antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers
Blin, Kai; Medema, Marnix H.; Kazempour, Daniyal; Fischbach, Michael A.; Breitling, Rainer; Takano, Eriko; Weber, Tilmann
Published in: Nucleic Acids Research
DOI: 10.1093/nar/gkt449

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

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

Publication date: 2013

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
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.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Download date: 15-05-2021
antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers

Kai Blin1, Marnix H. Medema2,3, Daniyal Kazempour1, Michael A. Fischbach4, Rainer Breitling3,5,*, Eriko Takano2,5,* and Tilmann Weber1,*

1Interfaculty Institute of Microbiology and Infection Medicine Tübingen, Eberhard Karls University Tübingen, 72076, Germany, 2Department of Microbial Physiology, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, 9747 AG, The Netherlands, 3Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, 9747 AG, The Netherlands, 4Department of Bioengineering and Therapeutic Sciences and California Institute for Quantitative Biosciences, University of California San Francisco, CA 94158, USA and 5Manchester Institute of Biotechnology, Faculty of Life Sciences, University of Manchester, M1 7DN, UK

Received February 24, 2013; Revised April 24, 2013; Accepted May 2, 2013

ABSTRACT

Microbial secondary metabolites are a potent source of antibiotics and other pharmaceuticals. Genome mining of their biosynthetic gene clusters has become a key method to accelerate their identification and characterization. In 2011, we developed antiSMASH, a web-based analysis platform that automates this process. Here, we present the highly improved antiSMASH 2.0 release, available at http://antismash.secondarymetabolites.org/. For the new version, antiSMASH was entirely re-designed using a plug-and-play concept that allows easy integration of novel predictor or output modules. antiSMASH 2.0 now supports input of multiple related sequences simultaneously (multi-FASTA/GenBank/EMBL), which allows the analysis of draft genomes comprising multiple contigs. Moreover, direct analysis of protein sequences is now possible. antiSMASH 2.0 has also been equipped with the capacity to detect additional classes of secondary metabolites, including oligosaccharide antibiotics, phenazines, thiopseudolactones, phosphonates and furans. The algorithm for predicting the core structure of the cluster end product is now also covering lantipeptides, in addition to polyketides and non-ribosomal peptides. The antiSMASH ClusterBlast functionality has been extended to identify sub-clusters involved in the biosynthesis of specific chemical building blocks. The new features currently make antiSMASH 2.0 the most comprehensive resource for identifying and analyzing novel secondary metabolite biosynthetic pathways in microorganisms.

INTRODUCTION

Many microorganisms produce secondary metabolites with interesting bioactivities, including antibiotics, anticancer agents and many other drugs (1).

For decades, the only way to identify and characterize such bioactive secondary metabolites involved a labor- and time-consuming procedure: one had to isolate new bacterial or fungal strains, cultivate them under different conditions, identify, isolate, purify and test any bioactive molecules that were produced and perform a complete chemical structure elucidation. The rapidly decreasing cost of whole-genome sequencing technologies enables new approaches that can greatly accelerate this process using bioinformatics analysis of the genome sequences of potential producer strains (2–4), before or in parallel with the biological/chemical isolation process. The fact that the biosynthetic pathways for many secondary metabolites are encoded by highly modular compact gene clusters facilitates this kind of analysis (5,6).

In recent years, many individual algorithms have been developed that cover specific steps in the bioinformatics analysis of secondary metabolite biosynthesis based on microbial genome sequences [for review (7,8)]. For example, ClustScan (9), CLUSEAN (10), SBSPKS (11) and SMURF (12) are tools for the identification and/or analysis of the enzymatic domains in multi-modular polyketide synthases and/or non-ribosomal peptide synthases.
synthetases, which are the key enzymes for the synthesis of the largest classes of clinically important secondary metabolites. These include, e.g. non-ribosomal peptide antibiotics like penicillin and polyketide macrolides like the immunosuppressant tacrolimus. NRPSpredictor (13,14), NRPSSP (15) and the PKS/NRPS predictive BLAST Server (16) are sophisticated tools for the prediction of substrate specificities of key biosynthetic steps, allowing an approximate prediction of the chemical structure of bioactive end compounds based on the genome sequence (Table 1).

In 2011, we released the first version of the ‘antibiotics and secondary metabolite analysis shell’ (antiSMASH), a web server and stand-alone software, which combines automated identification of secondary metabolite gene clusters in genome sequences with a large collection of compound-specific analysis algorithms (17). Within the past two years, antiSMASH has become the standard tool to analyze genomes of bacteria and fungi for their potential to produce secondary metabolites. Since the start of the service, the stand-alone software has been downloaded >3200 times, and >28 000 antiSMASH jobs have been submitted to the antiSMASH web server; the monthly data volume currently processed is >12 Gb. antiSMASH also supports the manual PKS/NRPS cluster curation effort of the ClusterMine360 database (18) by providing a standardized annotation basis.

Here, we present version 2.0 of antiSMASH. The software has been entirely restructured internally, and it now uses a plug-and-play concept for easier maintainability and extensibility. A number of novel cluster detection and analysis features have been added to cover the broadest possible range of secondary metabolite classes. Finally, the web-based user interface was completely redesigned for better usability and a wider range of possible input files, allowing, e.g. the analysis of unassembled draft genomes and metagenomic sequences.

### MATERIALS AND METHODS

#### Implementation of new features

The basic steps of an antiSMASH analysis have been described by Medema et al. (17): first, potential biosynthetic gene clusters are identified by comparing each gene product encoded on the uploaded DNA sequence against a manually curated collection of profile hidden Markov models (pHMMs). These pHMMs describe key biosynthetic enzymes of the 24 secondary metabolite classes detectable by antiSMASH, using the HMMer3 software (19). Key enzymes encoded in each gene cluster are assigned to secondary metabolite-specific clusters of orthologous groups (smCOGs). Depending on the class of the detected secondary metabolite gene cluster, further detailed analyses are performed: the domains of multimodular polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) are identified by a pHMM-based approach. Specificities of enzymes are determined by analyzing active site residues using integrated third-party algorithms and tools, such as the methods of Minowa et al. (20) and NRPSpredictor2 (14).
for the prediction of NRPS adenylation domain specificities. Based on these data, a core chemical structure of the putative biosynthesis product is generated and displayed. In addition, an integrated version of MultiGeneBlast (21), ClusterBlast, is used to identify similar gene clusters in a comprehensive gene cluster database. antiSMASH 2.0 can be either installed locally on Windows, Mac OS X or Linux computers, or be accessed via the internet at http://antismash.secondarymetabolites.org (recommended). The use of the antiSMASH web server is free of charge and does not require registration or login data. Voluntarily, the users can provide an email address, which is used to send information and the link of the results, once the computing of the antiSMASH 2.0 results is finished. The data are stored on the server for 30 days and are deleted afterward.

Although the general strategy of antiSMASH has not changed in version 2.0, many improvements have been implemented in the new version, which we outline here.

**New file and input options**

antiSMASH 2.0 now makes it easier to work with draft genomes consisting of a large number of individual sequence records: support has been added for multi-GenBank, multi-EMBL, as well as multi-FASTA files. If the NCBI download option yields a whole-genome shotgun (WGS) master or supercontig record, antiSMASH 2.0 will download all constituent single WGS records from NCBI as well and combine all of them into a single output (Figure 1). For prokaryotic FASTA inputs, antiSMASH 2.0 now also offers the option to perform the initial search for gene cluster signature genes on all open reading frames of >60 nt throughout all six translation frames of a nucleotide sequence, before running the standard gene prediction with Glimmer. This avoids that mistakes in the gene prediction stage lead to false negatives in the gene cluster prediction stage. After the gene prediction stage, all open reading frames that match to pHMMs in the antiSMASH pHMM library are retained in the gene cluster output, even if they were not predicted as genes by Glimmer.

In addition to nucleotide sequences, antiSMASH 2.0 can now also be used to analyze PKS, NRPS and lantipeptide precursor amino acid sequences directly: their protein sequences can either be analyzed by specifying their NCBI GenPept accession numbers or by pasting the FASTA sequences directly into an input field.

**Detection of secondary metabolite gene clusters in sequence data**

In addition to the secondary metabolite cluster types supported in the original release of antiSMASH (type I, II and III polyketides, non-ribosomal peptides, terpenes, lantipeptides, bacteriocins, aminoglycosides/aminocyclitols, β-lactams, aminocoumarins, indoles, butyrolactones, eotoines, siderophores, phosphoglycolipids, melanins and a generic class of clusters encoding unusual secondary metabolite biosynthesis genes), version 2.0 adds support for oligosaccharide antibiotics, phenazines, thiopseudocins, homoserine lactones, phosphonates and furans. The cluster detection uses the same pHMM rule-based approach as the initial release (17): in short, the pHMMs are used to detect signature proteins or protein domains that are characteristic for the respective secondary metabolite biosynthetic pathway. Some pHMMs were obtained from PFAM or TIGRFAM. If no suitable pHMMs were available from these databases, custom pHMMs were constructed based on manually curated seed alignments (Supplementary Table S1). These are composed of protein sequences of experimentally characterized biosynthetic enzymes described in literature, as well as their close homologs found in gene clusters from the same type. The models were curated by manually inspecting the output of searches against the non-redundant (nr) database of protein sequences. The seed alignments are available online at http://antismash.secondarymetabolites.org/download.html#extras. After scanning the genome with the pHMM library, antiSMASH evaluates all hits using a set of rules (Supplementary Table S2) that describe the different cluster types. Unlike the hard-coded rules in the initial release of antiSMASH, the detection rules and profile lists are now located in editable TXT files, making it easy for users to add and modify cluster rules in the stand-alone version, e.g. to accommodate newly discovered or proprietary compound classes without code changes. The results of gene cluster predictions by antiSMASH are continuously checked on new data arising from research performed throughout the natural products community, and pHMMs and their cut-offs are regularly updated when either false positives or false negatives become apparent.

The profile-based detection of secondary metabolite clusters has now been augmented by a tighter integration of the generalized PFAM (22) domain-based ClusterFinder algorithm (Cimermancic et al., in preparation) already included in version 1.0 of antiSMASH. This algorithm performs probabilistic inference of gene clusters by identifying genomic regions with unusually high frequencies of secondary metabolism-associated PFAM domains, and it was designed to detect ‘classical’ as well as less typical and even novel classes of secondary metabolite gene clusters. While antiSMASH 1.0 only generated the output of this algorithm in a static image, version 2.0 displays these additional putative gene clusters along with the other gene clusters in the HTML output. A key advantage of this is that these putative gene clusters will now also be included in the subsequent (Sub)ClusterBlast analyses.

**Metabolite-specific detection modules**

antiSMASH version 2.0 adds lantipeptide-specific chemical core structure analysis to the existing set of NRPS/PKS core prediction tools. If one or more open reading frames encoding putative lantipeptide prepropeptides are found, antiSMASH predicts the core peptide molecular mass and sequence after leader peptide cleavage. The leader peptide cleavage motifs are identified via pHMMs specific for cleavage sites of class I–IV lantipeptides, respectively. The best-matching profile determines the classification of the prepropeptide, and the cleavage site is calculated from the pHMM-sequence alignment.
To obtain the core peptide mass, all serine and threonine residues in the core peptide are assumed to be dehydrated to didehydro-alanine (Dha) and didehydro-butyrine (Dhb), the most frequent post-translational modification in lantipeptides. Reported masses are the monoisotopic masses of the most prevalent isotopomers. The number of lanthionine/methyl-lanthionine bridges is calculated from the number of cysteine, Dha and Dhb residues available for bridge formation (Blin et al., in preparation).

**SubclusterBlast**

Extending the ClusterBlast analysis that identifies homologous gene clusters across many published genome sequences, we have added a new option to identify operons related to the biosynthesis of precursors or specific
chemical moieties in a gene cluster’s end product. This new analysis module, SubclusterBlast, performs blastp searches of the amino acid translations of all cluster genes against a database containing 126 sub-clusters from gene clusters encoding known compounds (Figure 2). These sub-clusters code for the biosynthesis of precursors, such as 6-methylsalicylic acid, 3-amino-5-hydroxybenzoic acid, ethylmalonyl-CoA, deoxysugars and hydroxyphenylglycine, which are highly specific for certain classes of bioactive compounds. Hence, their presence in a genome allows more confident conclusions about the biosynthetic capacities of an organism. The hits are sorted in the same way as the ClusterBlast hits (17), but they are gathered with stricter thresholds: a minimal
percentage identity of 45% and a minimal sequence coverage of 40% are required. The highest-scoring subcluster hits are then displayed on the results page using an annotated vector graphic similar to the general ClusterBlast output.

Output and visualization
When antiSMASH has finished the computation of an analysis, it now provides an overview table that displays all identified secondary metabolite biosynthesis gene clusters with links to the respective prediction details, as a convenient starting point for further analysis (Figure 1). For nucleotide inputs consisting of multiple GBK/EMBL/FASTA entries, the results are separated per entry. Because of the large size of the antiSMASH results webpage in version 1.0, loading took a long time and sometimes even caused timeout error messages in the user’s web browser. Therefore, the visualization component of antiSMASH 2.0 was completely re-designed, resulting in a reduction of transfer data volume and greatly accelerated display, even for results containing many cluster hits.

The overall layout of the interactive results page has been retained (Figure 3): in the top section, the identified clusters are displayed as circles that serve as direct links to the clusters. In antiSMASH 2.0, the circles are color coded depending on the class of the identified cluster to ease navigation by the user. The individual cluster result pages are now reachable via the result URL, making it possible to both bookmark and direct other people to specific cluster pages. Individual cluster result pages contain an interactive graphical representation of the genes identified in the cluster. Again, color coding was added to represent the functional classes of the gene cluster genes according to an smCOG-based classification: biosynthesis, transport, regulation or other. For modular enzymes (NRPS, PKS) or lantipeptides, detailed annotation sections provide information on the domain organization and the putative cleavage sites and molecular weights, respectively. At the bottom of the page, graphical representations of the ClusterBlast results and—if available—the SubclusterBlast results are displayed. For several classes of antibiotics, where the analysis of the gene clusters allows the prediction of core structures of the biosynthetic products, a predicted structure and detailed information on the prediction source are displayed in a box on the right side of the results page (Figure 3). For lantipeptides and NRPS products, there is a direct link to the NORINE (24) peptide database. The information displayed on the interactive webpage is also annotated in EMBL- or GenBank-formatted sequence files, which can be downloaded and used with standard sequence analysis software. In addition, an archive containing all data including the webpage can be saved for later use.

Plug-and-play architecture
In antiSMASH 2.0, the software architecture has been completely re-designed to make it easily extendable: the core program reads in ‘analysis plug-ins’ that are either general or specific to a certain gene cluster type ‘output plug-ins’ facilitate the output of the results to HTML, GBK, EMBL, TXT and XLS files. To make it easy for users to customize antiSMASH for their own analyses, we provide a plug-in template from the download section of http://antismash.secondarymetabolites.org, which can be used to design custom plug-ins, e.g. for reading user-specific input formats or analyzing novel cluster types.

RESULTS AND DISCUSSION
With options to upload DNA sequences of both finished genomes and draft sequences, to make antiSMASH download published sequences from NCBI and to analyze amino acid sequences directly, antiSMASH 2.0 now covers all common types of input data. For draft genome data published in the NCBI genome database, antiSMASH can automatically download the records specified in the WGS summary record. As a test for the downloader, the recently published Oxytricha trifallax WGS record (Genbank accession no. AMCR00000000.1) consisting of 22,363 contigs was run via the internet interface, and the server handled the large amount of contigs and sequence data (67 Mb) without issues. For prokaryotic genome sequences, draft genome support increases the number of genomes that can be processed directly via NCBI accession numbers from 2570 to 8898, a ~2.5-fold increase of available sequences. One important caveat should be noted: when analyzing draft genomes, the number of detected gene clusters reported by antiSMASH can be artificially high because gene clusters can be fragmented across multiple contigs, and antiSMASH detects all fragments as separate gene clusters. On the other hand, some contigs with gene cluster fragments might be left undetected, if the subset of genes present on a contig does not suffice to match the criteria for gene cluster detection by antiSMASH.

antiSMASH 2.0 now supports 24 secondary metabolite cluster types via profile-based detection of their core biosynthetic genes (up from 19). In test runs on 28 known gene clusters encoding compounds of the newly added classes, all of them were detected successfully (Supplementary Table S3). To assess the general accuracy of the antiSMASH predictions, we selected the same test set of genomes as for the original version (17): the genomes of the proteobacterium Pseudomonas fluorescens Pf-5 (25), the actinomyecetes Streptomyces griseus IFO 13350 (26), Kitasatospora setae NBRC 14216T (27) and Salinispora tropica CNB-440 (28) and the fungus Aspergillus fumigatus AF293 (29) were analyzed with antiSMASH 2.0 and compared with the manually identified clusters referred to in the original publications. In all, 97.3% of clusters (108 of 111) that were assigned manually were also identified by antiSMASH 2.0. This is the same performance as with antiSMASH 1.0, which was expected, as the established cluster finding algorithm has not changed in version 2.0. In addition to the 35 clusters that were predicted by antiSMASH 1.0 but were missed in the original publications, four additional clusters were identified by the new detection modules of antiSMASH 2.0, increasing the percentage of newly found
gene clusters from 31.5 to 35.1% (Supplementary Table S4).

If further extension of the prediction ability is desired, new profiles can be added easily and without changes to the core code of the software using the new plug-and-play architecture of antiSMASH 2.0. The new version can also cast a wider net than the original version, by using improved ways to exploit the outputs of the ClusterFinder inclusive search algorithm for putative clusters (Cimermancic et al., in preparation). Although the inclusive algorithm is likely to identify too many clusters, the combination with homology search methods allows focusing on the ones with homology to previously identified secondary metabolite clusters.

A major goal of antiSMASH 2.0 was to increase usability. Because antiSMASH 1.0 loaded all the results simultaneously when loading/opening the HTML output file, it was slow for the typical large results files: e.g. loading the 35 cluster results for *Streptomyces tsukubaensis* NRRL18488 (Genbank accession no. AJSZ01000001) from a local hard drive took ~40 s on a fast PC. In contrast, antiSMASH 2.0 output for the same data now
loads in <2 s, even though more clusters (37) are detected. The reduced result page size has the added benefit of being accessible from smart phones and tablets (tested for iOS and Android).

antiSMASH 2.0 is currently the most comprehensive software for genome mining and analysis of secondary metabolite biosynthetic pathways, and it includes or provides direct links to the most significant other tools and algorithms for this task. The updates to the antiSMASH framework will enable it to be successfully used with the latest sequencing technologies and biochemical insights, whereas it will continue to be a key tool for state-of-the-art synthetic biology approaches towards secondary metabolism (23).

**SUPPLEMENTARY DATA**

Supplementary Data are available at NAR Online: Supplementary Tables 1–4 and Supplementary References [30,31].

**FUNDING**

German Ministry of Education and Research (BMBF) [0315585A to T.W.]; German Centre for Infection Research (DZIF) [8000-402-2 to T.W.]; Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs [STW 10463 to E.T.]; NWO-Vidi fellowship (to R.B.). Funding for open access charge: Deutsche Forschungsgemeinschaft and Open Access Publishing Fund of Tübingen University.

**Conflict of interest statement.** None declared.

**REFERENCES**


