenetics-based) algorithm (PrediCAT), benchmarked all algorithms systematically and constructed an ensemble prediction method (called SANDPUMA) that outperformed each method individually. To benefit from the latest insights in this field, we have now replaced the previ-

ous prediction algorithms with the SANDPUMA predictions; these provide not only the ensemble outputs, but also the individual outputs of the underlying SVM, ASM, PrediCAT and pHMM algorithms. Since the benchmark comparison had shown the Minowa method (22) to be the least reliable of all previously published methods, this algorithm was judged to be uninformative and has been removed from the antiSMASH pipeline.

In addition to the prediction of non-ribosomal peptide chemistry, antiSMASH now also provides chemical structure predictions for the products of bacterial terpene synthases (23). To this end, a terpene cyclase-specific version of PrediCAT (see Supplementary Figure S1 and Table S1) has been included, to predict terpene cyclization patterns (such as 1,6-, 1,10- or 1,11 cyclizations) based on phylogenetic relationships with known enzymes from a documented reference set of terpene cyclases: when a query enzyme forms a monophyletic clade with enzymes with a known cyclization chemistry, this cyclization pattern is assigned to the query as a prediction. These predictions (see Supplementary Figure S1 for accuracy assessment) are then reported alongside the name of and sequence identity to the most closely related experimentally characterized homolog. It should be noted that the predictions are only performed for those terpene BGCs that encode mono-, sesqui- or diterpene cyclases (Pfams PF01397 and/or PF03936) and not for those that (only) encode phytoene synthases, tetraterpene cyclases, oxidosqualene cyclases, tryptophan dimethylallyltransferases, geranylgeranyl diphosphate (GGPP) synthases and/or lycopene cyclases.

Improved RiPP BGC identification and structure prediction. Ribosomally synthesized and Post-translationally modified Peptides (RiPPs) constitute a growing area of natural products research. antiSMASH supports researchers in predicting 15 distinct classes of RiPP BGCs. Previously, antiSMASH predicted only lanthipeptide precursors using a relatively limited pHMM-based approach. The current version of antiSMASH now provides a more sophisticated prediction and classification for class I lanthipeptides as well as lasso peptides, sactipeptides and thiopeptides. Given that RiPPs start as gene-encoded precursor peptides prior to post-translational modification, amino acid sequence prediction provides a wealth of information regarding the structure of the final product. However, the open-reading frames (ORFs) encoding these peptides are often overlooked by automated analysis and can be highly sequence variable, necessitating the need for current precursor identification methods.

To assist in identifying the precursor peptide-encoding gene, antiSMASH now utilizes the algorithm from the genome-mining platform Rapid ORF Description and Evaluation Online (RODEO) (24), which uses a combination of heuristic scoring, SVM and motif analysis to evaluate all candidate precursor peptides in a putative RiPP BGC. To broaden its applicability, the RODEO algorithm was extended to perform precursor prediction not only for lasso peptides, but also for thiopeptides, class I lanthipeptides and sactipeptides (see Supplementary Text 1 and Figures S2–4). When submitting an annotated nucleotide sequence to antiSMASH, the algorithm evaluates small genes that are already part of this annotation, as well as all other small ORFs in intergenic regions across the predicted cluster, in order to mitigate issues with gene prediction.

For the RiPP classes analyzed by the RODEO algorithm, antiSMASH reports: (i) the respective class of RiPP (e.g. lasso peptide or thiopeptide, etc.), (ii) a predicted leader peptide cleavage site and (iii) any potential C-terminal proteolytic processing. Given the post-translational simplicity of lasso peptides, a molecular mass is also calculated, accounting for the number of disulfide bridges. For thiopeptides, the macrocycle size and potential amidation are predicted as well. Molecular weight predictions are not given for the other RiPP subclasses owing to their extensive and variable post-translational modifications.

Tran s-*AT PKS domain alignments.* Several key classes of natural products are produced by multimodular enzymatic assembly lines. Standard similarity searches (as performed in antiSMASH's ClusterBlast module) do not reveal major insights between the natural product structures and the genes for the corresponding multidomain proteins that encode their biosynthetic enzymes. In order to better address this issue, we have now included an assembly line alignment method for *trans*-AT PKS (E. Helfrich, X. Lu *et al.* manuscript in preparation), which uses reference phylogenies of ketosynthase (KS) domains to assign KS domains from identified gene clusters into clades that correspond to a certain type of polyketide chemistry. Based on this classification, the encoded assembly line is then aligned to reference assembly lines from known BGCs in MIBiG (14) based on a distance metric that involves the Jaccard index, Goodman–Kruskal gamma function and domain duplication index of KS domain clades at empirically determined weights of 0.5, 0.25 and 0.25, respectively (see also (25)). The assembly lines that are most closely related to the query are then selected and clustered using Unweighted Pair Group Method with arithmetic mean clustering with the same metric and displayed in a visual alignment, in which each KS domain clade is annotated with a distinct color and a text description of the associated chemistry (Figure 2). This analysis allows for a rapid assessment of biochemical relationships between the products of these assembly lines, in order to identify new variants of known molecules or to find novel polyketide scaffolds.

TTA codon annotation. *Streptomyces* and related genera are important producers of clinically used antibiotics, such as tetracyclines or erythromycin, or drugs to treat parasitic worms such as avermectin. These bacteria have GC-contents of >70% and thus a skew toward higher GC triplets in their codon usage. While genes involved in primary metabolism almost exclusively use CTC codons to code for Leu, key genes in secondary metabolism and cell differentiation often contain TTA codons. As the expression of the TTA-codon specific Leu-tRNA-gene *bldA* is tightly controlled and the Leu-tRNA only accumulates in later stages of growth, this offers an additional level of regulation (26–28). The expression of the BGCs therefore does not only require activation at the transcriptional level, but also the presence of the TTA-specific Leu-tRNA. This must be considered, for example, for heterologous BGC expres-

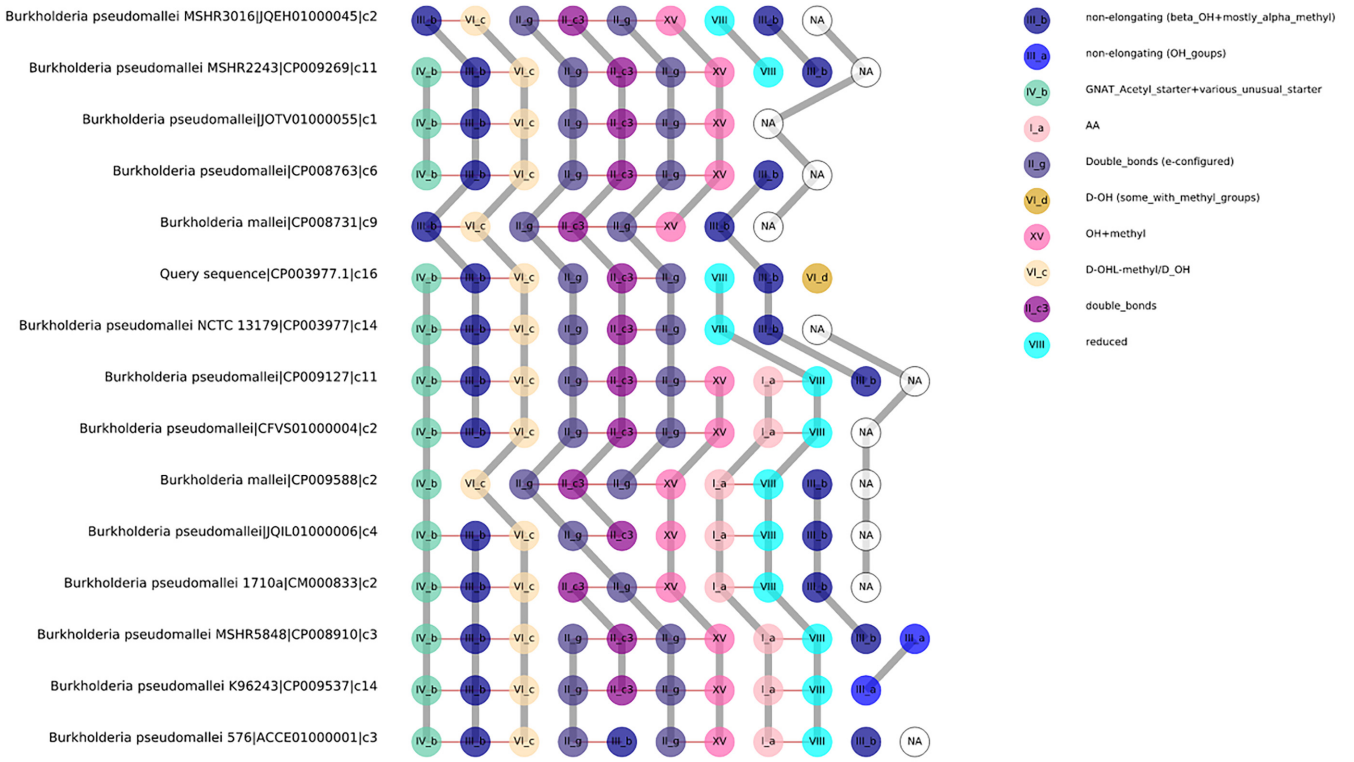

Figure 2. Visualization of *trans*-AT PKS assembly-line alignments. The top 15 most closely related assembly lines are visualized together with the query sequence (which represents the identified BGC currently in view). When clicking on a domain, its location (amino acid coordinates) within the parent protein are displayed and clicking on the gray connecting edges will trigger a display of the sequence identity between homologous domains based on a MAFFT multiple sequence alignment.

sion in other streptomycete hosts or metabolic engineering approaches. Therefore, a new feature was included in antiSMASH version 4 to automatically scan all BGCs for the presence of TTA codons and annotate these in the graphical cluster overview and the GenBank/EMBL result files.

Usability and efficiency improvements. antiSMASH comes with an updated, larger ClusterBlast database for comparative gene cluster analysis. In order to keep the runtime of the ClusterBlast analysis at acceptable levels with the much larger database, antiSMASH now uses the BLAST-compatible DIAMOND algorithm (29) to calculate results for ClusterBlast (against all ±220,000 BGCs currently detected in NCBI GenBank) and KnownClusterBlast (against experimentally characterized BGCs from MIBiG (14). ClusterBlast results are now cross-referenced to the antiSMASH database (13), whenever present there, through hyperlinks on the matched clusters; this allows researchers to quickly get a more complete view of these BGCs. Also, for each gene in a predicted gene cluster, an individual BLAST search is now automatically run against all proteins encoded in BGCs deposited in MIBiG (14); this helps researchers to predict functions of individual genes based on similarity of their encoded amino acid sequence to those of experimentally characterized proteins, even when the rest of the surrounding gene clusters are not similar.

In order to simplify selecting the correct input settings, separate submission pages were created for fungal sequences (http://fungismash.secondarymetabolites.org/) and plant sequences (http://plantismash.secondarymetabolites.org/). The main antiSMASH website is now focused on bacterial and archaeal sequences. The metabolic modeling functionality along with an EC number prediction option that were introduced in antiSMASH version 3 were removed again, as they led to extremely long run times and high server load. An updated version with improved reaction rules for secondary metabolite biosynthetic pathways will be released as a separate, but still closely linked program.

In addition to GenBank- and EMBL-formatted files, gene annotations can now also be added to FASTA sequences by also uploading a GFF3-formatted file. To assist job submission and retrieval from third-party tools running upstream or downstream analyses such as the CRISPR single guide RNA finding tool CRISPy-web (30) or the Antibiotics Resistance Target Seeker service (31), the antiSMASH web component now supports a REST-like (32) web API.

CONCLUSIONS AND FUTURE PERSPECTIVES

With the new features now introduced (Table 1), the antiSMASH framework continues to improve through the concerted action of researchers in the natural products community. A number of additional features are still in development, including application of the visual assembly line alignments to NRPSs, detailed gene cluster boundary prediction through phylogenetic profiling and detection of putative resistance genes inside BGCs.

Table 1. Overview of analyzes integrated into antiSMASH

Rule-based detection of BGCs	
Aminocoumarins	Microcin
Aminoglycosides / aminocyclitols	Microviridin
Aryl polyenes	Non-ribosomal peptides
Bacteriocins	Nucleosides
Beta-lactams	Oligosaccharide
Bottromycin	Others
Butyrolactones	Phenazine
ClusterFinder fatty acid	Phosphoglycolipids
ClusterFinder Saccharide	Phosphonate
Cyanobactins	Polyunsaturated fatty acids
(Dialkyl)resorcinols	Trans-AT type I PKS
Ectoines	Type I PKS
Furan	Type II PKS
Fused (Pheganomycin-like)	Type III PKS
Glycocin	Other (unusual) PKS
Head-to-tail cyclised peptide	Proteusin
Homoserine lactone	Sactipeptide
Indoles	Siderophores
Ladderane lipids	***Terpene***
Lantipeptides	***Thiopeptide***
Linear azol(in)e-containing peptides (LAPs)	
Lasso peptide	
Linaridin	
Melanins	
Rule-independent detection of BGCs	
ClusterFinder	
Cluster specific analyses	
Domain structure of PKSs and NRPSs[c] ***NRPS: A-domain specificity prediction (SANDPUMA)***	
PKS: AT specificity prediction	
Identification of conserved active site motifs; stereochemistry-determining motifs	
Prediction of core chemical structure (NRPS, PKS, lanthipeptides, ***lasso peptides, thiopeptides***) smCOG secondary metabolism-related gene family prediction	
TTA codon annotation for actinomycetes	
Improved prediction of gene cluster borders for fungal BGCs (CASSIS)	
Genome-wide analyses	
Protein family detection (PFAM) search	
Genome comparisons	
ClusterBlast (identification of similar clusters in sequence genomes)	
SubClusterBlast (identification of conserved operons with known function)	
KnownClusterBlast (identification of similar characterized gene clusters)	
TransAT-PKS Domain Alignments	
Links to other Web-resources	
antiSMASH-DB	
MIBiG repository	
NCBI BLAST+	
NaPDoS	
Norine	
Output file formats	
Genbank	
EMBL	
BiosynML	
Tab-delimited text files	
Input file formats	
FASTA (nucleotide or protein)	
FASTA + GFF3	
Genbank / Genpept	
EMBL	
Direct download via NCBI accession number	

With regard to chemistry prediction of the products of NRPSs and PKSs, we have opted to be conservative for the moment. The recently introduced PRISM pipeline (11) does a great job of automatically predicting a wide range of possible products of each BGC, which facilitates automated matching to large-scale metabolomic data. However, the majority of antiSMASH users still rely on manual comparison of BGCs with smaller-scale experimental data; we feel that this approach benefits more from reliable predictions of substructures and substrate specificities (and refraining from making lower-confidence combinatorial predictions). In this respect, PRISM and antiSMASH offer complementary functionalities and the user can opt to use either pipeline based on the intended research purposes.

We continue to strive for interoperability with other services. For example, antiSMASH predictions are also available through the Joint Genome Institute's IMG-ABC (33) as well as Genoscope's framework MicroScope (34); connections to EFI-EST (35) and other tools are being investigated. Also, we remain committed to collaborating with other researchers worldwide and invite expert feedback as well as technical contributions from the community to improve this important piece of software.

AVAILABILITY

antiSMASH is available from http://antismash.secondarymetabolites.org/. This website is free and open to all users and there is no login requirement. Source code is available from https://bitbucket.org/antismash/antismash/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Novo Nordisk Foundation (to S.Y.L., T.W.); The Netherlands Organization for Scientific Research (NWO) VENI Grant [863.15.002 to M.H.M.]; Graduate School for Experimental Plant Sciences (EPS) (to M.H.M.); Ministry of Science, ICT and Future Planning through the National Research Foundation (NRF) of Korea [NRF-2012M1A2A2026556 to H.U.K., S.Y.L.]; International Leibniz Research School for Microbial and Molecular Interactions (ILRS), as part of the excellence graduate school Jena School for Microbial Communication (JSMC), supported by the Deutsche Forschungsgemeinschaft (DFG) [to T. Wo.]; Collaborative Research Centre ChemBioSys (CRC 1127 ChemBioSys), by the DFG (to E.S.); NIH National Research Service Award [T32 GM008505 to M.G.C.]; David and Lucile Packard Fellowship for Science and Engineering (to D.A.M.); Department of Chemistry at the University of Illinois at Urbana–Champaign Fellowship (to C.J.S.); NIH Chemical Biology Interface Training Program Fellowship [T32 GM070421 to C.J.S.]; Google Summer of Code grant (to M.N.); Warwick Integrative Synthetic Biology Centre (WISB), and Manchester Synthetic Biology Research Centre (SYNBIOCHEM) funded under the UK Research Councils' 'Synthetic Biology for Growth' programme [BB/M017982/1 (WISB), BB/M017702/1 (SYNBIOCHEM)]. Funding for open access charge: Netherlands Organization for Scientific Research (NWO).
Conflict of interest statement. None declared.

REFERENCES

1. Ziemert,N., Alanjary,M. and Weber,T. (2016) The evolution of genome mining in microbes - a review. *Nat. Prod. Rep.*, **33**, 988–1005.

2. Weber,T. (2014) In silico tools for the analysis of antibiotic biosynthetic pathways. *Int. J. Med. Microbiol.*, **304**, 230–235.
3. 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.
4. Medema,M.H. and Fischbach,M.A. (2015) Computational approaches to natural product discovery. *Nat. Chem. Biol.*, **11**, 639–648.
5. Li,M.H.T., Ung,P.M.U., Zajkowski,J., Garneau-Tsodikova,S. and Sherman,D.H. (2009) Automated genome mining for natural products. *BMC Bioinformatics*, **10**, 185.
6. 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.
7. 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.
8. 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.
9. Ziemert,N., Podell,S., Penn,K., Badger,J.H., Allen,E. and Jensen,P.R. (2012) The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. *PLoS One*, **7**, e34064.
10. Johnston,C.W., Skinnider,M.A., Wyatt,M.A., Li,X., Ranieri,M.R.M., Yang,L., Zechel,D.L., Ma,B. and Magarvey,N.A. (2015) An automated Genomes-to-Natural Products platform (GNP) for the discovery of modular natural products. *Nat. Commun.*, **6**, 8421.
11. Skinnider,M.A., Dejong,C.A., Rees,P.N., Johnston,C.W., Li,H., Webster,A.L.H., Wyatt,M.A. and Magarvey,N.A. (2015) Genomes to natural products prediction informatics for secondary metabolomes (PRISM). *Nucleic Acids Res.*, **43**, 9645–9662.
12. Kautsar,S.A., Suarez Duran,H.G., Blin,K., Osbourn,A. and Medema,M.H. (2016) plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.*, doi:10.1093/nar/gkx305.
13. 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.
14. Medema,M.H., Kottmann,R., Yilmaz,P., Cummings,M., Biggins,J.B., Blin,K., de Bruijn,I., Chooi,Y.H., Claesen,J., Coates,R.C. *et al.* (2015) Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.*, **11**, 625–631.
15. Wolf,T., Shelest,V., Nath,N. and Shelest,E. (2016) CASSIS and SMIPS: promoter-based prediction of secondary metabolite gene clusters in eukaryotic genomes. *Bioinformatics*, **32**, 1138–1143.
16. 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.
17. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
18. Röttig,M., Medema,M.H., Blin,K., Weber,T., Rausch,C. and Kohlbacher,O. (2011) NRPSpredictor2–a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.*, **39**, W362–W367.
19. Prieto,C., García-Estrada,C., Lorenzana,D. and Martín,J.F. (2012) NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics*, **28**, 426–427.
20. Khayatt,B.I., Overmars,L., Siezen,R.J. and Francke,C. (2013) Classification of the adenylation and acyl-transferase activity of NRPS and PKS systems using ensembles of substrate specific hidden Markov models. *PLoS One*, **8**, e62136.
21. Baranašić,D., Zucko,J., Diminic,J., Gacesa,R., Long,P.F., Cullum,J., Hranueli,D. and Starcevic,A. (2014) Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing. *J. Ind. Microbiol. Biotechnol.*, **41**, 461–467.
22. Minowa,Y., Araki,M. and Kanehisa,M. (2007) Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. *J. Mol. Biol.*, **368**, 1500–1517.
23. Dickschat,J.S. (2016) Bacterial terpene cyclases. *Nat. Prod. Rep.*, **33**, 87–110.
24. Tietz,J.I., Schwalen,C.J., Patel,P.S., Maxson,T., Blair,P.M., Tai,H.C., Zakai,U.I. and Mitchell,D.A. (2017) A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat. Chem. Biol.*, **13**, 470–478.
25. Nguyen,D.D., Melnik,A.V., Koyama,N., Lu,X., Schorn,M., Fang,J., Aguinaldo,K., Lincecum,T.L. Jr, Ghequire,M.G.K., Carrion,V.J. *et al.* (2016) Indexing the *Pseudomonas* specialized metabolome enabled the discovery of poaeamide B and the bananamides. *Nat. Microbiol.*, **2**, 16197.
26. Hackl,S. and Bechthold,A. (2015) The Gene *bldA*, a regulator of morphological differentiation and antibiotic production in *Streptomyces*. *Arch. Pharm.*, **348**, 455–462.
27. Leskiw,B.K., Bibb,M.J. and Chater,K.F. (1991) The use of a rare codon specifically during development? *Mol. Microbiol.*, **5**, 2861–2867.
28. Leskiw,B.K., Lawlor,E.J., Fernandez-Abalos,J.M. and Chater,K.F. (1991) TTA codons in some genes prevent their expression in a class of developmental, antibiotic-negative, *Streptomyces* mutants. *Proc. Natl. Acad. Sci. U.S.A.*, **88**, 2461–2465.
29. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
30. Blin,K., Pedersen,L.E., Weber,T. and Lee,S.Y. (2016) CRISPy-web: an online resource to design sgRNAs for CRISPR applications. *Synth. Syst. Biotechnol.*, **1**, 118–121.
31. Alanjary,M., Kronmiller,B., Adamek,M., Blin,K., Weber,T., Huson,D.H., Philmus,B. and Ziemert,N. (2017) The Antibiotic Resistant Target Seeker (ARTS), an exploration engine for antibiotic cluster prioritization and novel drug target discovery. *Nucleic Acids Res.*, doi:10.1093/nar/gkx360.
32. Fielding,R.T. and Taylor,R.N. (2002) Principled design of the modern web architecture. *ACM Trans. Internet Technol.*, **2**, 115–150.
33. Hadjithomas,M., Chen,I.M.A., Chu,K., Ratner,A., Palaniappan,K., Szeto,E., Huang,J., Reddy,T.B.K., Cimermančič,P., Fischbach,M.A. *et al.* (2015) IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. *Mbio*, **6**, e00932.
34. Vallenet,D., Calteau,A., Cruveiller,S., Gachet,M., Lajus,A., Josso,A., Mercier,J., Renaux,A., Rollin,J., Rouy,Z. *et al.* (2017) MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. *Nucleic Acids Res.*, **45**, D517–D528.
35. Gerlt,J.A., Bouvier,J.T., Davidson,D.B., Imker,H.J., Sadkhin,B., Slater,D.R. and Whalen,K.L. (2015) Enzyme function initiative-enzyme similarity tool (EFI-EST): a web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta*, **1854**, 1019–1037.