The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters

Blin, Kai; Pascal Andreu, Victòria; de Los Santos, Emmanuel L C; Del Carratore, Francesco; Lee, Sang Yup; Medema, Marnix H.; Weber, Tilmann

Published in:
Nucleic acids research

Link to article, DOI:
10.1093/nar/gky1060

Publication date:
2019

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

Link back to DTU Orbit

Citation (APA):
The antiSMASH database version 2: a comprehensive resource on secondary metabolite biosynthetic gene clusters

Kai Blin1, Victòria Pascal Andreu2, Emmanuel LC de los Santos3, Francesco Del Carratore4, Sang Yup Lee1,5, Marnix H. Medema2 and Tilmann Weber1,*

1Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kemitorvet, Building 220, 2800 Kgs. Lyngby, Denmark, 2Bioinformatics Group, Wageningen University, Wageningen, Netherlands, 3Warwick Integrative Synthetic Biology Centre, University of Warwick, Coventry, UK, 4Faculty of Science and Engineering, Manchester Institute of Biotechnology, University of Manchester, Manchester, UK and 5Department of Chemical and Biomolecular Engineering (BK21 Plus Program), Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Republic of Korea

Received September 15, 2018; Revised October 15, 2018; Editorial Decision October 16, 2018; Accepted October 17, 2018

ABSTRACT

Natural products originating from microorganisms are frequently used in antimicrobial and anticancer drugs, pesticides, herbicides or fungicides. In the last years, the increasing availability of microbial genome data has made it possible to access the wealth of biosynthetic clusters responsible for the production of these compounds by genome mining. antiSMASH is one of the most popular tools in this field. The antiSMASH database provides pre-computed antiSMASH results for many publicly available microbial genomes and allows for advanced cross-genome searches. The current version 2 of the antiSMASH database contains annotations for 6200 full bacterial genomes and 18,576 bacterial draft genomes and is available at https://antismash-db.secondarymetabolites.org/.

INTRODUCTION

A majority of antibacterial and antifungal drugs, as well as drugs for many other indications, are derived from microbial natural products (1). Traditionally, bioactive natural compounds were identified via classical isolation and analysis approaches. The increasing availability of genomic data in the last two decades allows us to complement these approaches with genome mining to identify and characterize biosynthetic pathways for natural products in genome and metagenome data (2). Specialized software to support researchers in their search for natural products has been available for some years (for a comprehensive overview/list of such tools, please see (3-5)). Since its initial release in 2011, antiSMASH (6-9) has established itself as a standard tool for secondary metabolite genome mining and is currently the most widely used software pipeline for this task.

antiSMASH uses a rule-based cluster detection approach to identify 45 different types of secondary metabolite biosynthetic pathways via their core biosynthetic enzymes. For nonribosomal peptide synthases, type I polyketides, terpenes, lanthipeptides, thiopeptides, sactipeptides and lasopeptides, antiSMASH can also provide more detailed predictions of the compounds produced by the respective biosynthetic gene clusters (BGCs). Identified clusters are compared to a database of clusters previously predicted by antiSMASH using the built-in ClusterBlast algorithm. A similar algorithm, KnownClusterBlast is used to compare the identified cluster against the manually curated set of known BGCs from the MIBiG (10) database. Secondary metabolite clusters of orthologous group (smCoG) classification is used to assign functions to gene products in the predicted BGCs.

As antiSMASH is a genome mining pipeline designed to analyze individual genomes, we developed the antiSMASH database (11) to provide interconnections and cross-genome search functionality based on antiSMASH results for many publicly available microbial genomes. Moreover, it provides users with instant access to full antiSMASH results of publicly available genome sequences. Here we present version 2 of the antiSMASH database. The database content of version 1, which was generated with version 3 of antiSMASH, was updated with annotation of the current antiSMASH 4.2.1 release. This implies that the antiSMASH database now includes updated detection rules, updated ClusterBlast database links, TTA codon prediction, NRPS-A domain predictions by the up-to-date SANDPUMA software (12), classification of ter-

*To whom correspondence should be addressed. Tel: +45 24 89 61 32; Email: tiwe@biosustain.dtu.dk

© The Author(s) 2018. 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.
for complex network research, InterJournal, Complex Sys-
age (Csardi G, Nepusz T: The igraph software package <
20 kb and assembly anomalies). We then performed the
pre-filtered this set to remove poor quality genomes (N50
prokaryotic genomes (https://doi.org/10.1101/225342). We
precomputed fastANI similarity scores of ninety thousand
antiSMASHdatabase.Toselectdraftgenomesforaddition
somal assemblies, we added a set of draft genomes to the
antiSMASH database.

L50. This resulted in a total of 6,200 complete genomes for
assemblies with the highest contig N50 and lowest contig L50. In order to maintain consist-
ency with the complete and chromosomal set, only draft
genomes that had corresponding RefSeq assemblies were
included in the database. The following resulted in an addi-
tional 18,576 draft genome entries that were added to the
database.

MATERIALS AND METHODS
Selection of included genomes
Microbial genome resources are growing rapidly and, de-
spite taxonomically novel genomes being released fre-
quently, there is a lot of sequence redundancy in the NCBI
genome databases, i.e. thousands of sequences of mostly
pathogenic bacteria such as Pseudomonas aeruginosa or Es-
erichia coli. Therefore, with the objective of creating a rep-
resentative set of genomes that are non-redundant, we de-
signed an approach to effectively update the antiSMASH
database, maintaining its high quality and adequately rep-
resenting natural diversity without significantly decreasing
the overall pipeline performance in terms of speed.

Genomes categorized as ‘draft genomes’ are fragmented
in multiple contigs. As many secondary metabolite biosyn-
thetic gene cluster contain repetitive sequences, this im-
plies that many BGCs end up being split on multiple con-
tigs without any linkage information, leading to low-quality
BGC data. Consequently, in order to minimize this issue
we prioritized the inclusion of NCBI RefSeq genomes that
were annotated with the assembly level ‘complete genome’
or ‘chromosome’ present in the database on April 2018
(10 863 genomes in total). We then estimated the distance
between selected assemblies using fastANI (Average Nu-
cleotide Identity) (https://github.com/ParBLISS/FastANI).
FastANI uses a hash-based algorithm to estimate the av-
erage nucleotide identity between pairs of genomic assem-
bies. A network was generated with each genome as a
node, and weighted edges between nodes corresponding
to the fastANI estimate between genomes. We used a fas-
tANI similarity score of 99.6 as a cutoff for having an edge
between nodes. Nodes were then assigned to commun-
ities using the multilevel community structure algorithm
(https://arxiv.org/abs/0803.0476) in the igraph Python pack-
age (Csardi G, Nepusz T: The igraph software package
for complex network research, InterJournal, Complex Sys-
genome from each community was chosen by prioritizing
assemblies with the highest contig N50 and lowest contig
L50. This resulted in a total of 6,200 complete genomes for
the antiSMASH database.

In order to supplement the set of complete and chromo-
somal assemblies, we added a set of draft genomes to the
antiSMASH database. To select draft genomes for addition
to the database, we started with a previously published set
of precomputed fastANI similarity scores of ninety thousand
prokaryotic genomes (https://doi.org/10.1101/225342). We
pre-filtered this set to remove poor quality genomes (N50
< 20 kb and assembly anomalies). We then performed the
same procedure as with the complete and chromosomal as-
semblies to group the draft genomes into communities. A representative genome from each community was chosen by
prioritizing assemblies based on assembly level (scaffold >
contig), and then selecting assemblies with the highest con-
tig N50 and lowest contig L50. In order to maintain consist-
ency with the complete and chromosomal set, only draft
genomes that had corresponding RefSeq assemblies were
included in the database. The following resulted in an addi-
tional 18,576 draft genome entries that were added to the
database.

antiSMASH annotations and data import
Based on the selection criteria mentioned above, the
assemblies were downloaded from the NCBI servers in Gen-
Bank format using the ncbi-genome-download tool (https:
//github.com/kbli/ncri-genome-download/). GNU paral-
lel (14) was used to run multiple docker containers of anti-
SMASH 4.2.1 simultaneously. Different analysis param-
eters were used for the full and partial genome set. For
full genomes, ClusterBlast, KnownClusterBlast, SubClus-
terBlast, ActiveSiteFinder, TTA codon detection in automatic
mode, secondary metabolite clusters of ortholo-
gous groups prediction, and cluster-specific detailed an-
notations were run (command line flags: –clusterblast –
knownclusterblast–subclusterblast–asf–tttauto–smcogs-
notree). For draft genomes, antiSMASH was run in fast
mode, skipping the detailed annotations. Additionally,
KnownClusterBlast, TTA codon detection in automatic
mode, and secondary metabolite clusters of orthologous
groups prediction were run (command line flags: –minimal

The SQL schema of the (https://github.com/antismash/
db-schema/) antiSMASH database was updated to
accommodate the annotation changes and additional
features/predictions that were introduced by antiSMASH
version 4. The antiSMASH results in GenBank format
were loaded into the SQL schema using the import script
available at https://github.com/antismash/db-import/.

RESULTS AND DISCUSSION
With an update to the PGAP annotation pipeline used by
the NCBI, the annotation issues causing us to use records
from GenBank instead of RefSeq for version 1 of the anti-
SMASH database have largely been resolved. Hence, with
version 2 of the database, we have switched to using RefSeq
genomes to obtain more unified gene annotations.

The antiSMASH database 2 contains BGCs identified
in 6,200 full genomes (an increase of 58%) and adds 18
576 draft genomes. Annotations in the database are gen-
erated by antiSMASH version 4.2.1, the most recent re-
lease of antiSMASH (9). New in the antiSMASH 4.2.1
release are detection rules for N-acyl amino acids, poly-
brominated diphenyl ethers, and PPY-like pyrones. De-
tailed cluster product predictions have been added for lasso
peptides, thiopeptides, sactipeptides (based on RODEO
(15)), non-ribosomal peptide synthases (based on SAND-
PUMA (12)) and terpenes. The ClusterBlast and Known-
ClusterBlast databases have been updated.
### Table 1. Overview on BGC numbers in version 1 and version 2 of the antiSMASH database

<table>
<thead>
<tr>
<th>Overall database statistics</th>
<th>Version 1 counts</th>
<th>Version 2 counts</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full (high quality) genomes</td>
<td>3907</td>
<td>6200</td>
<td>58</td>
</tr>
<tr>
<td>Number of BGCs in full genomes</td>
<td>22 292</td>
<td>32 548</td>
<td>46</td>
</tr>
<tr>
<td>Draft genomes</td>
<td>0</td>
<td>18 576</td>
<td>New</td>
</tr>
<tr>
<td>Number of BGCs in draft genomes</td>
<td>0</td>
<td>119 558</td>
<td>New</td>
</tr>
<tr>
<td>BGCs in total</td>
<td>22 292</td>
<td>152 106</td>
<td>682</td>
</tr>
</tbody>
</table>

### Table 2. Changes in cluster counts of the different BGC types between version 1 and version 2 of the antiSMASH database (excluding data from draft genomes)

<table>
<thead>
<tr>
<th>Gene cluster types</th>
<th>Version 1 counts</th>
<th>Version 2 counts</th>
<th>% change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NRPS</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonribosomal peptide</td>
<td>5878</td>
<td>7893</td>
<td>34</td>
</tr>
<tr>
<td><strong>Terpenes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Terpene</td>
<td>3362</td>
<td>5018</td>
<td>49</td>
</tr>
<tr>
<td><strong>Polyketides</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type I polyketide</td>
<td>2608</td>
<td>3302</td>
<td>27</td>
</tr>
<tr>
<td>Type III polyketide</td>
<td>742</td>
<td>1141</td>
<td>54</td>
</tr>
<tr>
<td>HgI-E-type polyketide</td>
<td>590</td>
<td>768</td>
<td>30</td>
</tr>
<tr>
<td>Trans-AT polyketide</td>
<td>512</td>
<td>623</td>
<td>22</td>
</tr>
<tr>
<td>Type II polyketide</td>
<td>173</td>
<td>307</td>
<td>77</td>
</tr>
<tr>
<td><strong>RiPPs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bacteriocin/RiPP</td>
<td>3323</td>
<td>5198</td>
<td>56</td>
</tr>
<tr>
<td>Lanthipeptide</td>
<td>857</td>
<td>1121</td>
<td>31</td>
</tr>
<tr>
<td>Thiopipeptide</td>
<td>122</td>
<td>1097</td>
<td>799</td>
</tr>
<tr>
<td>Lasso peptide</td>
<td>351</td>
<td>562</td>
<td>60</td>
</tr>
<tr>
<td>Sactipeptide</td>
<td>59</td>
<td>318</td>
<td>439</td>
</tr>
<tr>
<td>Microviridin</td>
<td>18</td>
<td>70</td>
<td>289</td>
</tr>
<tr>
<td>Head-to-tail cyclised</td>
<td>22</td>
<td>52</td>
<td>136</td>
</tr>
<tr>
<td>(subtilosin-like)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteusin</td>
<td>13</td>
<td>39</td>
<td>200</td>
</tr>
<tr>
<td>Microcin</td>
<td>5</td>
<td>3</td>
<td>–40</td>
</tr>
<tr>
<td>Bottromycin-like</td>
<td>1</td>
<td>2</td>
<td>100</td>
</tr>
<tr>
<td><strong>Other</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>1887</td>
<td>2322</td>
<td>23</td>
</tr>
<tr>
<td>Siderophore</td>
<td>1399</td>
<td>1745</td>
<td>25</td>
</tr>
<tr>
<td>Homoserine lactone</td>
<td>1084</td>
<td>1608</td>
<td>48</td>
</tr>
<tr>
<td>Aryl polyene</td>
<td>988</td>
<td>1595</td>
<td>61</td>
</tr>
<tr>
<td>Ectoine</td>
<td>424</td>
<td>794</td>
<td>87</td>
</tr>
<tr>
<td>Butyrolactone</td>
<td>189</td>
<td>392</td>
<td>107</td>
</tr>
<tr>
<td>Phosphonate</td>
<td>248</td>
<td>342</td>
<td>38</td>
</tr>
<tr>
<td>Resorcinol</td>
<td>184</td>
<td>261</td>
<td>42</td>
</tr>
<tr>
<td>Ladderane</td>
<td>113</td>
<td>217</td>
<td>92</td>
</tr>
<tr>
<td>Phenazine</td>
<td>152</td>
<td>210</td>
<td>38</td>
</tr>
<tr>
<td>Melanin</td>
<td>45</td>
<td>113</td>
<td>151</td>
</tr>
<tr>
<td>N-acetyl amino acid cluster</td>
<td>0</td>
<td>110</td>
<td>New</td>
</tr>
<tr>
<td>Indole</td>
<td>48</td>
<td>104</td>
<td>117</td>
</tr>
<tr>
<td>Cyanobactin</td>
<td>30</td>
<td>77</td>
<td>157</td>
</tr>
<tr>
<td>Polyunsaturated fatty acid</td>
<td>45</td>
<td>61</td>
<td>36</td>
</tr>
<tr>
<td>Oligosaccharide</td>
<td>40</td>
<td>54</td>
<td>35</td>
</tr>
<tr>
<td>Aminoglycoside/aminocyclitol</td>
<td>26</td>
<td>51</td>
<td>96</td>
</tr>
<tr>
<td>Nucleoside</td>
<td>23</td>
<td>49</td>
<td>113</td>
</tr>
<tr>
<td>Linamarin</td>
<td>17</td>
<td>35</td>
<td>106</td>
</tr>
<tr>
<td>beta-lactam</td>
<td>13</td>
<td>30</td>
<td>131</td>
</tr>
<tr>
<td>Aminocoumarin</td>
<td>3</td>
<td>10</td>
<td>233</td>
</tr>
<tr>
<td>Pheganomycin-like ligase</td>
<td>5</td>
<td>7</td>
<td>40</td>
</tr>
<tr>
<td>Phosphoglycolipid</td>
<td>1</td>
<td>4</td>
<td>300</td>
</tr>
<tr>
<td>Furan</td>
<td>2</td>
<td>3</td>
<td>50</td>
</tr>
<tr>
<td>Glycocin</td>
<td>14</td>
<td>3</td>
<td>–79</td>
</tr>
<tr>
<td>Polybrominated diphenyl ether</td>
<td>0</td>
<td>1</td>
<td>New</td>
</tr>
</tbody>
</table>
Figure 1. Statistic summary of the antiSMASH database version 2. (A) A phylogenetic tree constructed from the revised version of tree of life based on 120 conserved protein markers (16). The original tree was pruned by genome assembly id using ETE Toolkit (17), to only keep leaves that belong to genomes of the antiSMASH database version 2. The visualization and customization of the tree was performed with iTOL (18). As a result, 12 219 leaves from the total of 24,776 bacterial genomes are shown in this phylogeny. The colored ring represents the eight most abundant phyla; 97.6% of the genomes, and the bar plots in the outer ring the number of BGCs per genome. (B) Boxplots of the BGCs counts per phylum, with the values on top showing the total number of complete genomes per phylum. (C) Pie charts of the five major BGC classes per phylum showing the diversity of natural products produced by each group of bacteria.

The search builder has been extended to cover these new features. A new search field in the taxonomy browser makes it easier to navigate to species of interest in the much larger dataset.

The gene cluster data obtained in the queries can be downloaded. Depending on the type of search, different file formats are available. For gene cluster searches, the result table can be downloaded in tabular (CSV) format, alternatively it is possible to retrieve the DNA sequence of all matching clusters in FASTA format. Gene and protein domain searches offer a download the protein and nucleotide sequences of all matching genes or protein domains, respectively, or a tabular representation of the results. New options are provided to download specific chunks of the result data (for example only the first 1000 sequences) and to select between standard FASTA headers including the IDs and descriptive headers also including the query the hits were obtained with.

The selection of genomes available from NCBI still skews the perspective on the available diversity of biosynthetic gene clusters. While the antiSMASH database contains sequences from 33 different phyla, sequences from e.g. pro-
teobacteria are vastly overrepresented due to their significance as pathogens. The database now contains 32,548 biosynthetic gene clusters from the full genome dataset, an increase of 46% from version 1 (Table 1). Statistics from the 18,576 draft genomes certainly overpredict the number of identified clusters due to clusters being split over several contigs and counted multiple times, the fast-mode results still provide a good first estimate of the available biosynthetic diversity of the draft genomes. Of the 119,558 BGCs predicted on the draft genomes, over a third (41,482) are in contact with at least one contig edge and thus likely incomplete. In comparison, only ~1% of the clusters from the full genome dataset (390 in 32,548) are located on a contig edge. As the abundant fragmentation of clusters in draft genomes is skewing the numbers, the following statistics only count the results from the full genomes. See Table 2 for detailed cluster counts by BGC type and a comparison with the cluster counts from version 1.

In order to get an accurate taxonomic overview, the identified BGCs were mapped to a phylogenetic tree displaying approximately half of the genomes (12,219 complete and draft) that are included in the database (Figure 1A). The topology of the tree shows the microbial diversity chosen, ranging from well-characterized phyla to unclassified bacteria found in diverse ecosystems. Proteobacteria, Actinobacteria, Bacteroidetes, Firmicutes, Spirochaetes, Tenericutes, Cyanobacteria and Deinococcus-Thermus, the eight most abundant bacterial divisions in our database, accounting for 97.6% of genomes and all in the number of harbored BGCs (Figure 1B). High BGC numbers are characteristic for some groups of bacteria such as Actinobacteria (containing 13 clusters on average (full genomes) while others rarely possess one, like Tenericutes. These bacteria exhibit different distributions in terms of encoded secondary metabolite types as defined by antiSMASH (Figure 1C). For these statistics, the 45 BGC classes in antiSMASH have been condensed into five major groups: Non-Ribosomal Peptide Synthetase (NRPS), Polyketide, Ribosomally synthesized and post-translationally modified peptides (RiPP), terpenes and Others, clusters that do not belong to any of the aforementioned types. Terpenes, bacteriocins (a type of RiPP) and NRPS are the most common BGC types, all with higher number of representatives in the phylum Proteobacteria.

CONCLUSIONS

Genome mining is a valuable method to assess the biosynthetic potential of microorganisms. Since 2011, antiSMASH has assisted researchers with their secondary metabolite genome mining projects. The public web service has processed ~400,000 jobs, and the standalone tool has been downloaded over 10,000 times. The antiSMASH database both allows instant access to antiSMASH results for many publicly available genomes instead of waiting several hours for a de-novo antiSMASH run and allows advanced cross-genome searches for BGCs with specific features of interest.

In comparison to version 1, the updated version 2 of the antiSMASH database provides antiSMASH 4.2.1 annotations for 6,200 full genomes, which is an increase by 58%, and newly introduces data for 18,576 draft genomes. The graphical query builder allows researchers to interactively formulate searches to answer cross-genome research questions, while the results are presented in the familiar antiSMASH output format.

DATA AVAILABILITY

The antiSMASH database is available at https://antismashdb.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.

FUNDING

Novo Nordisk Foundation [NNF10CC1016517 to S.Y.L., NNF16OC0021746 to T.W.]; Netherlands Organization for Scientific Research [Veni grant number 863.15.002 to M.H.M.]; Warwick Integrative Synthetic Biology Centre, a UK Synthetic Biology Research grant from the BBSRC and EPSRC [BB/M017982/1 to E.L.C.d.L.S.]; Funding for open access charge: DTU/Novo Nordisk Foundation.

Conflict of interest statement. None declared.

REFERENCES


