Development of Bioinformatics Tools to Facilitate Genome Mining for Natural Products

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Diplom-Informatiker Kai Kristof Blin
aus Heidelberg

Tübingen
2013

Tag der mündlichen Qualifikation:	13.12.2013
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Wolfgang Wohlleben
2. Berichterstatter:	Prof. Dr. Harald Groß

Erklärung

Hiermit erkläre ich, dass ich diese Schrift selbständig und nur mit den angegebenen Hilfsmitteln verfasst habe. Ferner sind alle Stellen, die im Wortlaut oder dem Sinn nach den Werken anderer Autoren entnommen sind, durch Angabe der Quellen kenntlich gemacht. Eine detaillierte Abgrenzung meiner eigenen Leistungen von den Beiträgen meiner Kooperationspartner habe ich im Anhang vorgenommen.

. . . . . .

Unterschrift

Tübingen, den 24.11.2013

Abstract

The microbial secondary metabolism is a rich source of products with antibacterial, antifungal, anticancer and immunosuppressant activities that have found their way into clinical applications. After a golden area of natural product discovery lasting until the late 1970s, the rate of new discoveries turning into drugs suitable for clinical use has dropped and many pharmaceutical companies have abandoned natural product research.

The late 1990s saw the rise of a new technological advance that laid the foundation to a revival of natural products research using improved and more directed methodologies. High-throughput full genome sequencing allows to identify the potential biosynthetic capabilities of a producer organism by genome mining for natural products. The central part of this new approach is to not only identify clusters but also to predict the product of the biosynthesis.

antiSMASH, a software pipeline to predict a large number of different secondary metabolite gene clusters from genomic data was designed and implemented. antiSMASH is a tool that takes genomic DNA input sequences and generates an interactive HTML output page containing the predictions for 24 different secondary metabolite classes. Predictions include the polyketide backbone structure for polyketide synthase (PKS) products, the polypeptide structure for nonribosomal peptitde synthetase (NRPS) products and the molecular mass and post-translational modification to lanthipeptide core peptides. A public web service for running antiSMASH is available under `http://antismash.secondarymetabolites.org/`. Alternatively, antiSMASH can also be downloaded and run locally. In order to provide high-quality analyses for the antiSMASH pipeline, a machine-learning based prediction algorithm capable of predicting the composition of NRPS products was updated and improved in predictive power. Additionally, a novel prediction algorithm for the products of lanthipeptide synthases was developed and integrated into the antiSMASH prediction pipeline.

After a proof-of-concept implementation, a large-scale refactoring project was undertaken to ensure that good software engineering practice was observed in the antiSMASH code base. The refactoring ensures the long-term sustainability, stability and accessibility of the antiSMASH codebase.

Zusammenfassung

Der mikrobielle Sekundärmetabolismus ist eine reichhaltige Quelle von Produkten mit antibakterieller, antimykotischer und immunsuppressiver Wirkung, von denen viele den Weg zur klinischen Anwendung gefunden haben. Nach einigen erfolgreichen Jahrzehnten der Naturstoffforschung nimmt die Rate der klinisch relevanten Neuentdeckungen seit den späten 1970ern stetig ab, viele Pharmakonzerne haben die Naturstoffforschung gänzlich aufgegeben.
In den späten 1990ern wurden einige technologische Fortschritte gemacht, die den Grundstein legten für eine Renaissance der Naturstoffforschung mit verbesserten und zielgerichteteren Methoden. Hochdurchsatzsequenzierungen von Gesamtgenomen erlauben es die möglichen biosynthetischen Fähigkeiten von Poduzentenorganismen mittels genombasierter Naturstoffsuche abzuschätzen. Das Hauptziel der genombasierten Naturstoffsuche ist es, Gencluster nicht nur zu identifizieren sondern auch die biosynthetischen Produkte vorherzusagen.
antiSMASH, eine Softwarepipeline zur Vorhersage einer großen Zahl an unterschiedlichen Sekundärmetabolit-Genclustern aus Genomdaten, wurde entworfen und umgesetzt. antiSMASH ist ein Werkzeug das genomische DNS-Sequenzen als Eingabe akzeptiert und daraus eine interaktive HTML-Seite mit den Vorhersagen über 24 unterschiedliche Sekundärmetabolitklassen generiert. Zu den Vorhersagen gehören das Poliketid-Rückgrat von PKS-Produkten, das Polypeptid-Rückgrat von NRPS-Produkten und die molare Masse und posttranslationale Veränderungen an Lanthipeptid-Kernpeptiden. Ein öffentlich zugänglicher Webdienst mit antiSMASH ist unter `http://antismash.secondarymetabolites.org/` verfügbar. Alternativ kann antiSMASH auch heruntergeladen und lokal ausgeführt werden. Um für die Pipeline qualitativ hochwertige Vorhersagen treffen zu können, wurde ein auf maschinellem Lernen basierender Algorithmus für die Vorhersage vo NRPS-Produkten aktualisiert und in seiner Vorhersageleistung verbessert. Zusätzlich wurde ein neuartiger Algorithmus für die Vorhersage von Produkten der Lanthipeptid-Synthetasen etwickelt und in die antiSMASH-Pipeline integriert.
Nach einer initialen Machbarkeitsstudie wurde ein großangelegtes Refactoring unternommen, um sicherzustellen dass die Prinzipien guter Softwareingenieurpraxis im Quelltext von antiSMASH beachtet wurden. Dieses Refactoring stellt die langfristige Zukunftstauglichkeit, Stabilität und Benutzbarkeit von antiSMASH sicher.

Acknowledgements

First and foremost, I want to express my deepest gratitude to Prof. Dr. Wolfgang Wohlleben. Thank you for giving me the opportunity to play my tiny part in the great struggle against resistant super-bugs, for your inspiration and support.

I am indebted to Prof. Dr. Harald Groß for his unhesitating offer to be my second referee on a really short notice. Thank you for your advice and your support when I decided to not heed all of it.

My heartfelt thanks go to Dr. Tilmann Weber. Thank you for being a friend as well as a boss. Thank you for all the kind words, the encouragement and the long nights we spent on getting papers finished. Thank you for coming all the way down from Copenhagen to be on my thesis defence committee.

I'm greatful to Dr. Kay Nieselt for offering to be the fourth member of my thesis defence committee. Thank you for frequently inviting an exiled bioinformatician back to the Sand for all those bioinformatics events.

Special thanks to Dr. Marnix Medema for all great work we got done in antiSMASH, especially for all the long nights we spent struggling to get antiSMASH 1 up and running in time for the paper deadline. It was great crunching all that code with you, but let's never do it in that short a timeframe again. Working on antiSMASH 2 was so much more relaxed, thank you for that as well.

I would also like to thank my great co-workers. Firstly, the Secondary Metabolite Genomics group. Ewa, Sabrina, Demi, Denise, Thomes, Thomas and Andi, thank you for giving me the opportunity to "play" in the lab, for your patience with all my questions when getting started and for never hiding the pipettes well enough. Secondly, the "Balhimycinis", Evi, Robi, Melanie, Siegrid, Valentina, HaJö and Marius, thanks for adopting me into the group with the best food in the department. HaJö, I so understand why you were cranky in your final weeks in the lab right now. Melli, thanks for being a worthy opponent, very much appreciated. Special thanks go to the "Mikrobio Stammtisch" group. Annika, thanks for herding us to a pub regularly, we never seemed to really make this work without you. Last but not least, thanks to all the other great people in the department. I really, really enjoyed my time with you, and I was always greatful to see you turn up at my theatre performances. Annika, I know that was usually your fault, thanks again for that. Günther, I'm afraid you'll have to recruit new people for the espresso break, I'll miss the opportunity to talk about sci-

ence, football or both.
Thanks to all of my friends, those on stage and backstage, and those not connected to the theatre at all. Thank you Oliver for patiently listening to my rants about academia in the long nights hanging lights, and thank you for casting me in some of the most fun parts I've played on stage. You've made theatre into the big part of my life that it is now. Let's get that CNC machine finished, now that this dissertation is over! Thanks to Julia, Maresa, Andrea, Tatjana, Christoph, Burkhard, Max and Tobias for being such an awesome set of friends.
Last, but certainly not least, thank you to my family, especially my parents. Thanks for your unerring support no matter what path I decided to try for the journey of my life, and thanks for your advice whenever I got stuck at some of the crossroads. Thank you Regi for walking this journey with me for all these years, I couldn't have done this without you.

Contents

Acronyms

A	adenylation
AA	amino acid
ACP	acyl carrier protein
API	application programming interface
AT	acyl transferase
AviCys	*S*-[(*Z*)-2-aminovinyl]-D-cysteine
AviMeCys	*S*-[(*Z*)-2-aminovinyl]-3-methyl-D-cysteine
C	condensation
COG	cluster of orthologous groups
Cy	heterocyclisation
Cys	cysteine
DH	dehydratase
Dha	dehydroxyalanine
Dhb	dehydroxybutyrine
E	epimerisation
ER	enoyl reductase
KR	keto reductase
KS	keto synthase
Lan	lanthionine
MeLan	methyl-lanthionine
MSA	multiple sequence alignment

NGS	next-generation sequencing
NRPS	nonribosomal peptitde synthetase
ORF	open reading frame
PCP	peptidyl carrier protein
PCR	polymerase chain reaction
PDB	Protein Data Base
pHMM	profile Hidden Markov Model
PKS	polyketide synthase
RiPP	ribosomally synthetised and post-translationally modified peptide
Ser	serine
smCOG	secondary metabolite cluster of orthologous groups
SVM	support vector machine
TE	thioesterase
Thr	threonine
TSVM	transductive support vector machine

1 Publications

1.1 Published Manuscripts

1.1.1 Medema et al. (2011) antiSMASH

Published online 14 June 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W339–W346***
doi:10.1093/nar/gkr466

antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences

Marnix H. Medema[1,2], Kai Blin[3], Peter Cimermancic[4], Victor de Jager[5,6,7], Piotr Zakrzewski[1,2], Michael A. Fischbach[4], Tilmann Weber[3], Eriko Takano[1,*] and Rainer Breitling[2,8]

[1]Department of Microbial Physiology, [2]Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747AG Groningen, The Netherlands, [3]Mikrobiologie/Biotechnologie, Interfakultäres Institut für Mikrobiologie und Infektionsmedizin, Eberhard Karls Universität Tübingen, Auf der Morgenstelle 28, 72076 Tübingen, Germany, [4]Department of Bioengineering and Therapeutic Sciences and California Institute for Quantitative Biosciences, University of California San Francisco, 1700 4th Street, San Francisco CA 94158, USA, [5]Laboratory of Microbiology, Wageningen University, 6703HB Wageningen, [6]Netherlands Bioinformatics Centre and [7]Centre for Molecular and Biomolecular Informatics, Nijmegen Centre for Molecular Life Sciences, Radboud University Nijmegen Medical Centre, 6500HB Nijmegen, The Netherlands and [8]Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, G12 8QQ, Glasgow, UK

Received February 28, 2011; Revised May 9, 2011; Accepted May 21, 2011

ABSTRACT

Bacterial and fungal secondary metabolism is a rich source of novel bioactive compounds with potential pharmaceutical applications as antibiotics, anti-tumor drugs or cholesterol-lowering drugs. To find new drug candidates, microbiologists are increasingly relying on sequencing genomes of a wide variety of microbes. However, rapidly and reliably pinpointing all the potential gene clusters for secondary metabolites in dozens of newly sequenced genomes has been extremely challenging, due to their biochemical heterogeneity, the presence of unknown enzymes and the dispersed nature of the necessary specialized bioinformatics tools and resources. Here, we present antiSMASH (antibiotics & Secondary Metabolite Analysis Shell), the first comprehensive pipeline capable of identifying biosynthetic loci covering the whole range of known secondary metabolite compound classes (polyketides, non-ribosomal peptides, terpenes, aminoglycosides, aminocoumarins, indolocarbazoles, lantibiotics, bacteriocins, nucleosides, beta-lactams, butyrolactones, siderophores, melanins and others). It aligns the identified regions at the gene cluster level to their nearest relatives from a database containing all other known gene clusters, and integrates or cross-links all previously available secondary-metabolite specific gene analysis methods in one interactive view. antiSMASH is available at http://antismash.secondarymetabolites.org.

INTRODUCTION

Microbial secondary metabolites offer great potential for the development of new medicines. They belong to a wide variety of chemical classes, and many of them have cholesterol-lowering, anti-tumor or antibiotic activities. The rapid decrease in the cost of genome sequencing now allows the discovery of hundreds or even thousands of gene clusters encoding the biosynthetic machinery for these compounds (1). However, laboratory research cannot keep pace with the speed of genomic discovery, as the experimental characterization of each gene cluster is still very laborious. Therefore, effective *in silico* identification of the most promising targets within genomes is essential for the successful mining of the genomic riches available. Manual annotation is very labor-intensive and time-consuming, leading to incomplete annotations. Automatic annotation of secondary metabolite clusters may enhance accuracy as well as completeness of the annotation. A few *in silico* methods have been published thus far to automate the analysis of secondary metabolism in

*To whom correspondence should be addressed. Tel: +31503632143; Fax: +31503632154; Email: e.takano@rug.nl
Correspondence may also be addressed to Rainer Breitling. Tel: +441413307374; Email: rainer.breitling@glasgow.ac.uk

bacterial genomes. The first of these was ClustScan (2), which allows the uploading of genomic data to a server for the semi-automatic detection and annotation of polyketide synthase (PKS) and non-ribosomal peptide synthetase (NRPS) gene clusters. Additionally, Anand *et al.* (3) recently published the SBSPKS toolbox for structure-based PKS analysis. Li *et al.* (4) constructed the NP.searcher web server, which is specialized in predicting the possible chemical structures resulting from a subset of gene cluster types. Unfortunately, all these tools are largely limited to the analysis of the core genes for type I polyketide (PK) and non-ribosomal peptide (NRP) biosynthesis. Thus far, accessory genes as well as core genes for many other secondary metabolite scaffolds have largely been neglected in computational approaches, even though some very good but also very specific tools are available for bacteriocin (5) and type III PKS (6) detection. For fungal genomes, the SMURF tool (7) has recently become available, which is capable of generating a somewhat more comprehensive list of secondary metabolite biosynthesis gene clusters, but this tool offers little further detailed analysis. CLUSEAN (8) currently offers the most comprehensive analysis by including a full genome annotation, but it is difficult to operate for the non-specialist and requires intensive manual analysis of the output.

Here, we present a software pipeline for secondary metabolite gene cluster identification, annotation and analysis which is comprehensive, rapid and user-friendly (Figure 1). It can be run either from a web server (http://antismash.secondarymetabolites.org/) or as a stand-alone version on a standard desktop computer. It can rapidly detect all known classes of secondary metabolite biosynthesis gene clusters, provide detailed NRPS/PKS functional annotation, and predict the chemical structure of NRPS/PKS products with higher accuracy than existing methods. Additionally, by constructing a database of all currently known secondary metabolite biosynthesis gene clusters throughout the tree of life, we were able to equip the tool with a comparative gene cluster analysis module. In this module, evolutionary similarities between a queried gene cluster and other gene clusters are detected and visualized in order to be able to rapidly infer functions of genes and operons based on homology. Finally, from the genes within this database of gene clusters, we constructed secondary metabolism Clusters of Orthologous Groups (smCOGs). These are used in yet another module to predict and categorize the functions of accessory genes, and to calculate phylogenetic trees for each gene with a seed alignment of its smCOG protein family. Our benchmark results show that our method reliably detects gene clusters of a wide variety of biosynthetic types, and that it is able to significantly enhance manual genome annotations of secondary metabolite biosynthesis.

METHODS AND IMPLEMENTATION

File and options input

The input front end of the antiSMASH web server allows uploading of sequence files of a variety of types (FASTA, GBK, or EMBL files). Alternatively, a GenBank/RefSeq accession number can be provided, which is used by the web server to automatically obtain the associated file from GenBank. If the user chooses to use a FASTA input file, gene prediction is performed by Glimmer3 (9)—using its long-orfs tool to construct a gene model based on the input sequence itself—or by GlimmerHMM (10) when

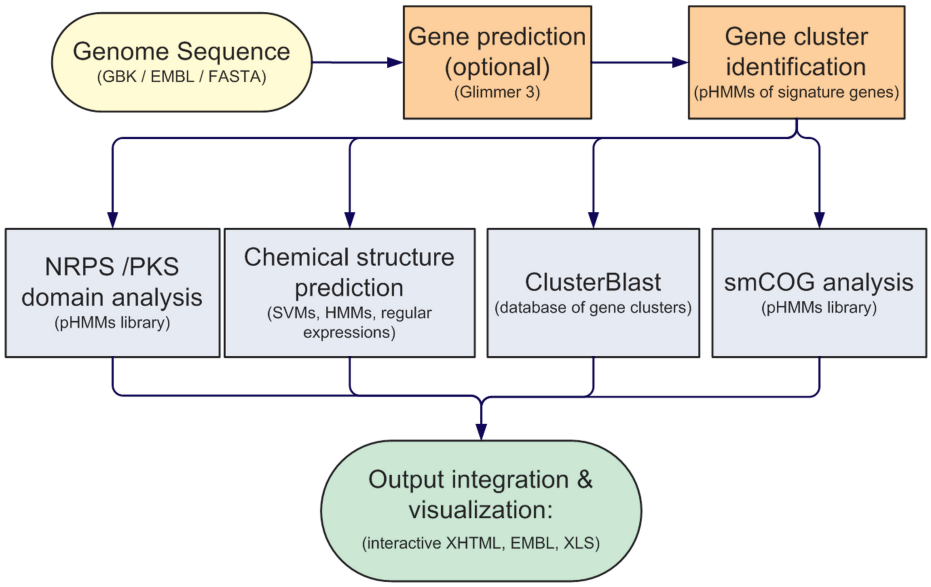


Figure 1. Outline of the pipeline for genomic analysis of secondary metabolites. Genes are extracted or predicted from the input nucleotide sequence, and gene clusters are identified with signature gene pHMMs. Subsequently, several downstream analyses can be performed: NRPS/PKS domain analysis and annotation, prediction of the core chemical structure of PKSs and NRPSs, ClusterBlast gene cluster comparative analysis, and smCOG secondary metabolism protein family analysis. The output is visualized in an interactive XHTML web page, and all details are stored in an EMBL file for additional analysis and editing in a genome browser. A Microsoft Excel file with an overview of all detected gene clusters and their details is also generated.

eukaryotic input data is submitted. Before starting the antiSMASH analysis run, the user can select the gene cluster types he or she wants to search for. Additionally, he can select which of the downstream analysis modules to include. For those users who, e.g. work with proprietary data, a stand-alone version with a Java graphical user interface is available with the same input options as the web version. Finally, expert users may choose to directly run the Python-based pipeline program from the command line in order to batch analyze a larger number of inputs.

Detection of secondary metabolite biosynthesis gene clusters

Using the HMMer3 tool (http://hmmer.janelia.org/), the amino acid sequence translations of all protein-encoding genes are searched with profile Hidden Markov Models (pHMMs) based on multiple sequence alignments of experimentally characterized signature proteins or protein domains (proteins, protein subtypes or protein domains which are each exclusively present in a certain type of biosynthetic gene clusters). Using both existing pHMMs (5,11–13) and new pHMMs from seed alignments, we constructed a library of models specific for type I, II and III PK, NRP, terpene, lantibiotic, bacteriocin, aminoglycoside/aminocyclitol, beta-lactam, aminocoumarin, indole, butyrolactone, ectoine, siderophore, phosphoglycolipid, melanin and aminoglycoside biosynthesis signature genes. Additionally, we constructed a number of pHMMs specific for false positives, such as the different types of fatty acid synthases which show homology to PKSs. The final detection stage operates a filtering logic of negative and positive pHMMs and their cut-offs. The logic is based on knowledge of the minimal core components of each gene cluster type taken from the scientific literature. The cut-offs were determined by manual studies of the pHMM results when run against the NCBI non-redundant (nr) protein sequence database (ftp://ftp.ncbi.nlm.nih.gov/blast/db). All technical details on the pHMM library and the detection rules are available in Supplementary Tables S1 and S2, respectively.

Gene clusters are defined by locating clusters of signature gene pHMM hits spaced within <10 kb mutual distance. To include flanking accessory genes, gene clusters are extended by 5, 10 or 20 kb on each side of the last signature gene pHMM hit, depending on the gene cluster type detected. As a consequence of this greedy methodology, gene clusters that are spaced very closely together may be merged into 'superclusters'. These gene clusters are indicated in the output as 'hybrid clusters'; they may either represent a single gene cluster which produces a hybrid compound that combines two or more chemical scaffold types, or they may represent two separate gene clusters which just happen to be spaced very closely together.

NRPS/PKS domain architecture analysis

NRPS/PKS domain architectures are analyzed (Figure 2) using another pHMM library comprising existing models (8,11–15) as well as newly constructed models specific for NRPS/PKS protein domains and functional/phylogenetic subgroups of these domains (Supplementary Table S3).

Conserved motifs within key PKS and NRPS domains are also detected using the pHMMs described earlier in the CLUSEAN package (8), and are written to the detailed downloadable EMBL output. PKS/NRPS gene names are annotated according to the domains and domain subtypes that the genes contain (e.g. 'hybrid NRPS-PKS', 'enediyne PKS', 'glycopeptide NRPS', 'trans-AT PKS', etc.).

Substrate specificity, stereochemistry and final structure predictions

Substrate specificity prediction of PKS and NRPS modules, based on the active sites of their respective acyltransferase (AT) and adenylation (A) domains, is performed by various available methods. PKS AT domain specificities are predicted using a 24 amino acid signature sequence of the active site (16), as well as with pHMMs based on the method of Minowa *et al.* (17), which is also used to predict co-enzyme A ligase domain specificities. NRPS A domain specificities are predicted using both the signature sequence method and the support-vector machines-based method of NRPSPredictor2 (18,19), and using the method of Minowa *et al.* (17). Finally, all predictions are integrated into a consensus prediction by a majority vote. Ketoreductase domain-based stereochemistry predictions for PKSs (2) are performed as well. An estimate of the biosynthetic order of PKS/NRPS modules is predicted based on PKS docking domain sequence residue matching [for type I modular PKSs, (3)] or assumed colinearity, and a final predicted core chemical structure is generated as a SMILES string (20), i.e. a unique text description of the chemical structure, and visualized in a picture file (Figure 2). To increase the reliability of the core structure prediction, monomers for which there was no consensus in the predictions are represented as generic amino acids or ketides with unspecified R-groups.

Secondary metabolite clusters of orthologous groups

In order to rapidly annotate the accessory genes surrounding the detected core signature genes in the various types of secondary metabolite biosynthesis gene clusters, we constructed a database of all gene clusters contained in the latest NCBI nt database (15 February 2011). To do so, pHMMs described above were used to detect all secondary metabolite biosynthesis gene cluster signature genes in the nr database. The accession numbers of all hits meeting the described cut-offs were extracted and used to download the corresponding GenPept files. If the taxonomy identifier included 'bacteria' or 'fungi', the nucleotide source accession number was extracted. The corresponding nucleotide GenBank files were then downloaded as well, and cross-checked for presence of the queried protein accession number. For each nucleotide GenBank file, gene clusters were detected as described above. Amino acid sequences of all genes contained within the gene clusters were written to a FASTA file with headers containing key information, and a summary of all detected gene clusters (nucleotide accession, nucleotide description, cluster number, cluster type, protein accession numbers)

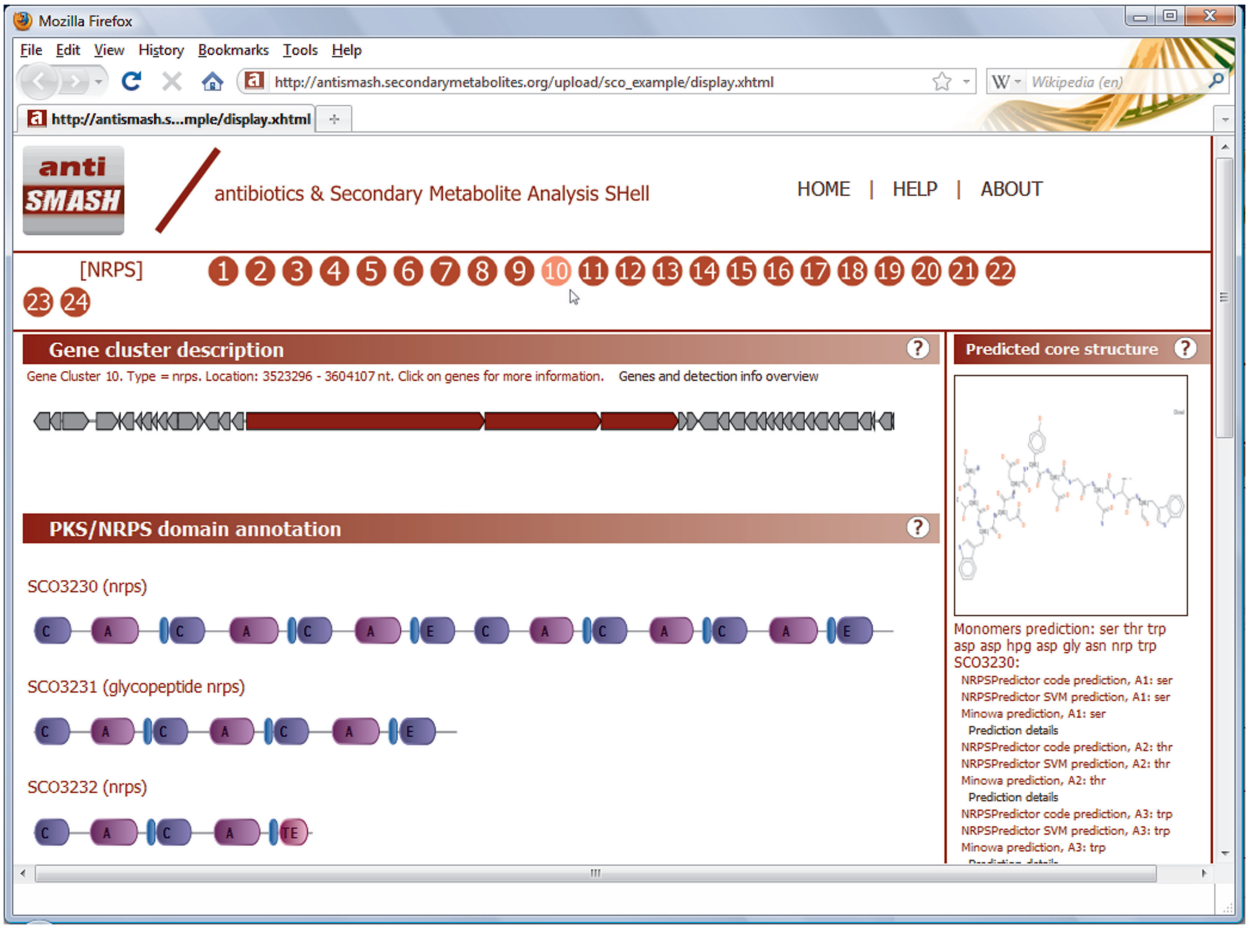


Figure 2. Interactive XHTML visualization of results. The numbers below the banner represent the gene clusters that were detected, the type of which is shown to the left of them at mouse-over. Once a gene cluster has been selected, the 'Gene cluster description' tab will display an SVG image with all genes within the approximate gene cluster, with the detected signature genes displayed in red. Locus tags appear on mouse-over, and on clicking a gene a small panel pops up with annotation information and cross-links to other web services. If PKS/NRPS proteins are encoded in the gene cluster, their domain annotations are given in the 'PKS/NRPS domain annotation' tab. More detailed domain annotation information and cross-links are provided on mouse-over. In the 'Predicted core structure' tab, a prediction of the core chemical structure is given for PKS or NRPS gene clusters based on the predictions displayed below it. All tabs contain a wide range of links to pop-ups which further detail the prediction information.

was written to a text file. To construct the smCOGs, clustering of all gene cluster proteins was performed using OrthoMCL (21), and consensus annotations were manually assigned based on the frequencies of the five most prevalent annotations of each smCOG in GenBank. For each smCOG, a seed alignment was created from 100 randomly picked sequences using MUSCLE 3.5 (22), and a pHMM of each smCOG was generated based on the conserved core of each alignment (Supplementary Figure S1). Within the antiSMASH software pipeline, the smCOG pHMMs are used for functional annotation of all accessory genes within the gene clusters. After assignment of an smCOG to a gene—based on the highest-scoring pHMM on its sequence above a certain *e*-value threshold—the predicted protein sequence is aligned to the smCOG seed alignment, and a rough neighbor-joining phylogenetic tree is calculated using FastTree 2 (23) and visualized with TreeGraph 2 (24) (Supplementary Figure S1).

ClusterBlast comparative gene cluster analysis

Secondary metabolite biosynthesis gene clusters are highly modular, and their genes are transferred frequently from one gene cluster to another during evolution (25,26). Therefore, when trying to obtain a functional understanding of a gene cluster, it is highly beneficial to be able to compare it with (parts of) other gene clusters which show similarity to it and which may have been characterized experimentally. In order to facilitate this, we applied our annotated database of gene clusters to link up protein sequences with their parent gene clusters and create a comparison tool—based on the most recent BLAST$^+$ implementation (27)—which ranks gene clusters by similarity to a queried gene cluster. Clusters are sorted first based on an empirical similarity score $S = h + H + s + S + B$, in which h is the number of query genes with a significant hit, H is the number of core query genes with a significant hit, s is the number of gene pairs with

conserved synteny, S is the number of gene pairs with conserved synteny involving a core gene, and B is a core gene bonus (three points given when at least one core gene has a hit in the subject cluster). If the similarity scores are equal, the hits are subsequently ranked based on the cumulative BlastP bit scores between the gene clusters. This feature enables a rapid assessment of the comparative genomics for each annotated cluster (Figure 3).

Genome-wide BLAST and Pfam analysis and prediction of potential unknown secondary metabolite biosynthesis gene cluster types

To facilitate further thorough manual genome analysis, antiSMASH has also been linked up to the whole-genome BLAST and Pfam analysis modules from the previously published CLUSEAN framework (8). The CLUSEAN results are integrated into an EMBL output file. Furthermore, as unknown biosynthetic gene cluster types are likely to exist which may be missed by the antiSMASH gene cluster detection module, the Pfam results are also used to predict genomic regions with a high probability of constituting secondary metabolite biosynthesis gene clusters in a more generalized fashion than the signature genes pHMMs method. For this, the genome sequence is converted to a string of predicted Pfam domains which is fed to a hidden Markov model (P. Cimermancic *et al.*, manuscript in preparation) with transitions between a gene cluster state and a rest-of-the-genome state. This model was trained on Pfam domain frequencies from a set of 473 cloned gene clusters (gene cluster state) and from the set of ~1100 genomes currently in the JGI IMG database (rest-of-the-genome state). The result of this analysis is visualized in a PNG graph.

Output and visualization

All pipeline analysis results are visualized in a user-friendly interactive XHTML page (Figure 2), which can be used to browse through the different gene clusters. For PKS and NRPS gene clusters, the predicted core chemical structures are shown as images. Gene cluster maps are drawn with scalable vector graphics (SVGs), to which interactive on-click and mouse-over functions are added through JavaScript to provide annotation information,

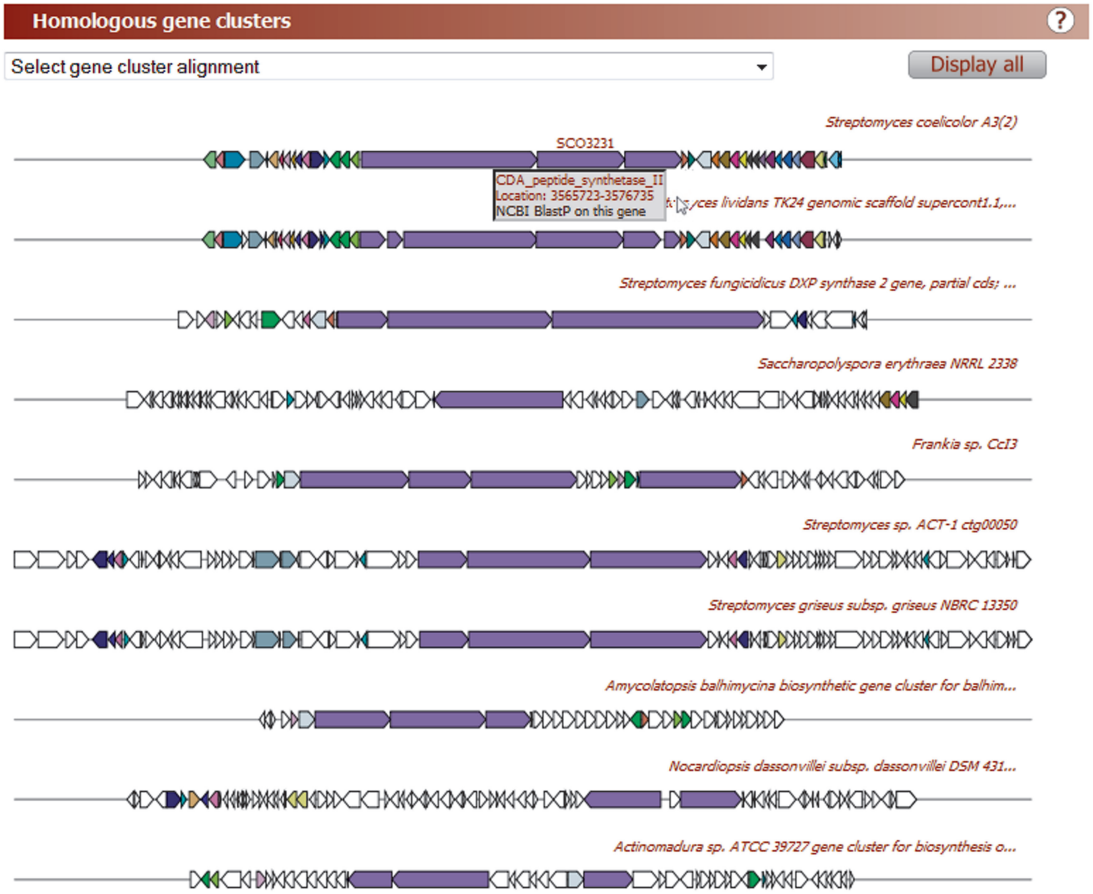


Figure 3. Example of ClusterBlast alignment of gene clusters homologous to the query gene cluster. In this case, the ten best hits to the calcium-dependent antibiotic NRPS gene cluster from *Streptomyces coelicolor* A3(2) are displayed. Homologous genes (BLAST *e*-value < 1E-05; 30% minimal sequence identity; shortest BLAST alignment covers over >25% of the sequence) are given the same colors. The 'select gene cluster alignment' drop-down menu provides links to one-by-one gene cluster alignments to each gene cluster hit. In the one-by-one gene cluster alignments, PubMed and/or PubChem links are provided for gene clusters associated with a known compound.

pipeline result scores, and BLAST hyperlinks. Detected signature genes on which the gene cluster identification is based are shown in a distinct color. ClusterBlast results are displayed in a similar way, as aligned gene cluster maps in which genes with mutual BLAST hits are given identical colors. Additionally, available at the bottom right of the page, fully annotated EMBL output files provide the user with the additional possibility to browse their genome in a genome browser such as Artemis (28).

RESULTS

Compared to previous software, the pipeline described here is uniquely comprehensive: it integrates all previously published analysis types into one tool and adds valuable novel functionalities (Table 1).

In order to measure the accuracy of the gene cluster predictions, we performed two independent benchmark evaluations of the method. First, we collected the sequences of cloned gene clusters of known compounds of biosynthetic types by searching both the GenBank/RefSeq databases and the scientific literature with a range of different keywords. From the resulting set of 484 cloned gene cluster GenBank files, 473 (97.7%) were correctly identified by antiSMASH, and 468 (96.7%) were given exactly the same annotation by antiSMASH as by the articles describing their experimental characterization (Figure 4 and Supplementary Table S4). In order to test for false positives as well, we also benchmarked the method on five well-annotated genomes from different taxonomic groups. Besides genomes of three different actinomycetes (the organisms on which the tool is likely to be used most often) these included a Proteobacterium (*Pseudomonas fluorescens* Pf-5) and a fungus (*Aspergillus fumigatus* Af293). In the five genomes, 97.3% of all 111 annotated gene clusters were detected by antiSMASH (Figure 5 and Supplementary Table S5). Under closer scrutiny, two of the three gene clusters that were missed by antiSMASH appeared to lack a complete set of genes associated with biosynthesis of a known chemical scaffold. More interestingly, 35 additional gene clusters were detected (31.5%) which had been missed during initial genome annotation and which after close inspection all appeared to have a high probability of being actual biosynthetic gene clusters. The cluster types that appeared to be frequently missed during the annotation of these genomes appeared to be butyrolactones (eight gene clusters missed), terpenes (seven gene clusters missed), NRPSs/PKSs (six gene clusters missed) and lantibiotics (five gene clusters missed), which suggests that the computational approach used can yield improvements even in finding gene clusters of common biosynthetic types.

We also compared the performance of antiSMASH with other existing tools. No similarly comprehensive tools are available, but NP.searcher and SMURF each offer automated gene cluster detection for a small subset of the cluster types detected by antiSMASH (NP.searcher detects bacterial NRPS/PKS gene clusters, and SMURF detects fungal NRPS, PKS, and dimethylallyl tryptophan synthase gene clusters). Our analysis of the results of these tools on four bacterial and two fungal genomes (Supplementary Table S6), respectively, showed that antiSMASH and SMURF performed equally well (both detect 74 gene clusters, with 93.4% overlap). Compared to NP.searcher, antiSMASH detected significantly more (47 versus 31, i.e. 51.6% more) NRPS/PKS gene clusters, while all NP.searcher-detected gene clusters were also picked up by antiSMASH. The gene clusters that were detected by antiSMASH but not by NP.searcher were all small NRPS-like or PKS-like gene clusters. None of the three tools gave predictions that were clear false positives, except one SMURF detection of a probable fatty acid synthase (GenBank ID CAP98191.1) labeled as PKS.

DISCUSSION AND CONCLUSIONS

antiSMASH not only provides a unique integration of previously widely dispersed tools, but it also achieves very high accuracy in its individual cluster annotations, which are enhanced by unique novel analyses such as BLAST-based gene cluster alignments and secondary metabolite COG phylogenetic trees for accessory genes. As the field of synthetic biology is opening up new ways to study these gene clusters in a high-throughput fashion (29), antiSMASH will enable experimental researchers to quickly pinpoint those gene clusters most interesting for further study, and swiftly collect secondary metabolite

Table 1. Comparison of different software tools for secondary metabolite biosynthesis analysis

Software	Open-source & stand-alone available	Covers full tree of life	NRPS/PKS detection	NRPS/PKS detailed functional domain annotation	NRP/PK core structure prediction	Detection of other biosynthetic classes	Gene cluster border prediction	Comparative gene cluster analysis	Prediction of all secondary metabolite-like genomic regions
ClustScan		+	+	+	+	±			
CLUSEAN	+		+	+					
NP.searcher	+	+	+		+				
SBSPKS		+	+	+					
SMURF			+			±	+		
antiSMASH	+	+	+	+	+	+	+	+	+

Comparison of functionalities of currently existing programs or software packages for secondary metabolite biosynthesis analysis.

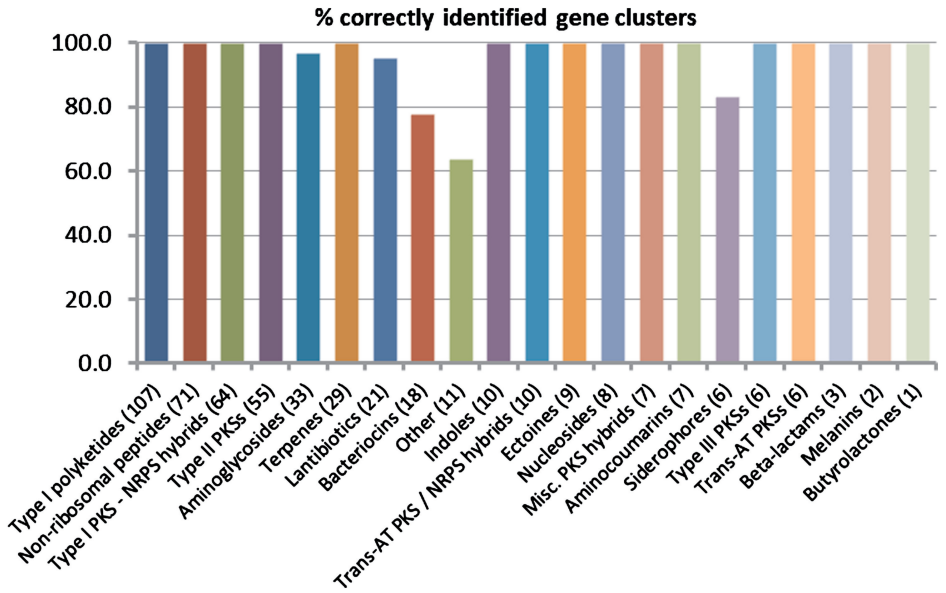


Figure 4. Benchmark results on a set of 473 cloned secondary metabolite biosynthesis gene clusters found in the GenBank nucleotide database. The numbers behind the names of the biosynthetic types indicate how many gene clusters of that type were in the benchmark set.

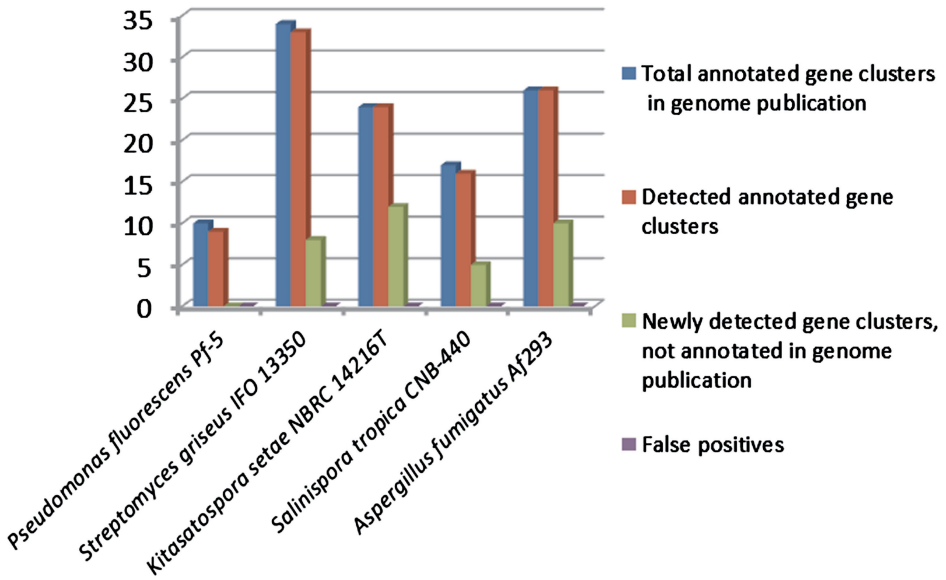


Figure 5. Benchmark results on five genome sequences. All except three annotated gene clusters from the five genome publications were detected; two of these annotated gene clusters (SGR5285-SGR5295 in *Streptomyces griseus* and Strop_3244-Strop_3253 in *Salinispora tropica*) appeared to lack core genes for biosynthesis of a known secondary metabolite chemical scaffold. The one certain gene cluster which was not detected was a small gene cluster for the biosynthesis of hydrogen cyanide from *Pseudomonas fluorescens* Pf-5.

BioBricks for the (re-)design of gene clusters. Moreover, the new comparative analyses that antiSMASH offers provide unprecedented possibilities to interpret the functions of both complete gene clusters and their particular genes in their evolutionary context. The approaches developed are likely to soon allow global analysis of all small molecule biosynthesis gene clusters throughout the tree of life, so that we can acquire a more and more comprehensive understanding of how nature itself designs novel bioactive compounds.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Mike Li for kindly providing a script for the conversion of strings of amino acid and polyketide residues into SMILES strings. The authors

thank Marc Röttig and Oliver Kohlbacher for providing NRPSpredictor2.

FUNDING

The Dutch Technology Foundation STW, which is the applied science division of NWO and the Technology Programme of the Ministry of Economic Affairs (STW 10463); GenBioCom program of the German Ministry of Education and Research (BMBF) (grant 0315585A); Rosalind Franklin Fellowship, University of Groningen (to E.T.); NWO-Vidi Fellowship (to R.B.); NIH DP2 Award (OD007290) (to M.A.F.); Travel grant from the Boehringer Ingelheim Fonds (to M.H.M.). Funding for open access charge: STW (STW 10463).

Conflict of interest statement. None declared.

REFERENCES

1. Walsh,C.T. and Fischbach,M.A. (2010) Natural products version 2.0: connecting genes to molecules. *J. Am. Chem. Soc.*, **132**, 2469–2493.
2. Starcevic,A., Zucko,J., Simunkovic,J., Long,P.F., Cullum,J. and Hranueli,D. (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and *in silico* prediction of novel chemical structures. *Nucleic Acids Res.*, **36**, 6882–6892.
3. Anand,S., Prasad,M.V., Yadav,G., Kumar,N., Shehara,J., Ansari,M.Z. and Mohanty,D. (2010) SBSPKS: Structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.*, **38**, W487–W496.
4. Li,M.H., Ung,P.M., Zajkowski,J., Garneau-Tsodikova,S. and Sherman,D.H. (2009) Automated genome mining for natural products. *BMC Bioinformatics*, **10**, 185.
5. de Jong,A., van Heel,A.J., Kok,J. and Kuipers,O.P. BAGEL2: Mining for bacteriocins in genomic data. *Nucleic Acids Res.*, **38**, W647–W651.
6. Mallika,V., Sivakumar,K.C., Jaichand,S. and Soniya,E.V. (2010) Kernel based machine learning algorithm for the efficient prediction of type III polyketide synthase family of proteins. *J. Integr. Bioinform*, **7**, 143.
7. Khaldi,N., Seifuddin,F.T., Turner,G., Haft,D., Nierman,W.C., Wolfe,K.H. and Fedorova,N.D. (2010) SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.*, **47**, 736–741.
8. Weber,T., Rausch,C., Lopez,P., Hoof,I., Gaykova,V., Huson,D.H. and Wohlleben,W. (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J. Biotechnol.*, **140**, 13–17.
9. Delcher,A.L., Bratke,K.A., Powers,E.C. and Salzberg,S.L. (2007) Identifying bacterial genes and endosymbiont DNA with glimmer. *Bioinformatics*, **23**, 673–679.
10. Majoros,W.H., Pertea,M. and Salzberg,S.L. (2004) TigrScan and GlimmerHMM: two open source *ab initio* eukaryotic gene-finders. *Bioinformatics*, **20**, 2878–2879.
11. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
12. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: Recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
13. Yadav,G., Gokhale,R.S. and Mohanty,D. (2009) Towards prediction of metabolic products of polyketide synthases: An *in silico* analysis. *PLoS Comput. Biol.*, **5**, e1000351.
14. Ansari,M.Z., Sharma,J., Gokhale,R.S. and Mohanty,D. (2008) In silico analysis of methyltransferase domains involved in biosynthesis of secondary metabolites. *BMC Bioinformatics*, **9**, 454.
15. Rausch,C., Hoof,I., Weber,T., Wohlleben,W. and Huson,D.H. (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol. Biol.*, **7**, 78.
16. Yadav,G., Gokhale,R.S. and Mohanty,D. (2003) Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *J. Mol. Biol.*, **328**, 335–363.
17. 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.
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.*, doi: 10.1093/nar/gkr323.
19. Rausch,C., Weber,T., Kohlbacher,O., Wohlleben,W. and Huson,D.H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.*, **33**, 5799–5808.
20. Weininger,D. (1988) SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, **28**, 31–36.
21. Li,L., Stoeckert,C.J. Jr and Roos,D.S. (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**, 2178–2189.
22. Edgar,R.C. (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
23. Price,M.N., Dehal,P.S. and Arkin,A.P. (2010) FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**, e9490.
24. Stover,B.C. and Muller,K.F. (2010) TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics*, **11**, 7.
25. Fischbach,M.A., Walsh,C.T. and Clardy,J. (2008) The evolution of gene collectives: how natural selection drives chemical innovation. *Proc. Natl Acad. Sci. USA*, **105**, 4601–4608.
26. Donadio,S., Sosio,M., Stegmann,E., Weber,T. and Wohlleben,W. (2005) Comparative analysis and insights into the evolution of gene clusters for glycopeptide antibiotic biosynthesis. *Mol. Genet. Genomics*, **274**, 40–50.
27. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: Architecture and applications. *BMC Bioinformatics*, **10**, 421.
28. Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.A. and Barrell,B. (2000) Artemis: Sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
29. Medema,M.H., Breitling,R., Bovenberg,R. and Takano,E. (2011) Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nat. Rev. Microbiol.*, **9**, 131–137.

1.1.2 Blin et al. (2013b) antiSMASH2

Nucleic Acids Research Advance Access published June 3, 2013

Nucleic Acids Research, 2013, 1–9
doi:10.1093/nar/gkt449

antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers

Kai Blin[1], Marnix H. Medema[2,3], Daniyal Kazempour[1], Michael A. Fischbach[4], Rainer Breitling[3,5,*], Eriko Takano[2,5,*] and Tilmann Weber[1,*]

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

*To whom correspondence should be addressed. Tel: +49 7071 29 78841; Fax: +49 7071 29 5979; Email: tilmann.weber@biotech.uni-tuebingen.de
Correspondence may also be address to Rainer Breitling. Tel: +44 1613 065117; Email: rainer.breitling@manchester.ac.uk
Correspondence may also be address to Eriko Takano. Tel: +44 1613 064419; Email: eriko.takano@manchester.ac.uk

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

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)

Table 1. Overview of the capabilities of various software tools for the analysis of biosynthetic gene clusters

Features	antiSMASH 2.0	antiSMASH 1.0	CLUSEAN	SMURF	ClustScan	NaPDoS	NP.searcher	NRPSpredictor2	NRPSSP	SBSPKS
Open-source and stand-alone available	X	X	X				X	X		X
Covers bacteria, archaea and fungi	X	X			X		X	X	X	X
NRPS/PKS detection	X	X	X	X	X	X	X	X	X	X
NRPS/PKS detailed functional domain annotation	X	X	X		X			X		X
NRP/PK core structure prediction	X	X			X		X		X	
Lantipeptide core structure prediction	X									
Detection of other biosynthetic classes	X	X		X						
Gene cluster border prediction	X	X		X						
Comparative gene cluster analysis	X	X								
Sub-cluster analysis	X									
Prediction of putative novel gene cluster types	X	X				X		X		
Protein sequence input	X					X			X	X
Nucleotide sequence input	X	X	X	X	X	X	X			
Multi-contig input	X					X				
PKS structural modeling										X
NRPS/PKS domain phylogenomic analysis	(X)[a]					X				

antiSMASH 2.0 combines by far the most functionalities into a single framework and adds four key new features compared with antiSMASH 1.0. The phylogenomic analysis embedded in NaPDoS can be accessed through direct links from the relevant C and KS domains shown in the antiSMASH output page.
[a]Support for NRPS/PKS phylogenomic analysis via NaPDoS cross-reference.

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, ectoines, siderophores, phosphoglycolipids, melanins and a generic class of clusters encoding unusual secondary metabolite biosynthesis genes), version 2.0 adds support for oligosaccharide antibiotics, phenazines, thiopeptides, 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.

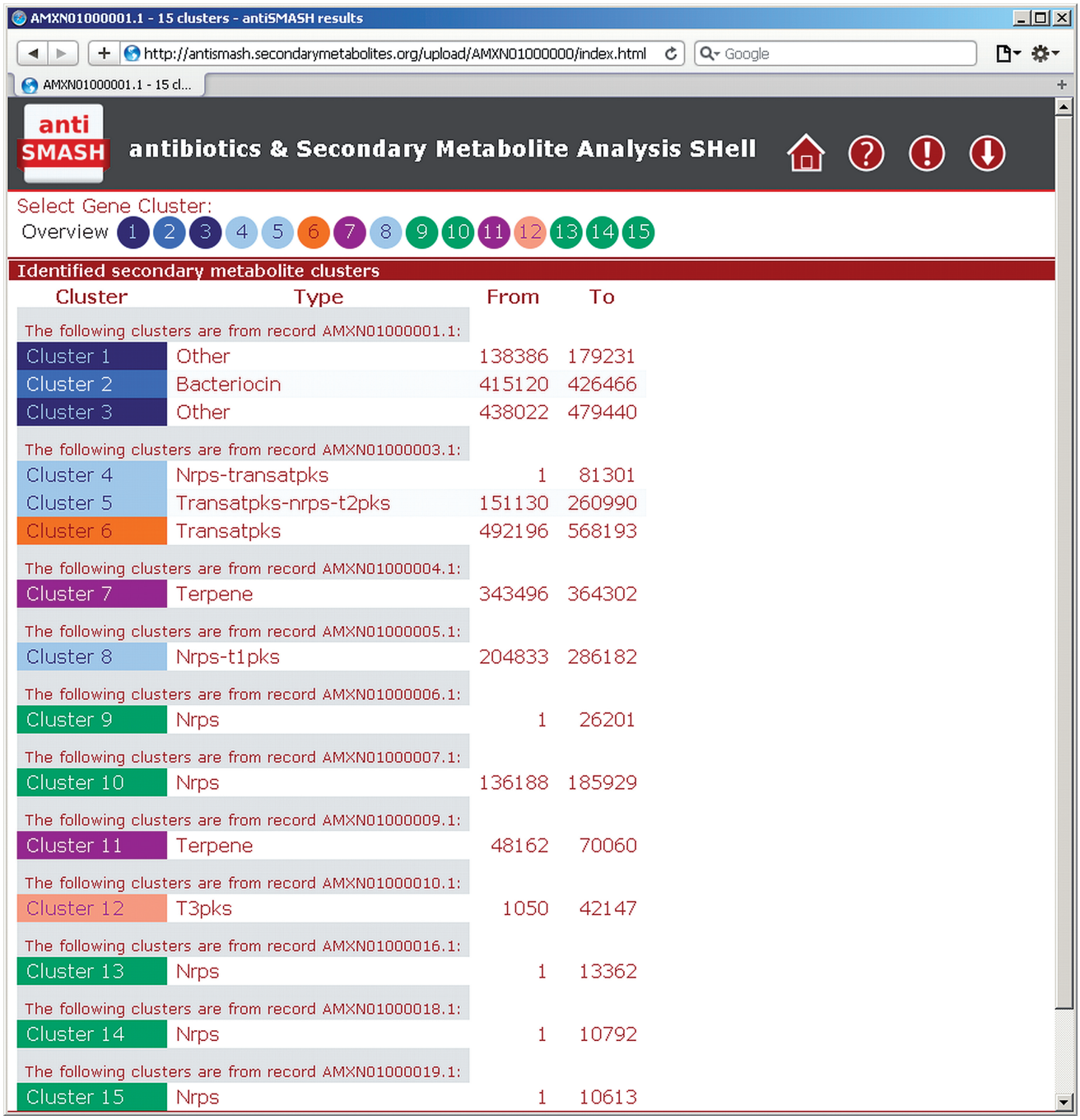


Figure 1. Overview page of the antiSMASH results. antiSMASH 2.0 gives an overview of all the output results in a single page, showing all the detected biosynthetic gene clusters with their type classifications and nucleotide positions. For inputs consisting of multiple entries/contigs, the clusters are separated by input entry/contig. Gene cluster types are signified by specific colors.

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 didehydrobutyrine (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-lantionine 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

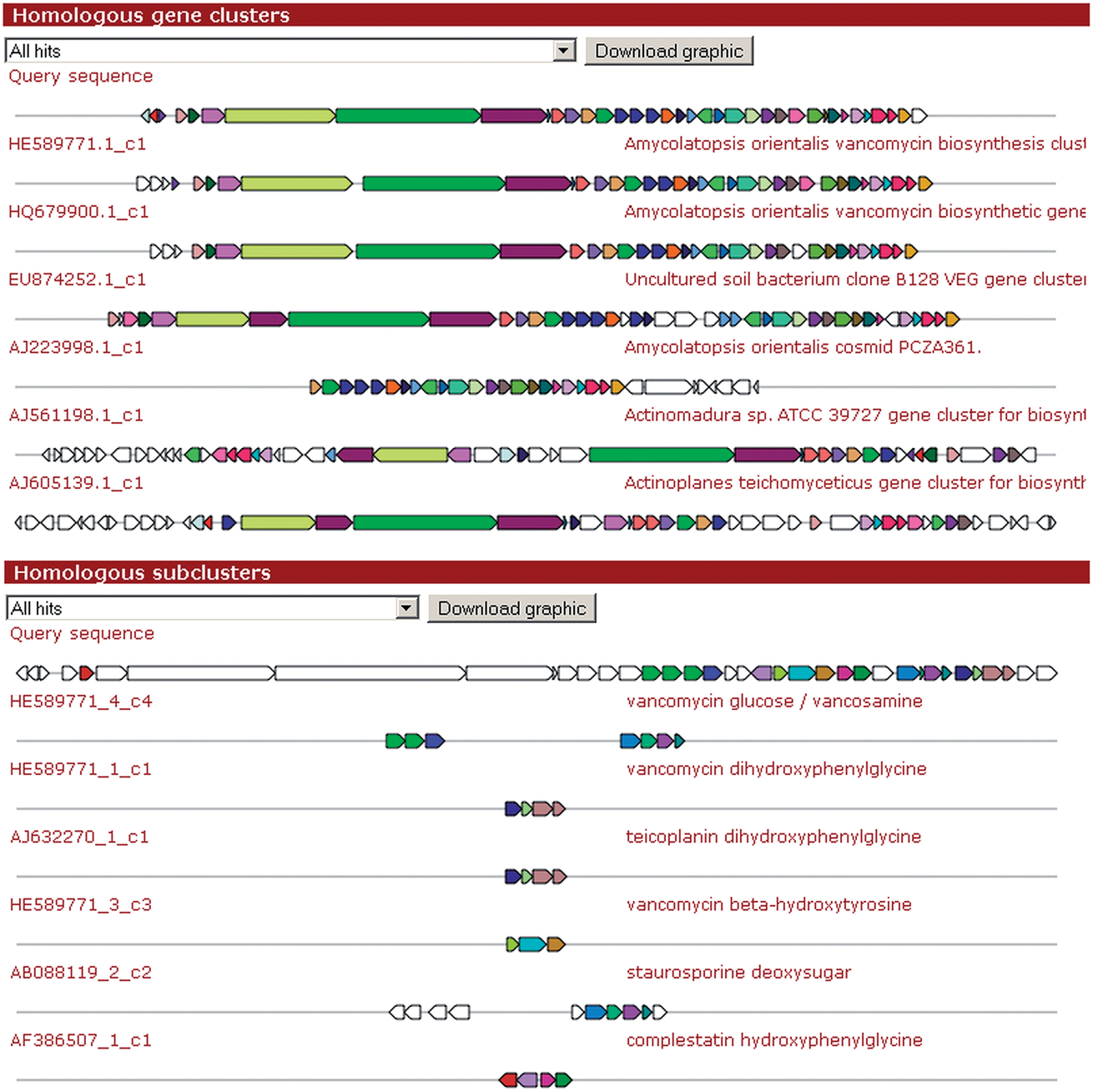


Figure 2. ClusterBlast and SubclusterBlast outputs for the balhimycin (23) biosynthesis gene cluster. The top six hits of each analysis module are shown. The ClusterBlast module shows the homology between the balhimycin gene cluster and the vancomycin, VEG, A40926 and teicoplanin biosynthesis gene clusters. Homologous genes are shown in identical colors, whereas white-colored genes have no BLAST hits between the gene clusters. The novel SubclusterBlast module can identify homologous sub-clusters encoding the biosynthesis of specific chemical moieties. In this case, SubclusterBlast is able to identify the dihydroxyphenylglycine (dHpg), hydroxyphenylglycine (Hpg) and hydroxytyrosine (Bht) precursor biosynthesis sub-clusters, as well as the vancosamine-like sugar biosynthesis sub-cluster.

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 actinomycetes *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

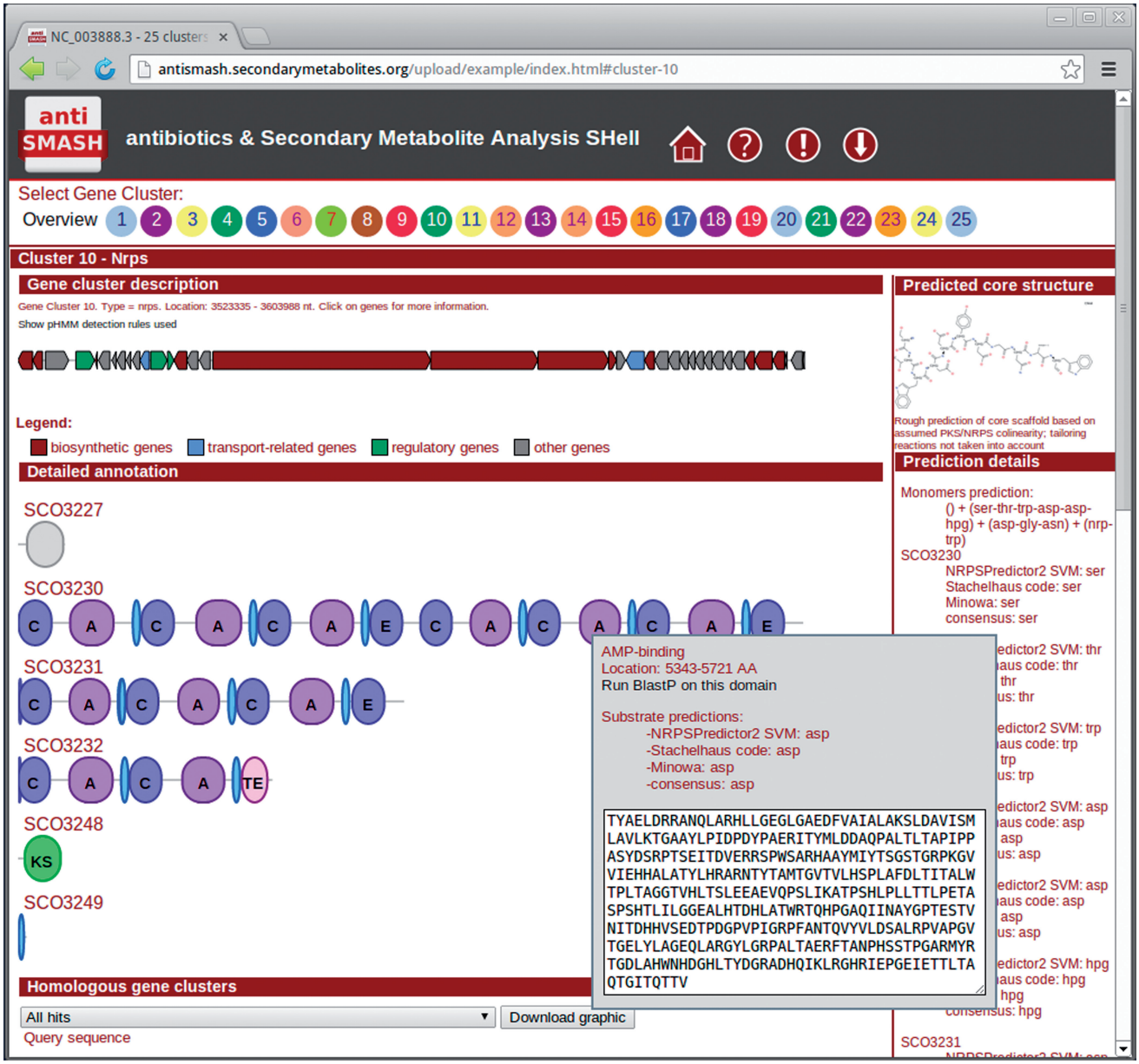


Figure 3. Top part of a gene cluster overview in the re-designed antiSMASH 2.0 output. The gene cluster shown is the calcium-dependent antibiotic biosynthesis gene cluster from *Streptomyces coelicolor* A3(2). The gene cluster-type–specific coloring of the numbered gene cluster buttons makes it easier to navigate through large result files. smCOG-based coloring of biosynthetic, transport-related and regulatory genes within the gene cluster make it easier to interpret the architecture of the gene cluster.

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 (DFG) and Open Access Publishing Fund of Tübingen University.

Conflict of interest statement. None declared.

REFERENCES

1. Newman,D.J. and Cragg,G.M. (2012) Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J. Nat. Prod.*, **75**, 311–335.
2. Crawford,J.M. and Clardy,J. (2012) Microbial genome mining answers longstanding biosynthetic questions. *Proc. Natl Acad. Sci. USA*, **109**, 7589–7590.
3. Scheffler,R.J., Colmer,S., Tynan,H., Demain,A.L. and Gullo,V.P. (2013) Antimicrobials, drug discovery, and genome mining. *Appl. Microbiol. Biotechnol.*, **97**, 969–978.
4. Zotchev,S.B., Sekurova,O.N. and Katz,L. (2012) Genome-based bioprospecting of microbes for new therapeutics. *Curr. Opin. Biotechnol.*, **23**, 941–947.
5. Medema,M.H., Breitling,R., Bovenberg,R. and Takano,E. (2011) Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nat. Rev. Microbiol.*, **9**, 131–137.
6. Medema,M.H., van Raaphorst,R., Takano,E. and Breitling,R. (2012) Computational tools for the synthetic design of biochemical pathways. *Nat. Rev. Microbiol.*, **10**, 191–202.
7. Weber,T. (2013) In silico tools for the analysis of antibiotic biosynthetic pathways. *Int. J. Med. Microbiol.* (epub ahead of print).
8. Fedorova,N.D., Moktali,V. and Medema,M.H. (2012) Bioinformatics approaches and software for detection of secondary metabolic gene clusters. *Methods Mol. Biol.*, **944**, 23–45.
9. Starcevic,A., Zucko,J., Simunkovic,J., Long,P.F., Cullum,J. and Hranueli,D. (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. *Nucleic Acids Res.*, **36**, 6882–6892.
10. Weber,T., Rausch,C., Lopez,P., Hoof,I., Gaykova,V., Huson,D.H. and Wohlleben,W. (2009) CLUSEAN: a computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters. *J. Biotechnol.*, **140**, 13–17.
11. Anand,S., Prasad,M.V., Yadav,G., Kumar,N., Shehara,J., Ansari,M.Z. and Mohanty,D. (2010) SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Res.*, **38**, W487–W496.
12. Khaldi,N., Seifuddin,F.T., Turner,G., Haft,D., Nierman,W.C., Wolfe,K.H. and Fedorova,N.D. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genet. Biol.*, **47**, 736–741.
13. Rausch,C., Weber,T., Kohlbacher,O., Wohlleben,W. and Huson,D.H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.*, **33**, 5799–5808.
14. 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.
15. Prieto,C., Garcia-Estrada,C., Lorenzana,D. and Martin,J.F. (2012) NRPSsp: non-ribosomal peptide synthase substrate predictor. *Bioinformatics*, **28**, 426–427.
16. Bachmann,B.O. and Ravel,J. (2009) Chapter 8. Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol.*, **458**, 181–217.
17. 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.
18. Conway,K.R. and Boddy,C.N. (2013) ClusterMine360: a database of microbial PKS/NRPS biosynthesis. *Nucleic Acids Res.*, **41**, D402–D407.
19. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
20. 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.
21. Medema,M.H., Takano,E. and Breitling,R. (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. *Mol. Biol. Evol.*, **30**, 1218–1223.
22. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
23. Pelzer,S., Süßmuth,R.D., Heckmann,D., Recktenwald,J., Huber,P., Jung,G. and Wohlleben,W. (1999) Identification and analysis of the balhimycin biosynthetic gene cluster and its use for manipulating glycopeptide biosynthesis in *Amycolatopsis mediterranei* DSM5908. *Antimicrob. Agents Chemother.*, **43**, 1565–1573.
24. Caboche,S., Pupin,M., Leclere,V., Fontaine,A., Jacques,P. and Kucherov,G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.*, **36**, D326–D331.
25. Paulsen,I.T., Press,C.M., Ravel,J., Kobayashi,D.Y., Myers,G.S., Mavrodi,D.V., DeBoy,R.T., Seshadri,R., Ren,Q., Madupu,R. *et al.* (2005) Complete genome sequence of the plant commensal *Pseudomonas fluorescens* Pf-5. *Nat. Biotechnol.*, **23**, 873–878.
26. Ohnishi,Y., Ishikawa,J., Hara,H., Suzuki,H., Ikenoya,M., Ikeda,H., Yamashita,A., Hattori,M. and Horinouchi,S. (2008) Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. *J. Bacteriol.*, **190**, 4050–4060.
27. Ichikawa,N., Oguchi,A., Ikeda,H., Ishikawa,J., Kitani,S., Watanabe,Y., Nakamura,S., Katano,Y., Kishi,E., Sasagawa,M. *et al.* (2010) Genome sequence of *Kitasatospora setae* NBRC

14216T: an evolutionary snapshot of the family *Streptomycetaceae*. *DNA Res.*, **17**, 393–406.
28. Udwary,D.W., Zeigler,L., Asolkar,R.N., Singan,V., Lapidus,A., Fenical,W., Jensen,P.R. and Moore,B.S. (2007) Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proc. Natl Acad. Sci. USA*, **104**, 10376–10381.
29. Nierman,W.C., Pain,A., Anderson,M.J., Wortman,J.R., Kim,H.S., Arroyo,J., Berriman,M., Abe,K., Archer,D.B., Bermejo,C. *et al.* (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. *Nature*, **438**, 1151–1156.
30. Yadav,G., Gokhale,R.S. and Mohanty,D. (2009) Towards prediction of metabolic products of polyketide synthases: an in silico analysis. *PLoS Comput. Biol.*, **5**, e1000351.
31. de Jong,A., van Heel,A.J., Kok,J. and Kuipers,O.P. (2010) BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Res.*, **38**, W647–W651.

1.1.3 Röttig et al. (2011) NRPSPredictor2

Nucleic Acids Research Advance Access published May 9, 2011

Nucleic Acids Research, 2011, 1–6
doi:10.1093/nar/gkr323

NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity

Marc Röttig[1,*], Marnix H. Medema[2,3], Kai Blin[4], Tilmann Weber[4], Christian Rausch[5] and Oliver Kohlbacher[1]

[1]Applied Bioinformatics, Center for Bioinformatics, Department of Computer Science, University of Tübingen, Sand 14, 72076 Tübingen, Germany, [2]Department of Microbial Physiology, [3]Groningen Bioinformatics Center, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 7, 9747AG Groningen, The Netherlands, [4]Interfaculty Institute of Microbiology and Infection Medicine, University of Tübingen, Auf der Morgenstelle 28 and [5]Algorithms in Bioinformatics Group, Center for Bioinformatics/ Department of Computer Science, University of Tübingen, Sand 14, 72076 Tübingen, Germany

Received March 15, 2011; Revised April 12, 2011; Accepted April 20, 2011

ABSTRACT

The products of many bacterial non-ribosomal peptide synthetases (NRPS) are highly important secondary metabolites, including vancomycin and other antibiotics. The ability to predict substrate specificity of newly detected NRPS Adenylation (A-) domains by genome sequencing efforts is of great importance to identify and annotate new gene clusters that produce secondary metabolites. Prediction of A-domain specificity based on the sequence alone can be achieved through sequence signatures or, more accurately, through machine learning methods. We present an improved predictor, based on previous work (NRPSpredictor), that predicts A-domain specificity using Support Vector Machines on four hierarchical levels, ranging from gross physicochemical properties of an A-domain's substrates down to single amino acid substrates. The three more general levels are predicted with an F-measure better than 0.89 and the most detailed level with an average F-measure of 0.80. We also modeled the applicability domain of our predictor to estimate for new A-domains whether they lie in the applicability domain. Finally, since there are also NRPS that play an important role in natural products chemistry of fungi, such as peptaibols and cephalosporins, we added a predictor for fungal A-domains, which predicts gross physicochemical properties with an F-measure of 0.84. The service is available at http://nrps.informatik.uni-tuebingen.de/.

INTRODUCTION

Non-ribosomally synthesized peptides are a class of highly important metabolites in the secondary metabolisms of bacteria and fungi (1,2). Important representatives of this family are mostly antibiotics like penicillin or vancomycin but also the immunosuppressant cyclosporin. The precursor peptides of these compounds are synthesized by non-ribosomal peptide synthetases (NRPSs), which are multi-modular megasynthetases with molecular weights up to 2.3 MDa (*tex1* NRPS from *Trichoderma virens*). NRPSs act as an assembly line that produces the final peptide by a chain of reactions occuring along that line. The primary sequence of the peptide product is determined by the sequential arrangement of minimal repetitive modules of an NRPS. The minimal module consists of three domains termed adenylation domain (A-domain), peptidyl carrier domain (PCP-domain) and condensation domain (C-domain). The A-domain is responsible for the recruitment of the amino acid monomers that are to be incorporated into the final product. Several hundred different A-domain substrate specificities have been biochemically characterized and each A-domain recruits a specific amino acid as monomer. Accordingly, the sequential order of A-domains along the assembly line determines (in the majority of cases) the primary sequence of the final peptide product. A comprehensive source of NRPS peptides and monomers is the NORINE database assembled by Caboche *et al.* which currently features over 1000 peptide products and over 500 monomers (3) . The cross linking between each adjacent monomer is carried out by the help of the other two domains that synthesize the peptide bond between these monomers. The minimal module is often equipped with additional

*To whom correspondence should be addressed. Tel: +49 7071 29 70464; Fax: +49 7071 29 5152; Email: roettig@informatik.uni-tuebingen.de

domains that allow for modifications of the recruited amino acid monomers like epimerization, methylation or formylation.

The structure–function relationship for monomer recruitment by A-domains has been further elucidated by Stachelhaus *et al.* and Challis *et al.* by examining the crystal structure of the peptide synthetase gramicidin S synthetase 1 (GrsA, PDB-ID: 1AMU) (4–6). The structure of the GrsA adenylation domain was determined with a co-crystallized phenylalanine monomer and thus delivers additional structural information about the binding pocket of the A-domain, which enabled Stachelhaus *et al.* to propose a specificity conferring-code of A-domains by relating the active site configuration of A-domains to the corresponding substrates.

The specificity-conferring code was based on 10 active site residues and it could be used to predict the putative substrates of A-domains for which only the sequence was known. Many NRPS services like the NRPS-PKS knowledgebase, the NP.searcher or the system devised by Bachmann *et al.* make use of this specificity-conferring code to predict putative A-domain substrates (7–9). The specificity-conferring code was further refined by Rausch *et al.* (10) by not only considering these 10 residues but by using all active site residues within 8 Å of the amino acid substrate. A predictor, NRPSpredictor, based on Transductive Support Vector Machines (TSVMs) was built on these 34 active site residues to predict A-domain specificity. In the following part of this article we will present details about the new version of this predictor, termed NRPSpredictor2, namely the improved prediction performance, simplified descriptor set used for signature encoding and estimation of the applicability domain of the predictor.

MATERIALS AND METHODS

Method outline

The predictions of substrate specificity are based on the configuration of the residues in the active site of an A-domain. We therefore made use of an A-domain crystal structure (PDB-ID: 1AMU) as a template to determine these active site residues. The positions of these residues were then located in the A-domain sequences of our training data set, and for each domain we extracted those positions. Having labeled sequence data, we applied machine learning methods, namely SVMs, to train predictors of substrate specificity. The predictions are based on numerical representations of the extracted signatures. The predictors were trained as detectors for each known substrate specificity in a one-versus-rest scheme, so every predictor that gives a positive prediction signals that the query A-domain might activate the corresponding substrate. Using this scheme, a query A-domain might yield positive signals from more than one predictor and thereby giving the user additional information about possible substrate promiscuity of the A-domain or ambiguity of the prediction.

Training data

The starting point for this work were the 397 labeled A-domains collected by Rausch *et al.* for which the specificity had been harvested from scientific literature describing their experimental characterization (10). We added 79 labeled bacterial A-domains and 100 labeled fungal A-domains to the database of NRPSpredictor. Furthermore, we added 4282 unlabeled bacterial and 814 unlabeled fungal A-domains to the data set (see Supplementary Material S1). These A-domains were retrieved from the UniProt database by an automated BLAST search for A-domains that are embedded within a minimal NRPS module, which requires the existence of an A-domain (Pfam-ID: PF00501), C-domain (Pfam-ID: PF00668) and PCP-domain (Pfam-ID: PF00550) (11,12).

Signature extraction

The set of all active site amino acids, called the signature, was identified by extracting all residues within 8 Å of the substrate phenylalanine in the crystal structure of GrsA (PDB-ID: 1AMU). These 34 positions were then extracted from the set of training sequences using an A-domain profile HMM and selecting relevant positions from the alignment. The specificity conferring code proposed by Stachelhaus *et al.* is a subset of these 34 residues and is also reported by the web server (6). Handling of protein structures, extraction of signatures and further processing was carried out using the Active Site Classification (ASC) software (13).

Encoding

NRPSpredictor2 makes use of two feature encodings for amino acids: one is the original encoding proposed by Rausch *et al.* based on 12 AAindex (14) descriptors and the other is a reduced encoding based on three z-scales descriptors devised by Wold *et al.* (15). The z-scales descriptors represent the following physicochemical properties: hydrophobicity (WOLS870101), size (WOLS870102) and electronic properties (WOLS870103). Each signature can be embedded in $\mathbb{R}^n$ by encoding each residue into a descriptor tuple and concatenating these tuples. The predictive models are then trained on the transformed data.

SVMs

SVMs are classifiers based on the maximum margin principle (16,17). During SVM training a hyperplane in feature space is determined that gives the largest possible margin between the positive and negative class, thereby yielding an intuitively robust classifier. The hyperplane gives a decision surface defined by $f(x) = \Sigma_i\ y_i\alpha_i\ k(x, x_i)$ whose functional value is zero for data points directly on the hyperplane, +1 or more for data points in the positive half-space and −1 or less for points in the negative half-space. The margin is determined by the geometric distance of points with functional value of +1 or −1 (support vectors) to the hyperplane. NRPSpredictor2 uses the RBF kernel $k(x, y) = \exp(-\gamma||x-y||^2)$ and the linear kernel $k(x, y) = x'y$ on the physico-chemical feature vectors. For the training of SVMs a set of labeled data points

(x_i, y_i) is needed where x_i is from $\mathbb{R}^n$ and the labels y_i are in (+1, −1) for two-class problems.

TSVMs

TSVMs extend classical SVMs by the property of making use of unlabeled data to train more robust classifiers, especially in the case of scarce labeled training data (18). TSVMs try to determine a separating hyperplane that does not cut clusters of data by forcing the hyperplane to go through low data density regions. This is enforced by keeping the margin clear of unlabeled data points. However, the objective function of TSVMs is not that easily optimized as the classical SVM objective, hence heuristics have to be used to optimize the objective. For NRPSpredictor2 we make use of the SVMlight package that offers such an heuristic to train TSVM classifiers (18).

Prediction levels and predictor quality

NRPSpredictor2 was designed to predict the putative substrate specificity on four different hierarchical levels for bacterial A-domains and on one level for fungal A-domains. The bacterial levels are: gross physico-chemical properties of the substrate (hydrophobic–aromatic, hydrophobic–aliphatic and hydrophilic), large clusters, small clusters and on a single amino acid level (Table 1). The fungal predictor predicts only on the gross physico-chemical properties level (hydrophobic-aromatic, hydrophobic-aliphatic and hydrophilic) due to the lack of sufficient fungal training data to allow further subdivision of substrate clusters. However, within the web server we trigger the bacterial models to give also more fine grained predictions for fungal signatures. An overview of the set of bacterial prediction levels is given in Table 1. For many substrates there are only very few labeled A-domains, like the 2-amino-butyric acid (Abu) specificity with less than five known A-domain sequences. For these specificities no SVM-model was built. Instead, we make use of the Nearest-Neighbor Rule to get a specificity prediction, by reporting for each query the substrate specificity of the most similar active-site signature (based on the Stachelhaus code) in our database, along with the sequence identity.

Predictor validation

To quantify the performance of the NRPSpredictor2 we used the F-measure as quality criterion, which is defined as the harmonic mean of precision and recall. The precision is defined by $prec = tp/(tp+fp)$ and the recall (or sensitivity) is defined by $rec = tp/(tp+fn)$, where tp, fp and fn are the number of true positives, false positives and false negatives, respectively. The precision (or positive predictive value) measures how reliable a positive prediction of a substrate specificity detector is and the recall measures how good the detector is in finding the true positives. To determine the performance on new test data we applied a repeated external validation scheme. We split the whole data set into half, selected and trained a SVM model on one half of the data and evaluated the predictor performance on the other half, the independent test set. This procedure was repeated on 10 shuffled versions of the whole data set to get a more robust average of the predictor performance on new test data.

Applicability domain

The applicability domain of a predictor is a concept that helps to give for each predictor query a feedback whether that query is too far away from the data used during training or whether that instance lies within the, say, 95% support volume of the training data. Predictions for queries that do not lie within the applicability domain of the model should be handled with more care. To model the applicability domain of our model we made use of the 1-Class SVM concept as described by Schölkopf *et al.* (19). Therefore, we modelled the 95% support of our data using the 1-Class SVM functionality of LIBSVM. We selected values for γ and ν in such a way as to achieve a recall of ~95% on left out data and then trained a 1-class SVM for the whole data set using these parameters to describe the 95% support volume in feature space of our data.

RESULTS

Predictor quality

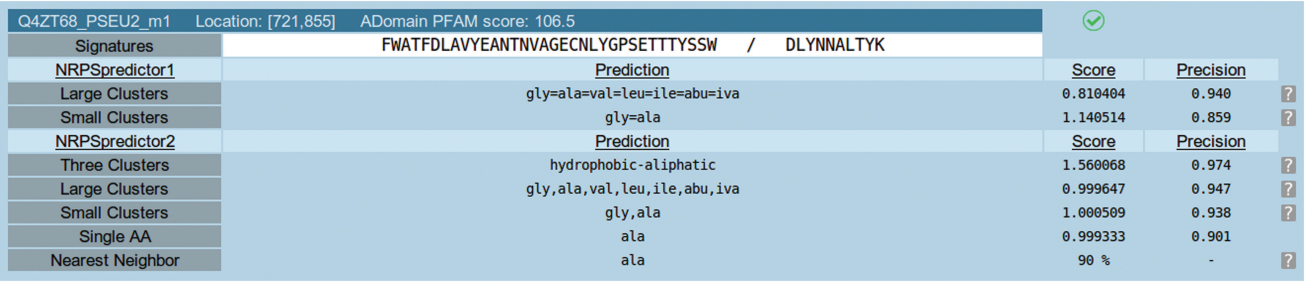
The quality of each bacterial predictor as determined by our model validation is given in Table 1. It can be observed that the predictors at the highest hierarchical level are the best-performing ones. At the level of gross physico-chemical properties we have an average F-measure of $F = 0.94$, whereas the average F-measure at the most fine-grained level (single substrates) is $F = 0.80$. Generally, the average performance as quantified by the F-measure is $F = 0.94$ for the three class level, $F = 0.93$ for the large clusters level, $F = 0.89$ for the small clusters level and $F = 0.80$ for the single substrate level. The fungal predictor has an average F-measure of $F = 0.84$ at the three class level. Table 1 also gives for each prediction task the best performing kernel, feature encoding and SVM type (classic or TSVM).

A general trend is that, except from the more exotic aromatic substrates, like the hydroxy-benzoic derivatives that can be predicted very well, the other more common aromatic substrates are predicted less reliably. One reason might be the observed promiscuity of the A-domains utilizing these substrates (10). When compared with the original version of the NRPSpredictor (Table 1) the new version could improve the performance (F-measure) on the large cluster level and on the small clusters level by roughly one percentage point. While the original NRPSpredictor was able to predict the membership to clusters of amino acids only, NRPSpredictor2 also can predict single amino acid specificities. The newly introduced applicability domain gives further information on the quality of the specificity prediction. Upon request of many colleagues working on fungal NRPSs, a predictor specific for fungal NRPS sequences was included in NRPSpredictor2.

Table 1. Prediction levels and predictor quality (bacterial)

Classname	Members	Type	NRPSpredictor2			NRPSpredictor1
			F	Prec.	Rec.	*F*
Three class						
Hydrophobic aliphatic	Ala, Gly, Val, Leu, Ile, Abu, Iva Ser, Thr, Hpg, Dhpg, Cys, Pro, Pip	W,R,T	0.974	0.974	0.974	–
Hydrophilic	Arg, Asp, Glu, His, Asn, Lys, Gln, Orn, Aad	W,R,T	0.940	0.940	0.940	–
Hydrophobic aromatic	Phe, Tyr, Trp, Dhb, Phg, Bht	W,R,T	0.890	0.889	0.892	–
Large clusters						
Hydroxy-benzoic acid derivates	Dhb, Sal	W,R,T	0.982	1.000	0.967	0.982
Polar, uncharged (aliphatic with -SH)	Cys	R,R,T	0.976	0.975	0.975	0.954
Aliphatic chain or phenyl group with -OH	Ser, Thr, Dhpg, Hpg	R,R,T	0.968	0.967	0.969	0.963
Aliphatic chain with H-bond donor	Asp, Asn, Glu, Gln, Aad	W,R,C	0.958	0.969	0.950	0.942
Apolar, aliphatic	Gly, Ala, Val, Leu, Ile, Abu, Iva	W,R,T	0.940	0.947	0.934	0.940
Aromatic side chain	Phe, Trp, Phg, Tyr, Bht	W,R,T	0.881	0.881	0.881	0.881
Cyclic aliphatic chain (polar NH2 group)	Pro, Pip	R,R,T	0.867	0.867	0.867	0.811
Long positively charged side chain	Orn, Lys, Arg	W,R,T	0.864	0.898	0.833	0.861
		Ø	0.930	–	–	0.917
Small clusters						
2-amino-adipic acid	Aad	W,L,C	1.000	1.000	1.000	1.000
Dhb, Sal	Dhb, Sal	W,L,C	1.000	1.000	1.000	0.940
Polar, uncharged (hydroxy-phenyl)	Dhpg, Hpg	R,L,T	1.000	1.000	1.000	0.981
Cys	Cys	R,L,T	0.983	0.983	0.983	0.950
Serine-specific	Ser	W,R,T	0.972	1.000	0.947	0.936
Threonine-specific	Thr	W,L,C	0.969	0.978	0.961	0.942
Asp-Asn	Asp, Asn	W,L,C	0.948	0.969	0.931	0.942
Orn and hydroxy- Orn specific	Orn	R,L,T	0.900	0.900	0.900	0.800
Aliphatic, branched hydrophobic	Val, Leu, Ile, Abu, Iva	W,R,T	0.893	0.892	0.895	0.887
Tiny, hydrophilic, transition to aliphatic	Gly, Ala	W,L,C	0.886	0.938	0.843	0.859
Pro-specific	Pro	R,L,T	0.882	0.938	0.833	0.900
Polar aromatic ring	Tyr, Bht	W,R,T	0.857	0.892	0.825	0.793
Glu-Gln	Glu, Gln	W,L,C	0.813	0.850	0.791	0.860
Arg-specific	Arg	W,L,C	0.740	1.000	0.600	0.800
Unpolar aromatic ring	Phe, Trp	W,L,C	0.538	0.608	0.500	0.671
		Ø	0.892	–	–	0.884
Single substrates						
Aad	Aad	W,R,T	1.000	1.000	1.000	–
Cys	Cys	R,R,T	1.000	1.000	1.000	–
Hpg	Hpg	R,R,T	0.974	1.000	0.950	–
Ser	Ser	W,R,T	0.962	0.993	0.933	–
Thr	Thr	W,R,T	0.949	0.976	0.922	–
Dhb	Dhb	W,R,T	0.947	1.000	0.900	–
Dhpg	Dhpg	W,R,T	0.943	0.967	0.925	–
Asn	Asn	R,R,T	0.939	0.934	0.944	–
Orn	Orn	R,R,T	0.933	0.933	0.933	–
Ile	Ile	R,R,T	0.918	1.000	0.850	–
Gly	Gly	R,R,T	0.906	0.902	0.910	–
Ala	Ala	W,R,T	0.878	0.901	0.856	–
Arg	Arg	W,R,T	0.833	0.833	0.833	–
Iva	Iva	W,R,T	0.814	0.933	0.725	–
Val	Val	W,R,T	0.801	0.828	0.777	–
Leu	Leu	W,R,T	0.784	0.782	0.787	–
Pro	Pro	W,R,T	0.755	0.792	0.722	–
Bht	Bht	W,R,T	0.717	0.782	0.675	–
Glu	Glu	R,R,T	0.704	0.760	0.657	–
Pip	Pip	W,R,T	0.700	0.800	0.625	–
Asp	Asp	R,R,T	0.700	0.700	0.700	–
Tyr	Tyr	W,R,T	0.696	0.671	0.725	–
Gln	Gln	W,R,T	0.689	0.775	0.620	–
Phe	Phe	W,R,T	0.688	0.740	0.643	–
Lys	Lys	R,R,T	0.400	0.500	0.333	–
Trp	Trp	W,R,T	0.320	0.400	0.267	–

The column type gives the best performing predictor encoded by three letters: the first letter represents the used encoding (W: Wold, R: Rausch), the second letter the used kernel (L: linear, R: RBF) and the third letter the used SVM type (C: classical SVM T: transductive SVM). The columns F, Prec. and Rec. give the *F*-measure, Precision and Recall of the best predictor, respectively. Aad: 2-amino-adipic-acid; Bht: beta-hydroxy-tyrosine; Hpg: 4-hydoxy-phenyl-glycine; Dhb: 2,3-dihydroxy-benzoic acid; Dhpg: 3,5-dihydroxy-phenyl-glycin; Iva: isovaline; Orn: ornitine; Pip: pipecolic acid; Sal: salicylic acid.



Q4ZT68_PSEU2_m1 Location: [721,855] ADomain PFAM score: 106.5			
Signatures	FWATFDLAVYEANTNVAGECNLYGPSETTTYSSW / DLYNNALTYK		
NRPSpredictor1	Prediction	Score	Precision
Large Clusters	gly=ala=val=leu=ile=abu=iva	0.810404	0.940
Small Clusters	gly=ala	1.140514	0.859
NRPSpredictor2	Prediction	Score	Precision
Three Clusters	hydrophobic-aliphatic	1.560068	0.974
Large Clusters	gly,ala,val,leu,ile,abu,iva	0.999647	0.947
Small Clusters	gly,ala	1.000509	0.938
Single AA	ala	0.999333	0.901
Nearest Neighbor	ala	90 %	-

Figure 1. NRPSpredictor2 prediction report for one extracted A-domain. On top, the ID of the parent sequence, location of the A-domain within the sequence and the bit score of the PFAM-HMM are given. The green checkmark signals that the signature sequence lies within the applicability domain of the model. The extracted 8 Å signature and Stachelhaus code are given directly below. Subsequently, the list of predictions is given along with the score of the respective SVM predictors. For each predictor we also report the reliability of that predictor as determined during model validation. The last row gives the nearest sequence neighbor in the NRPSpredictor2 database (based on Stachelhaus code) and the respective sequence identity.

Web server

Users of the NRPSpredictor2 web server can submit their data as full NRPS sequences in multi-FASTA format and the signatures will be extracted automatically. Another option is to directly supply the extracted signatures and request a prediction from the predictor, thus users are not required to disclose the full NRPS sequence. After short extraction and prediction phases the user receives a list of detected A-domains along with the predictions of NRPSpredictor2 at each hierarchical level. For user convenience we report the predictions of the original version of the NRPSpredictor. A typical report for one particular extracted A-domain is given in Figure 1. For each extracted A-domain the ID of the parent sequence is given with the number of the A-domain added as suffix. The exact location of the A-domain within the parent sequence is also reported, along with the bit score of the Pfam HMM that extracted this domain. The result of the applicability check is given by either a green checkmark (as shown in Figure 1) if the query signatures lies within the applicability domain of our predictor or as red X if the signature is most likely outside the applicability domain of the model. In this case the prediction should be taken with caution. Finally, the specificity predictors that give positive predictions for this signature are listed for each hierarchical level. The scores of the SVMs along with the precision of the SVM predictors, determined during model validation, are given in the last two columns. The last row gives the nearest neighbor to the query signature found in our database of annotated A-domain signatures (based on Stachelhaus code) along with the sequence identity. Using this rule NRPSpredictor2 can even detect specificities for which no SVM model could be learned, due to scarcity of labeled training data.

DISCUSSION

We have presented the NRPSpredictor2 that predicts A-domain substrate specificity based on sequence and structural information about the active site of the domain. The new predictor comes with an improved prediction performance over the previous version and also with two new prediction levels, namely the gross physico-chemical properties level and the detailed prediction level, which predicts the single amino acid likely to be activated by the given A-domain. The performance improvement was mainly due to the additional labeled training data as well as the use of an additional encoding of A-domain signatures (Wold encoding). The transductive SVM method, which makes use of unlabeled data, is very important in the settings with scarce training data per class, as can be seen in the most detailed prediction tasks (single amino acid level) where the transductive SVM is the best performing type of SVM. In the upper prediction levels classical SVMs quite often suffice to build a well-performing predictive model. In some of these cases the use of a transductive SVM might even hurt performance due to the heuristic training procedure that may yield suboptimal models, when compared to the classical SVM models, which use only labeled training data. We also created a new web interface for the predictor, allowing prediction of either bacterial or fungal sequences based on full NRPS sequences or already extracted signatures. For comparison purposes the web server also reports the predictions of the original NRPSpredictor. Finally, NRPSpredictor2 has also been incorporated into antiSMASH, a new comprehensive pipeline for secondary metabolite gene cluster detection and annotation, which allows users to rapidly analyze complete NRPS gene clusters or even whole genomes containing multiple NRPS gene clusters (M. H. Medema *et al.*, submitted for publication).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Rainer Breitling for constructive comments and suggestions.

FUNDING

German Ministry for Education and Research (BMBF) [0315585A (GenBioCom) to T.W.]. The work of MHM was supported by the Dutch Technology Foundation (STW), which is the applied-science division of The Netherlands Organisation for Scientific Research (NWO) and the Technology Programme of the Ministry of Economic Affairs (grant STW 10463). Funding for open access charge: University of Tübingen.

Conflict of interest statement. None declared.

REFERENCES

1. Marahiel,M.A., Stachelhaus,T. and Mootz,H.D. (1997) Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chem. Rev.*, **97**, 2651–2674.
2. Schwarzer,D., Finking,R. and Marahiel,M.A. (2003) Nonribosomal peptides: from genes to products. *Nat. Prod. Rep.*, **20**, 275–287.
3. Caboche,S., Pupin,M., Leclere,V., Fontaine,A., Jacques,P. and Kucherov,G. (2008) NORINE: a database of nonribosomal peptides. *Nucleic Acids Res.*, **36**, D326–D331.
4. Challis,G.L., Ravel,J. and Townsend,C.A. (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem. Biol.*, **7**, 211–224.
5. Conti,E., Stachelhaus,T., Marahiel,M.A. and Brick,P. (1997) Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *EMBO J.*, **16**, 4174–4183.
6. Stachelhaus,T., Mootz,H.D. and Marahiel,M.A. (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.*, **6**, 493–505.
7. Ansari,M.Z., Yadav,G., Gokhale,R.S. and Mohanty,D. (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. *Nucleic Acids Res.*, **32**, W405–413.
8. Li,M.H., Ung,P.M., Zajkowski,J., Garneau-Tsodikova,S. and Sherman,D.H. (2009) Automated genome mining for natural products. *BMC Bioinformatics*, **10**, 185.
9. Bachmann,B.O. and Ravel,J. (2009) Methods for *in silico* prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. *Methods Enzymol.*, **458**, 181–217.
10. Rausch,C., Weber,T., Kohlbacher,O., Wohlleben,W. and Huson,D.H. (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res.*, **33**, 5799–5808.
11. UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.*, **38**, D142–148.
12. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
13. Röttig,M., Rausch,C. and Kohlbacher,O. (2010) Combining structure and sequence information allows automated prediction of substrate specificities within enzyme families. *PLoS Comput Biol.*, **6**, e1000636.
14. Kawashima,S., Pokarowski,P., Pokarowska,M., Kolinski,A., Katayama,T. and Kanehisa,M. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202–D205.
15. Wold,S., Eriksson,L., Hellberg,S., Jonsson,J., Sjöström,M., Skagerberg,B. and Wikström,C. (1987) Principal property-values for 6 nonnatural amino-acids and their application to a structure activity relationship for oxytocin peptide analogs. *Can. J. Chem.*, **65**, 1814–1820.
16. Boser,B.E., Guyon,I.M. and Vapnik,V.N. (1992) *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. ACM, Pittsburgh, Pennsylvania, United States, pp. 144–152.
17. Cortes,C. and Vapnik,V. (1995) Support-vector networks. *Mach. Lear.*, **20**, 273–297.
18. Joachims,T. (1999) *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., pp. 200–209.
19. Schölkopf,B., Platt,J.C., Shawe-Taylor,J.C., Smola,A.J. and Williamson,R.C. (2001) Estimating the Support of a High-Dimensional Distribution. *Neural Comput.*, **13**, 1443–1471.

1.2 Submitted Manuscripts

Blin et al. (2013a) Improved lanthipeptide prediction

Improved lanthipeptide detection and prediction for antiSMASH

Kai Blin[1,2], Daniyal Kazempour[1], Wolfgang Wohlleben[1,2], Tilmann Weber[1,2,*]

1 Division of Microbiology/Biotechnology, Interfaculty Institute of Microbiology and Infection Medicine, Eberhard-Karls-University of Tübingen, Germany

2 German Centre for Infection Research, Tübingen, Germany

∗ **E-mail: Tilmann.Weber@biotech.uni-tuebingen.de**

Abstract

Lanthipeptides are a class of ribosomally synthesised and post-translationally modified peptide (RiPP) natural products from the bacterial secondary metabolism. Their name is derived from the characteristic lanthionine or methyl-lanthionine residues contained in the processed peptide. Lanthipeptides that possess an antibacterial activity are called lantibiotics. Whereas multiple tools exist to identify lanthipeptide gene clusters from genomic data, no programs are available to predict the post-translational modifications of lanthipeptides, such as the proteolytic cleavage of the leader peptide part or tailoring modifications based on the analysis of the gene cluster sequence.

antiSMASH is a software pipeline for the identification of secondary metabolite biosynthetic clusters from genomic input and the prediction of products produced by the identified clusters.

Here we present a novel antiSMASH module using a rule-based approach to combine signature motifs for biosynthetic enzymes and lanthipeptide-specific cleavage site motifs to identify lanthipeptide clusters in genomic data, assign the specific lanthipeptide class, predict prepeptide cleavage, tailoring reactions, and the processed molecular weight of the mature peptide products.

Introduction

Lanthipeptides

Lanthipeptides are polycyclic peptides named after the thioether-linked amino acids lanthionine and (2*S*,3*S*,6*R*)-3-methyllanthionine contained in the mature peptide. Formerly called lantibiotics from "lanthionine-containing antibiotics", the new name lanthipeptide was proposed to also include non-antibiotic peptides

of the same biosynthetic origin [1]. Lanthipeptides are ribosomally synthesised and post-translationally modified peptides (RiPPs). The extensive post-translational modifications enhance the stability of the mature peptide against proteolysis and temperature stress. Lanthipeptides are encoded on the genome as a precursor peptide containing a leader and a core peptide part. The lanthionine (Lan) and methyllanthionine (MeLan) residues are introduced in a two-step reaction. First, serine (Ser) and threonine (Thr) residues are dehydrated to dehydroalanine (Dha) and dehydrobutyrine (Dhb), respectively, usually with an intermediate phosphorylation step. In the second step, a Michael-type addition by cysteine (Cys) residues onto the dehydro amino acids then yields the thioether cross-links.

Depending on the biosynthetic enzymes installing the thioether cross-links, lanthipeptides are divided into different classes [2]. Currently, four lanthipeptide classes are know. In class I lanthipeptides, the dehydration is catalysed by a dedicated dehydratase commonly called LanB. Cyclisation is carried out by a cyclase called LanC. In specific gene clusters, the generic enzyme names might be replaced by a more specific name: for example in the nisin gene cluster, the LanB-type dehydratase is called NisB and the LanC-type cyclase is called NisC. For the remaining class II, III and IV lanthipeptides, both dehydration and cyclisation are catalysed by a single bi-functional enzyme. A class II LanM enzyme carries an N-terminal dehydratase domain with little sequence similarity to other characterised enzymes. The C-terminal cyclisation domain is similar to the LanC enzymes from class I lanthipeptide cyclases. The bi-functional enzymes for class III (LanKC) and IV (LanL) have a common N-terminal phosphoserine/phosphothreonine lyase domain and a central kinase domain. The C-terminal cyclisation domain in class III enzymes, while similar to the cyclisation domains from the other classes, lacks three zinc-bindinding residues that are conserved in the other classes. In class IV, those residues are present.

In addition to the introduction of Lan and MeLan, a number of further post-translational modifications may occur if the appropriate tailoring enzymes are present in the gene cluster. Among the modifications found in lanthipeptides is the formation of *S*-[(*Z*)-2-aminovinyl]-D-cysteine (AviCys) or *S*-[(*Z*)-2-aminovinyl]-(3*S*)-3-methyl-D-cysteine (AviMeCys) [3]. The formation of AviCys and AviMeCys is catalysed by an enzyme of the family of homo-oligomeric flavin-containing cysteine decarboxylases. The enzyme with the generic designation LanD catalyses the oxidative decarboxylation of a C-terminal cysteine residue to a reactive thio–enol intermediate, which then cyclises with a Dha or Dhb residue, respectively, yielding AviCys or AviMeCys. An example would be the AviCys residue in epidermin introduced by EpiD [4]. Another post-translational modification is the chlorination of tryptophan residues

catalysed by a flavin-dependent tryptophan halogenase designated LanH. This kind of reaction has been observed in the chlorination of tryptophan by MibH in microbisporicin biosynthesis [5]. If the cluster contains a cytochrome P450 oxygenase designated LanO, amino acids in the modified precursor peptide can be hydroxylated, as observed in the hydroxylation of proline in microbisporicin biosynthesis [5]. If the N-terminal amino acid is Dha and an oxidoreductase is present in the cluster, the N-terminal amino acid can be converted to lactate, observed in the epicidin 280 cluster [6].

antiSMASH

antiSMASH, the antibiotics and secondary metabolite analysis shell, is a software pipeline for the automated identification of secondary metabolite biosynthesis clusters. Initially, product prediction was only possible for non-ribosomal peptide synthase (NRPS) and polyketide synthase (PKS) gene clusters [7]. Earlier this year, we released antibiotics and secondary metabolite analysis shell (antiSMASH) 2.0 [8]. In the new release, the architecture of the software was redesigned, now making it possible to add new predictors as self-contained plug-ins. antiSMASH is available as a web service at http://antismash.secondarymetabolites.org and can also be downloaded to run standalone. It is released under the GNU Affero Public License version 3, an OSI-approved Open Source license.

Here we present the implementation of a lanthipeptide-specific analysis module for antiSMASH 2. The module is shipped with the antiSMASH 2.1 release and also running on the public web server.

Design, Implementation and Validation

Secondary metabolite clusters in antiSMASH are identified using Hidden Markov Models (HMMs) of protein motifs for key biosynthetic enzymes. Which profiles are required to be identified for a specific secondary metabolite type is described by a rules file containing one rule-set per cluster type. Rule-sets can be simple hits against a single profile, AND and OR combinations of multiple profiles, or a selection of more complex rules, e.g. requiring a match against a minimum of n hits of a set of profiles. New secondary metabolite types can be added by adding new profile HMMs and extending the rules file.

Once the cluster detection has identified a secondary metabolite cluster of a certain type, specific analysis modules can be run to generate a more detailed analysis of the pathway and the prediction of the product of a given cluster. Specific analysis modules are written as self-contained plug-ins that are

loaded from the user's `PYTHONPATH` at run-time.

Identification of Lanthipeptide Biosynthetic Gene Clusters

To make more detailed cluster information available to the downstream specific analysis module, the cluster detection rules have been extended to include domain-specific Pfam [9] HMMs for the N-terminal domain (PFAM: PF13575) of class II LanM enzymes, the central kinase domain (PFAM: PF00069) of class III and IV enzymes, LanD-type flavin-dependent decarboxylases (PFAM: PF02441), LanH-type flavin-dependent halogenases (PFAM: PF04820), LanO-type cytochrome P450 oxygenases (PFAM: PF00067) and EciO-type short chain dehydrogenases (PFAM: PF00106, PF13561) (see Table 1 for details).

Prediction of the Lanthipeptide Class

Lanthipeptide classes are assigned by determining the domains present in the biosynthetic enzymes. Characteristic for class I lanthipeptides is the separate LanB enzyme containing the dehydratase domain, so the class prediction checks for a hit against the `Lant_dehyd_N` or `Lant_dehyd_C` domains. The dehydratase domain of class II LanM-type enzymes is characteristic as well, so if the cluster contains this dehydratase domain (PFAM: PF13575), the lanthipeptide will be considered class II. Class III LanKC-type and class IV LanL-type enzymes are identified via the central kinase domain (PFAM: PF00069). To differentiate between class III and IV enzymes, the algorithm checks if the conserved zinc binding sites in the C-terminal cyclase domain are absent (class III) or present (class IV).

Cleavage Site prediction

In the final step in lanthipeptide biosynthesis, a protease cleaves the leader peptide part off the modified precursor peptide to yield the mature peptide. Depending on the class of the lanthipeptide, the cleavage site motives vary widely. In order to predict the cleavage site, we have created a manually curated set of HMMs, one for lanthipeptide classes I and II each (Tables 2, 3). Profiles for the HMMer 2.3.2 software [10] were generated using `hmmbuild profile.hmm alignment.fa; hmmcalibrate profile.hmm`. As the method depends on the size of the seed sequence data set, we decided not to include cleavage site predictions for class III (only six seed sequences available) and class IV (no experimentally verified sequences available) lanthipeptides. Once more seed sequences become available for these two classes,

adding cleavage site predictions using the same method will be straightforward.

Monoisotopic mass, molecular weight and alternative weights

Once the cleavage site is predicted, both the monoisotopic mass and the average molecular weight are calculated. For the calculation of these numbers it is assumed that all Ser and Thr residues are dehydrated to Dha and Dhb respectively. As a lack of dehydration is frequently observed but the mechanism behind this has not been elucidated, we also calculate alternative weights under the assumption that one up to n Ser or Thr are not dehydrated, where n is the number of Ser and Thr residues in the core peptide subtracted by the number of Cys residues in the core peptide. This upper bound is set to account for the observation that all Cys residues tend to participate in Lan or MeLan bridges with Dha or Dhb residues.

Predicting Tailoring Reactions

Tailoring reactions are not performed by the core biosynthetic enzymes that perform the dehydration and cyclisation but instead by additional enzymes also encoded on the cluster.

AviCys and AviMeCys formation

The unusual amino acids AviCys and AviMeCys are formed by oxidative decarboxylation of the C-terminal Cys residue. The resulting thio–enol intermediate cyclises with a Dha or Dhb side-chain respectively. This reaction is catalysed by a LanD-type flavin-dependent decarboxylase, identified by a hit against the PFAM PF02441 profile with a score ≥ 20. The formation of AviCys or AviMeCys reduces the predicted peptide weight by 46 Da.

Halogenation

Identified by a hit against the PFAM PF04820 profile with a score ≥ 20, LanH-type halogenases chlorinate an amino acid side chain, increasing the predicted peptide weight by 34 Da.

Hydroxylation

LanO-type cytochrome P450 oxygenases catalyse the regiospecific oxidation of non-activated hydrocarbons. The enzyme is identified by a hit against PFAM PF00067 with a score ≥ 60. The hydroxylation increases the predicted peptide weight by 16 Da.

Lactate formation

EciO-type short-chain dehydrogenases identified by a hit against PFAM PF00106 or PFAM PF13561 with a score ≥ 100 catalyse the final step of the conversion of the N-terminal Dha residue to lactate. This increases the predicted peptide weight by 2 Da.

Predicting the number of Lan and MeLan bridges

To predict the number of Lan and MeLan bridges, a simple heuristic is applied using the formula

$$|b| = min(|S| + |T|, |C|) - v, \qquad v = \begin{cases} 1 & \text{if AviCys or AviMeCys residue is present} \\ 0 & \text{otherwise} \end{cases}$$

where $|b|$ is the number of bridges, and $|S|, |T|, |C|$ is the number of amino acids Ser, Thr, and Cys in the core peptide.

Validation and Benchmarking

To validate the robustness of the cleavage site profiles, we used n-fold cross validation. For a seed alignment of size n, we built n different profiles by including $n - 1$ sequences, and then checked if a cleavage site was predictable and correct for the left out sequence. A cleavage site was predictable if the profile produced a hit with a score above the threshold. A cleavage site was considered correct if the prediction matched the ungapped seed sequence not used for building the profile.

To benchmark the overall performance of the prediction, we ran a number of lanthipeptide biosynthetic gene clusters through antiSMASH. We checked if the gene cluster was identified, the precursor peptide was detected, and finally the peptide mass was predicted correctly. Among the clusters run for benchmarking, we included the planosporicin and epilancin 15X clusters. The cleavage sites of both of these lanthipeptides were not part of the seed alignments.

Discussion

Only few tools are currently available that allow the automated identification of RiPPs. Apart from antiSMASH, there is BAGEL, recently released in version 3 [11]. BAGEL targets a large number of

different ribosomally synthesised peptides. For lanthipeptides, BAGEL only predicts the leader peptide and the class, but does not attempt to predict tailoring reactions, number of Lan and MeLan bridges or the molecular weight.

Determining the lanthipeptide class from the biosynthetic enzymes in the cluster is straightforward, and antiSMASH performs this tasks flawlessly on the benchmark data set (Table 4). Predicting the core peptide sequence is more difficult. The cleavage site motif of class II lanthipeptides (Table 3) is relatively uniform, largely consisting in two amino acids with small side chains that are preceded by alternating hydrophobic and hydrophilic residues. In fact, all the four of the 21 class II cleavage sites incorrectly predicted during validation (Table 5) differ from this pattern and contain a site that more closely matches the motif upstream of the actual cleavage site. Some class II core peptides like mersacidin [12] or lichenicidin A2 [13] lose an additional six amino acids at the N-terminus after the proteolytic cleavage, so it seems likely that the predicted cleavage sites may be accurate and some additional enzyme catalyses the N-terminal modifications. Class I leader peptides also carry a short motif of alternating hydrophobic and hydrophilic amino acids, usually called the `FNLD` motif. The spacer between this motif and the actual cleavage site varies. At position -2 in front of the cleavage site, many leader peptides carry a proline residue (Table 2). During validation, all cleavage sites were predicted correctly (Table 5). Due to the strong signal of the `FNLD` motif, class I prediction (stability 100 %) is even more robust than the class II prediction (stability 81 %) with the shorter motif. As a proof of concept, we used the gene clusters of the recently published planosporicin [14] and the lactate-containing epilancin 15X [15]. Both precursor peptides contain cleavage sites that are distinct from all the sequences included in the class I seed alignment. For both lanthipeptides the algorithm is able to correctly predict the mass, number of Lan/MeLan bridges and tailoring modifications.

The detection of enzymes responsible for tailoring reactions is central in the prediction of the mature peptide mass. antiSMASH correctly predicts the AviCys residues present in epidermin and microbisporicin, the two amino-vinylated peptides in the benchmark data-set (Table 4). The halogenation and hydroxylation of microbisporicin is also detected. A remaining issue in mass prediction is that not all Ser and Thr residues are dehydrated in all the peptides, resulting in mass predictions that are 18 Da too low per undehydrated amino acid. antiSMASH assists in detecting the presence of undehydrated residues by providing alternative mass predictions for lanthipeptides that carry more Ser and Thr than Cys residues.

Once the tailoring reactions have been predicted, the final step is the prediction of the number of

Lan and MeLan bridges. The naïve heuristic of counting the Cys and Ser/Thr residues and then using the smaller number fails if the mature peptide contains an AviCys or AviMeCys residue and needs to be adjusted accordingly. Using the advanced heuristic, antiSMASH correctly predicts the number or bridges in almost all residues of the benchmark data-set (Table 4). The heuristic only fails if two Cys residues form a disulphide bridge, a rare occurrence observed in e.g. thermophilin 1277 [16]. Unfortunately, the enzyme catalysing the formation of the disulphide bridge is not present on the gene cluster and thus can not be used to predict a disulphide bridge formation.

After all prediction steps are completed, the prediction details will be annotated into the antiSMASH output. For the HTML output (Figure 1), lanthipeptide class and leader / core peptide split predicted are shown in the "detailed annotation" section of the cluster page. The score of the class prediction, predicted monoisotopic mass and molecular weights, the number of bridges and the identified additional modifications are shown in the "prediction details" sidebar.

Conclusions

With the algorithm described in this paper, antiSMASH gains extensive lanthipeptide-specific predictive capabilities. antiSMASH is the only software currently available that will predict lanthipeptide class, core peptide cleavage, tailoring reactions, number of Lan and MeLan bridges, and the molecular weight of the mature peptide product.

List of abbreviations

antiSMASH antibiotics and secondary metabolite analysis shell

AviCys *S*-[(*Z*)-2-aminovinyl]-D-cysteine

AviMeCys *S*-[(*Z*)-2-aminovinyl]-(3*S*)-3-methyl-D-cysteine

Cys cysteine

Dha dehydroalanine

Dhb dehydrobutyrine

HMM	Hidden Markov Model
Lan	lanthionine
MeLan	methyllanthionine
NRPS	non-ribosomal peptide synthase
PKS	polyketide synthase
RiPP	ribosomally synthesised and post-translationally modified peptide
Ser	serine
Thr	threonine

Acknowledgments

Funding was provided by the German Ministry of Education and Research (BMBF) (Grant 0315585A to TW), German Centre for Infection Research (DZIF) (8000-402-2 to TW and WW), LAPTOP (Grant 245066 to WW). Funding for open access charge: Deutsche Forschungsgemeinschaft (DFG) and Open Access Publishing Fund of Tübingen University.

References

1. Knerr PJ, van der Donk WA (2012) Discovery, biosynthesis, and engineering of lantipeptides. Annual Review of Biochemistry 81: 1–27.

2. Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, et al. (2013) Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Natural Product Reports 30: 108-160.

3. Sit CS, Yoganathan S, Vederas JC (2011) Biosynthesis of aminovinyl-cysteine-containing peptides and its application in the production of potential drug candidates. Accounts of Chemical Research 44: 261–268.

4. Kellner R, Jung G, Hörner T, Zähner H, Schnell N, et al. (1988) Gallidermin: a new lanthionine-containing polypeptide antibiotic. European Journal of Biochemistry 177: 53–59.

5. Foulston LC, Bibb MJ (2010) Microbisporicin gene cluster reveals unusual features of lantibiotic biosynthesis in actinomycetes. Proceedings of the National Academy of Sciences 107: 13461–6.

6. Heidrich C, Pag U, Josten M, Metzger J, Jack RW, et al. (1998) Isolation, characterization, and heterologous expression of the novel lantibiotic epicidin 280 and analysis of its biosynthetic gene cluster. Applied and Environmental Microbiology 64: 3140–3146.

7. Medema MH, Blin K, Cimermancic P, de Jager V, Zakrzewski P, et al. (2011) antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Research 39: W339–W346.

8. Blin K, Medema MH, Kazempour D, Fischbach MA, Breitling R, et al. (2013) antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Research 41: W204–W212.

9. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, et al. (2012) The Pfam protein families database. Nucleic Acids Research 40: D290–D301.

10. Durbin R, Eddy SR, Krogh A, Mitchison G (1998) Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge University Press.

11. van Heel AJ, de Jong A, Montalbán-López M, Kok J, Kuipers OP (2013) BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Research 41: W448–W453.

12. Altena K, Guder A, Cramer C, Bierbaum G (2000) Biosynthesis of the lantibiotic mersacidin: organization of a type b lantibiotic gene cluster. Applied and Environmental Microbiology 66: 2565–2571.

13. Dischinger J, Josten M, Szekat C, Sahl HG, Bierbaum G (2009) Production of the novel two-peptide lantibiotic lichenicidin by *Bacillus licheniformis* DSM 13. PloS ONE 4: e6788.

14. Sherwood EJ, Bibb MJ (2013) The antibiotic planosporicin coordinates its own production in the actinomycete *Planomonospora alba*. Proceedings of the National Academy of Sciences 110: E2500–E2509.

15. Velásquez JE, Zhang X, van der Donk WA (2011) Biosynthesis of the antimicrobial peptide epilancin 15x and its N-terminal lactate. Chemistry & Biology 18: 857–867.

16. Kabuki T, Uenishi H, Seto Y, Yoshioka T, Nakajima H (2009) A unique lantibiotic, thermophilin 1277, containing a disulfide bridge and two thioether bridges. Journal of Applied Microbiology 106: 853–862.

17. Birri DJ, Brede DA, Nes IF (2012) Salivaricin D, a novel intrinsically trypsin-resistant lantibiotic from *Streptococcus salivarius* 5M6c isolated from a healthy infant. Applied and Environmental Microbiology 78: 402–10.

18. Wirawan RE, Klesse NA, Jack RW, Tagg JR (2006) Molecular and genetic characterization of a novel nisin variant produced by *Streptococcus uberis*. Applied and Environmental Microbiology 72: 1148–1156.

19. Zendo T, Fukao M, Ueda K, Higuchi T, Nakayama J, et al. (2003) Identification of the lantibiotic nisin Q, a new natural nisin variant produced by *Lactococcus lactis* 61-14 isolated from a river in Japan. Bioscience, Biotechnology, and Biochemistry 67: 1616–1619.

20. Fuchs SW, Jaskolla TW, Bochmann S, Kötter P, Wichelhaus T, et al. (2011) Entianin, a novel subtilin-like lantibiotic from *Bacillus subtilis* subsp. *spizizenii* DSM 15029T with high antimicrobial activity. Applied and Environmental Microbiology 77: 1698–1707.

21. Garg N, Tang W, Goto Y, Nair SK, van der Donk WA (2012) Lantibiotics from *Geobacillus thermodenitrificans*. Proceedings of the National Academy of Sciences 109: 5241–5246.

22. Wescombe PA, Tagg JR (2003) Purification and characterization of streptin, a type a1 lantibiotic produced by *Streptococcus pyogenes*. Applied and Environmental Microbiology 69: 2737–2747.

23. Castiglione F, Lazzarini A, Carrano L, Corti E, Ciciliato I, et al. (2008) Determining the structure and mode of action of microbisporicin, a potent lantibiotic active against multiresistant pathogens. Chemistry & Biology 15: 22–31.

24. Novak J, Caufield PW, Miller EJ (1994) Isolation and biochemical characterization of a novel lantibiotic mutacin from *Streptococcus mutans*. Journal of Bacteriology 176: 4316–4320.

25. Wescombe PA, Upton M, Dierksen KP, Ragland NL, Sivabalan S, et al. (2006) Production of the lantibiotic salivaricin a and its variants by oral streptococci and use of a specific induction assay to detect their presence in human saliva. Applied and Environmental Microbiology 72: 1459–1466.

26. Hynes WL, Ferretti J, Tagg J (1993) Cloning of the gene encoding Streptococcin A-FF22, a novel lantibiotic produced by *Streptococcus pyogenes*, and determination of its nucleotide sequence. Applied and Environmental Microbiology 59: 1969–1971.

27. Ryan MP, Jack RW, Josten M, Sahl HG, Jung G, et al. (1999) Extensive post-translational modification, including serine to D-alanine conversion, in the two-component lantibiotic, lacticin 3147. Journal of Biological Chemistry 274: 37544–37550.

28. Hyink O, Wescombe PA, Upton M, Ragland N, Burton JP, et al. (2007) Salivaricin A2 and the novel lantibiotic salivaricin B are encoded at adjacent loci on a 190-kilobase transmissible megaplasmid in the oral probiotic strain *Streptococcus salivarius* K12. Applied and Environmental Microbiology 73: 1107–1113.

29. Widdick D, Dodd H, Barraille P, White J, Stein T, et al. (2003) Cloning and engineering of the cinnamycin biosynthetic gene cluster from *Streptomyces cinnamoneus cinnamoneus* DSM 40005. Proceedings of the National Academy of Sciences 100: 4316–4321.

30. Herzner AM, Dischinger J, Szekat C, Josten M, Schmitz S, et al. (2011) Expression of the lantibiotic mersacidin in *Bacillus amyloliquefaciens* FZB42. PLoS ONE 6: e22389.

31. Boakes S, Cortés J, Appleyard AN, Rudd BA, Dawson MJ (2009) Organization of the genes encoding the biosynthesis of actagardine and engineering of a variant generation system. Molecular Microbiology 72: 1126–1136.

32. Holtsmark I, Mantzilas D, Eijsink V, Brurberg M (2006) Purification, characterization, and gene sequence of michiganin a, an actagardine-like lantibiotic produced by the tomato pathogen *Clavibacter michiganensis* subsp. *michiganensis*. Applied and Environmental Microbiology 72: 5814–5821.

Figure Legends

Figure 1. Example lanthipeptide output antiSMASH 2.1 output for the microbisporicin [5] gene cluster, showing the predicted leader/core peptide split and the predicted tailoring reactions and weights in the sidebar.

Tables

Table 1. Lanthipeptide-related HMM profiles and scores

Name	Description	Cutoff	File
LANC_like	LanC-like lantibiotics biosynthesis protein	17	LANC_like.hmm
DUF4135	Lantibiotic-associated domain	150	PF13575.hmm
Lant_dehyd_N	Lantibiotic dehydratase, N-terminus	20	Lant_dehyd_N.hmm
Lant_dehyd_C	Lantibiotic dehydratase, C-terminus	20	Lant_dehyd_C.hmm
Flavoprotein	Lantibiotic aminovinly flavoprotein	20	PF02441.hmm
Trp_halogenase	Tryptophan halogenase	20	PF04820.hmm
p450	P450 oxygenase	60	PF00067.hmm
Pkinase	Protein kinase domain	30	PF00069.hmm
adh_short	Short-chain dehydrogenase	100	PF00106.hmm
adh_short_C2	Short-chain dehydrogenase, C-terminus	100	PF13561.hmm
Antimicr18	Lantibiotic antimicrobial peptide 18	20	Antimicrobial18.hmm
Gallidermin	Gallidermin	20	Gallidermin.hmm
L_biotic_A	Lantibiotic, type A	20	L_biotic_typeA.hmm
TIGR03731	Lantibiotic, gallidermin/nisin family	18	TIGR03731.hmm
leader_d	Lantibiotic leader lacticin 481 group	20	LE-LAC481.hmm
leader_eh	Lantibiotic leader mersacidin cinnamycin group	20	LE-MER+2PEP.hmm
leader_abc	Lantibiotic leader LanBC modified	20	LE-LanBC.hmm
mature_d	Lantibiotic peptide lacticin 481 group	20	MA-LAC481.hmm
mature_ab	Lantibiotic peptide nisin epidermin group	20	MA-NIS+EPI.hmm
mature_a	Lantibiotic peptide nisin group	20	MA-NIS.hmm
mature_b	Lantibiotic peptide epidermin group	20	MA-EPI.hmm
mature_ha	Lantibiotic peptide two component alpha	20	MA-2PEPA.hmm
mature_h_beta	Lantibiotic peptide two component beta	20	MA-2PEPB.hmm
lacticin_l	lantibiotic peptide lacticin 481 group (dufour et al)	20	LE-DUF.hmm
lacticin_mat	lantibiotic leader lacticin 481 group (dufour et al)	20	MA-DUF.hmm
LD_lanti_pre	FxLD family lantipeptide	20	TIGR04363.hmm
strep_PEQAXS	Streptomyces PEQAXS motif lantipeptide	20	strep_PEQAXS.hmm

A list of the lanthipeptide-related HMM profiles

Table 2. Class I cleavage site motif sequences

Name	Position	Sequence
mutacin_1140	22...41	FAFDTTDTTIVASNDDPDTR
mutacin_Ny266	22...41	FTFDTTDTIVAESNDDPDTR
salivaricin_D	6...23	FNLDLVEVSK--SNTGASAR
nisin_U	6...24	FNLDLIKISK-ENNSGASPR
nisin_A	6...23	FNLDLVSVSKK--DSGASPR
nisin_Z	6...23	FNLDLLSVSKK--DSGASPR
nisin_Q	6...23	FNLDLVSVSKT--DSGASTR
gallidermin	11...30	FDLDVKVNAKESNDSGAEPR
epidermin	11...30	FNLDVKVNAKESNDSGAEPR
entianin	7...24	FDLDVVKVSKQ--DSKITPQ
Pep5	8...26	FDLEIKKETSQNTD-ELEPQ
epicidin_280	8...26	FDLEIKKDNME-NNNELEPQ
epilancin_K7	6...24	FDLNLNKGVETQK-SDLSPQ
geobacillin_I	7...23	FDLDIVVK-KQ--DDVVQPN
streptin	8...23	FDLDLKTNKK---D-TATPY
microbisporicin	17...33	LDLDLSIGVEE---ITAGPA

Sequences used to create the class I cleavage site motif

Table 3. Class II cleavage site motif sequences

Name	Position	Sequence
mutacin_II	19...52	EL-TILGG
variacin	15...47	ELDAILGG
salivaricin_A	22...51	ELMEVAGG
butyrivibriocin	16...48	ELEQILGG
streptococcin_A_FF22	20...51	ELDNLLGG
lichenicidin_A1	30...74	EQHSIAGG
lichenicidin_A2	27...72	ELKALVGG
thermophilin_1277	18...66	ELEMLIGG
lacticin_A1	16...59	FDEDVFGA
lacticin_A2	26...65	EGDESHGG
nukacin_KQ_131	23...57	ELNEVLGA
macedocin	18...51	ELDQIIGA
salivaricin_B	24...56	ELDNVLGA
haloduracin_A1	33...69	ILAGVNGA
haloduracin_A2	28...65	ELSSLAGS
cytolysin	12...68	EMEAIQGS
plantaricin_W	24...59	NLLNVNGA
cinnamycin	52...59	IAATEAFA
mersacidin	41...48	QMDKLVGA
actagardine	35...64	EDRTIYAA
michiganin_A	36...66	RRVVSPYM

Sequences used to create the class II cleavage site motif

Table 4. Benchmark results

Substance	Class	Predicted Mass (Da)	Actual Mass (Da)	# bridges	Source
Salivaricin D	I	3466.7	3467.5	4	[17]
Nisin U	I	3029.6	3029.0	5	[18]
Nisin A	I	3353.9	3354.5	5	[19]
Nisin Z	I	3330.9	3331.5	5	[19]
Nisin Q	I	3326.9	3327.3	5	[19]
Gallidermin	I	2164.0	2164	4	[4]
Epidermin	I	2164.0	2164	4	[4]
Entianin	I	3346.7	3346	5	[20]
Pep5	I	3487.1	3488	3	[4]
Epicidin 280	I	3135.6	3135	3	[6]
Geobacillin I	I	3261.5	3265	7	[21]
Streptin 1´	I	2441.9	2442	3	[22]
Microbisporicin A2	I	2232.4	2232	5	[23]
Mutacin II	II	3243.5	3244	3	[24]
Salivaricin A2	II	2366.6	2368	3	[25]
Streptococcin A-FF22	II	2796.1	2795	3	[26]
Lichencidin A1	II	3250.7	3251	4	[13]
Lichencidin A2	II	3632.8†	3021	4	[13]
Thermophilin 1277	II	3395.9‡	3428	2‡	[16]
Lacticin 3147 A1	II	3322.6	3322.3	4	[27]
Lacticin 3147 A2	II	2843.2	2847.5	3	[27]
Salivaricin B	II	2733.1	2740	3	[28]
Cinnamycin	II	2043.2	2041	3	[29]
Mersacidin	II	2399.0†	1826.3	3	[30]
Actagardine	II	1856.2	1860.5	4	[31]
Michiganin A	II	2145.5	2145	4*	[32]
Planosporicin	I	2193.3	2194	5	[14]
Epilancin 15X	I	3171.8	3171.7	3	[15]

†N-terminal removal of six amino acids not predicted

‡Contains a disulphide bridge

*Not shown experimentally

Benchmark of the antiSMASH lanthipeptide predictor

Table 5. Stability of the prediction motifs

Class	# Sequences	# Found	% Found	# Correct	% Correct
I	16	16	100	16	100
II	21	21	100	17	81

Stability of the prediction motifs

2 Introduction

2.1 Bioinformatics Approaches

2.1.1 Definition

When the term "bioinformatics" was coined initially (Hesper and Hogeweg 1970), it was meant to represent the study of informatic processes in biological systems – for a review, see (Hogeweg 2011). With the advent of large-scale shotgun sequencing projects, the meaning bioinformatics was frequently reduced to computational handling and analysis of genomic data. To avoid this narrow classification, the alternative term "computational biology" has emerged to represent all biological research performed with the aid of computational methods. In recent years, different *-omics* disciplines and the connection of the data produced have broadened the "bioinformatics" definition again. The US National Institute of Health defines bioinformatics as the "*[r]esearch, development, or application of computational approaches for expanding the use of biological [...] data, including those to acquire, store, organize, archive, analyze, or visualize such data*", and computational biology as "*[t]he development and application of data-analytical and theoretical methods, mathematical modeling and computational simulation techniques to the study of biological [...] systems*" (Huerta et al. 2000). Drawing a clear border between these definitions is difficult, and for this work, I will simply be using "bioinformatics" as a synonym for "computational biology", meaning the research and development of computational applications of data-analytical methods to acquire, study, store, or visualize biological data or systems.

2.1.2 History

Even before DNA sequencing was invented in the mid-seventies of the last century, people had started using computers to help answer biological questions. When Needleman and Wunsch (1970) pioneered sequence comparison and alignment algorithms, they were working on amino acid sequences. The concepts of the algorithms were easily adaptable to nucleotide sequences, and once DNA sequence data became available, the Needleman–Wunsch algorithm was quickly adapted. Beyond alignment, early uses of bioinformatics include the prediction of RNA secondary structure (Nussi-

nov and Jacobson 1980) and building phylogenies (Dayhoff and Eck 1969).
In the late nineties, building on the technological advances brought by the Human Genome project, sequencing whole genomes became a possibility. At first, the small length of sequencing reads made assembly of large genomes difficult, but later generations of the sequencing technologies gave larger and larger read sizes. Around 2008, the improvements to sequencing technology reached a level where sequence sizes were growing exponentially faster than the computing capacity to process them (Wetterstrand 2013). The field of bioinformatics is still adapting to this "data deluge" (Bell et al. 2009).

2.1.3 From Gene to Product, Bioinformatic Steps

Following the data from sequencing via genes to gene products, bioinformatics is involved in every step of the pipeline. After a sequencing run, short DNA snippets called "reads" need to be assembled into a contiguous sequence. This is done by trying to find the shortest common superstring of the overlapping sequence reads, while accounting for sequencing errors. As a mathematically optimal solution would require a prohibitive amount of computation, approximations and heuristics are employed to lower the time complexity of assemblies (see e.g. Chevreux 2005).
Once the genome sequence is assembled, the next step is annotating it. Usually this focuses on gene finding, but additional features like RNAs, promotors, ribosomal binding sites, secondary structure elements, and intron/exon layout also can be identified in this step. During gene finding, the usual approach is to identify the set of maximal non-overlapping open reading frames (ORFs) (see Majoros et al. 2004; Delcher et al. 2007). Frequently, those ORFs are then compared to databases of known genes to identify the correct start codons (e.g. Blattner et al. 1997). While annotations from related organisms might provide additional clues, relying on those annotations also increases the risk of propagating and cementing annotation errors (Richardson and Watson 2013).
Having identified genes, the next step frequently is functional annotation. By comparing translated protein sequences against databases of conserved proteins or protein domains (like PFAM, Punta et al. 2012), one can derive the function of a gene product. Of course the risk mentioned by Richardson and Watson (2013) again applies, basing automated annotations of function on automated annotations of func-

tion carries the risk of propagating annotation errors.

2.2 Genome Mining

In many prokaryotes, enzymes that act together to perform their function are also clustered together on the genome. This feature has been utilised since the early days of genetic engineering in secondary metabolite producers, e.g. to isolate secondary metabolite clusters onto plasmids that can then be transferred into heterologous hosts (Malpartida and Hopwood 1984). Locating secondary metabolites on plasmids also helped in early sequencing attempts, like the sequencing of the nisin structural gene (Kaletta and Entian 1989). Still, in the early days of genome mining, a lot of the methods depended on hybridisation experiments with plasmids carrying random fragments of genomic DNA (Chinault and Carbon 1979), and then using primers derived from the hybridisation sequences.

When the genome of an organism is available, different approaches become feasible. The first bacterial genome, *Haemophilus influenzae*, was sequenced using whole-genome shotgun sequencing (Fleischmann et al. 1995) in 1995, followed by the *Escherichia coli* K12 genome (Blattner et al. 1997) in 1997. To date, the GenBank database (Benson et al. 2013) contains 2666 completed and more than 12 000 partially sequenced and assembled prokaryotic genomes. Especially the latter number is expected to rise sharply over the next years. Data from the US National Human Genome Research Institute show that starting in January 2008, sequencing costs per raw megabase of DNA has dropped drastically (Wetterstrand 2013), making it feasible to sequence more and more organisms of interest.

The genome of the model Actinomycete *Streptomyces coelicolor* was sequenced in 2002 (Bentley et al. 2002). In addition to the three previously identified secondary metabolite gene clusters – the blue polyketide actinorhodin, the red oligopyrrole prodiginine, and the non-ribosomal peptide CDA – the *S. coelicolor* genome contained an additional 18 characteristic secondary metabolite clusters. Based on the genetic information about the clusters, a number of the additional secondary metabolites have since been elucidated (Bentley et al. 2002). Similar observations were made in the sequencing of *Streptomyces avermitilis* (Ikeda et al. 2003) and *Streptomyces griseus* (Ohnishi et al. 2008).

Walsh and Fischbach (2010) assumed that based on the knowledge gap between sec-

ondary metabolite clusters with known products from Streptomycetes and the number of cryptic and predicted clusters found on their genomes, we are missing about 90% of the secondary metabolite production capability. Even if only a small part of the cryptic gene clusters actually produces novel molecules, elucidating those clusters' biosynthetic steps will easily double the knowledge on biosynthetic processes.

2.3 Natural Products

Natural products are small molecules produced by living organisms. Many natural products have unusual chemical folds and show antibacterial, antifungal, antiparasitical, anticancer or immunosuppressive activities. After Alexander Fleming's discovery of penicillin (Fleming 1929), an antibacterial compound produced by a *Penicillium*, microorganisms have been in the focus of natural product research. When Waksman and coworkers isolated streptomycin (Schatz et al. 1944) from *Streptomyces griseus*, the group of Actinobacteria began to get increased attention. Streptomycetes and other Actinomycetales indeed have a great biosynthetic potential. About two thirds of the antibacterial compounds in use in the clinics today are produced by Actinomycetales (Demain 1999). Many different classes of natural product biosynthesis pathways have been identified. In the following, the three classes most relevant for this work will be introduced.

Polyketides

A big and important class are the polyketides. Interest in polyketide synthesis had started early last century, but the biosynthesis pathways continued to be elusive (for a historical overview, see Staunton and Weissman 2001). Polyketides are assembled from simple building blocks by polyketide synthases (PKSs). The building blocks usually are malonic acid derivatives coupled with coenzyme A. Polyketides are synthesised in a process similar to the fatty acid synthesis. Depending on the genetic and enzymatic organisation, different types of PKSs are distinguished, again in analogy to the fatty acid biosynthesis. Type I PKSs carry the different enzymatic domains covalently linked on large multifuntional proteins. In contrast, type II PKSs carry multiple discrete proteins catalysing one function each.
In the case of macrolide type I PKSs like the erythromycin synthase, multiple mod-

ules perform subsequent synthesis steps (Donadio et al. 1991; Bevitt et al. 1992). Each module contains at least the three enzymatic domains (keto synthase (KS), acyl transferase (AT) and acyl carrier protein (ACP)) required for a single chain extension step by Claisen condensation (Claisen 1887). Additional domains positioned between the AT and ACP domains can catalyse additional keto group modificiations (e.g. the reduction by a keto reductase (KR), dehydratiation by a dehydratase (DH) and further reduction by an enoyl reductase (ER)). In front of the first module, a starter module usually consisting of an AT and an ACP domain load and activate the starter substrate. Many PKSs contain a C-terminal thioesterase (TE) domain releasing the polyketide chain from the synthase. As the extender units forming the backbone of the polyketide are selected by the AT domains, it is possible to derive the structure of the polyketide backbone from substrate specificity predictions of the AT domains, the reductive domains and the order of those domains. As an exception to these canonical modules, some bacteria also contain so-called *trans*-AT PKSs, where the AT domain is not encoded in the same gene product as the rest of the PKS module but instead is located on a separate gene product, this occurs e.g. in the kirromycin gene cluster (Weber et al. 2008). Apart from the different location of the AT domain, *trans*-AT PKSs work similar to the canonical *cis*-AT PKSs (Piel 2010).

In contrast, type II PKSs like the actinorhodin synthase only contain a single module that iteratively builds the polyketide product (Malpartida and Hopwood 1984). This module is encoded on discrete enzymes carrying a KS_α, a KS_β, and an ACP domain respectively. The KS_α enzyme resembles the fatty acid keto synthase domains. The KS_β is thought to be responsible for controling the chain length (Tang et al. 2003).

Type III PKSs perform similar condensation reactions, but their domain setup and mode of action greatly differs from type I and II PKSs. Unlike the latter, type III PKSs directly utilise substrates and percursors, without the help of an ACP domain. This is possible because the whole set of condensation, cyclisation and aromatisation reactions is carried out in a single active site (Ferrer et al. 1999).

Nonribosomal Peptides

Nonribosomal peptides are polypeptides that are not of ribosomal origin but instead are synthesised by modular megaenzymes called nonribosomal peptitde synthetases (NRPSs) (for a review, see Schwarzer et al. 2003). They are not limited to the 20

proteinogenic amino acids but instead can contain non-proteinogenic amino acids such as hydroxyphenyl-glycine, dihydroxyphenyl-glycine (e.g. in balhimycin, Pelzer et al. 1999) or ornithine. So far, over 500 amino acids have been identified (Caboche et al. 2008).

Following similar principles as the type I PKSs, nonribosomal peptitde synthetases consist of modules carrying three domains responsible for synthesising the peptide backbone. The substrate amino acid integrated into the non-ribosomal peptide is selected and activated by the adenylation (A)-domain. The activated amino acid is then covalently bound to the peptidyl carrier protein (PCP), which takes care of transporting the amino acid residue to a condensation (C)-domain forming the peptide bond during chain elongation. The common organisation of these domains in a module usually is C–A–PCP, with an exception of the initiation module that does not need to perform a chain elongation step and thus only carries the A–PCP domains. Analogous to type I PKSs, a C-terminal thioesterase (TE)-domain releases the polypeptide chain, frequently catalysing a macrocyclisation.

Deviations from the C–A–PCP pattern are possible and usually indicate additional modifications. Epimerisations from L- to D-amino acids are catalysed by epimerisation (E)-domains, heterocyclisation (Cy)-domains can replace C-domains and introduce heterocyclisations, and methyltransferase-domains methylate amino acid side chains. In addition to tailoring domains integrated into the NRPS macromolecule, many NRPS gene clusters also contain standalone modification enzymes like P450 oxygenases, halogenases and methyltransferases.

The peptide backbone synthesised by NRPSs can be derived from the substrate specificity and the order of the A-domains. Unfortunately, the overall sequence similarity of A-domains with different substrate specifities in an NRPS cluster does not allow for a simple sequence-based approach to determining the substrate specificity. After obtaining a crystal structure of an A-domain of the gramicidin S synthase (Conti et al. 1997), it was possible to investigate the substrate binding pocket and to propose first models for predicting the substrate specificity based on the residues contained in the active site (Stachelhaus et al. 1999; Challis et al. 2000).

Lanthipeptides

Lanthipeptides are ribosomally synthetised and post-translationally modified peptides (RiPPs) undergoing extensive post-translational modifications. These modifications increase the stability of the peptide product against heat stress and proteolysis. The name lanthipeptide is derived from the characteristic lanthionine (Lan) and methyl-lanthionine (MeLan) residues contained in the processed peptide product. The best-known members of this secondary metabolite type – like nisin (Kaletta and Entian 1989) – possess an antibacterial activity and are called lantibiotics for "lanthionine-containing antibiotics". The signature Lan and MeLan residues are introduced in a two-step reaction, first dehydrating serine (Ser) and threonine (Thr) to dehydroxyalanine (Dha) and dehydroxybutyrine (Dhb), respecively. A Michael-type addition by cysteine (Cys) residues onto the dehydro amino acids yields the thioether cross-links in the second step.
Lanthipeptides are classified by biosynthetic enzymes catalysing these two steps (Table 1) and to date, four different classes are known (Knerr and van der Donk 2012).

Class	# Core Enzymes	Enzyme Name	Zn Binding Site
I	2	LanB & LanC	yes
II	1	LanM	yes
III	1	LanKC	no
IV	1	LanL	yes

Table 1: Lanthipeptide core biosynthesis enzyme properties

In lanthipeptides of class I, two separate enzymes catalyse the two steps: Dehydration is carried out by a dehydratase designated LanB. Cyclisation then is performed by a cyclase called LanC. To derive cluster-specific names from these generic names, the "Lan"-part is usually replaced by a cluster-specific identifier: e.g. NisB and NisC for the nisin gene cluster (Kuipers et al. 1993). In class II, III and IV lanthipeptides, bi-functional enzymes perform both the dehydration and condensation reactions. An enzyme called LanM in class II lanthipeptides carries an N-terminal dehydratase domain and a C-terminal cyclisation domain resembling class I LanC enzymes. The class III and IV bi-functional enzymes are very similar in their domain organisation. They carry an N-terminal phosphoserine/phosphothreonine lyase domain, a central ki-

nase domain and a C-terminal cyclase domain. In class III clusters the enzyme is called LanKC, in class IV clusters it is called LanL. LanKC enzymes lack three zinc-binding residues in the C-terminal cyclase domain that are conserved in the other three cluster type cyclases.

In addition to the core biosynthetic enzymes, further modifications can be introduced to form the finished lanthipeptide. Enzymes responsible for the formation of *S*-[(*Z*)-2-aminovinyl]-D-cysteine and *S*-[(*Z*)-2-aminovinyl]-(3*S*)-3-methyl-D-cysteine cyclisations (Sit et al. 2011) have been identified and given the generic designation LanD. A flavin-dependent tryptophan halogenase designated LanH is able to chlorinate tryptophan, as observed in the microbisporicin biosysnthesis (Foulston and Bibb 2010). A cytochrome P450 oxygenase designated LanO is able to synthesise (di)hydroxyproline, also observed in the microbisporicin gene cluster (Foulston and Bibb 2010). In core peptides carrying an N-terminal Dha residue, a short-chain dehydrogenase of the EciO type can convert the Dha residue into a lactate (Heidrich et al. 1998). A more detailed description of the enzyme classes responsible for lanthipeptide biosynthesis can be found in (Blin et al. 2013a).

3 Goals

1. The goal of this work was to create an easy to use, comprehensive secondary metabolite analysis software.

   Various software tools were available to perform predictions for specic types of secondary metabolite clusters. Most of these tools focus on the prediction of NRPS and PKS clusters. This study introduces the first analysis tool that has the aim to predict many different secondary metabolite gene clusters: antiSMASH. A central aim of antiSMASH is, besides covering a wide range of secondary metabolites, to allow easy access to both input and ouput data for the target audience of wet lab natural product researchers. Also covered in this study is the work on antiSMASH 2.0, which further improves both predictive capabilities and the usability of the antiSMASH pipeline.

2. The substrate specificity predictions provided by NRPSPredictor were improved.

   The initial release of NRPSPredictor (Rausch et al. 2005) in 2005 constructed support vector machine (SVM) models from the available NRPS A-domain sequences and their associated substrate specificities. In the six years since that publication, a large number of new A-domains have been identified and published. This study introduces an improved set of SVM models for the NRPSPredictor2 software.

3. An algorithm capable of predicting the post-translational modifications of lanthipeptide precursor peptides was designed and implemented.

   To the best of the author's knowlede, no published software tool is available to identify lanthipeptide gene clusters and then predict post-translational modifications to the precursor peptides based on the cluster layout. Predicting the post-translational modifications is central in assessing the molecular weight of the processed peptide product, which in turn is important for identifying this product in culture extract via mass spectrometric approaches. This study provides the first such software tool, based on novel prediction algorithms and fully integrated into the antiSMASH pipeline.

4 Results & Discussion

4.1 antiSMASH 1

antiSMASH – the antibiotics and secondary metabolite analysis shell – is a software pipeline designed to predict secondary metabolite gene clusters from genomic DNA sequences. While other software tools have been published aiming at specific secondary metabolite types (e.g. Ansari et al. 2004; Kamra et al. 2005; Rausch et al. 2005; Caboche et al. 2008; Weber et al. 2009), antiSMASH is the first Open Source Software pipeline to support a large number of secondary metabolite types (18 types in antiSMASH 1 (Medema et al. 2011), 24 types in antiSMASH 2 (Blin et al. 2013b)).

Users can either upload a DNA sequence in GenBank, EMBL or FASTA format, or specify an NCBI ID identifying the sequence to download from the GenBank (Benson et al. 2013) database. If no genes are annotated in the sequence record, (e.g. the input was a FASTA file), antiSMASH will automatically run a gene finding step (Glimmer (Delcher et al. 2007) for prokaryotic inputs, GlimmerHMM (Majoros et al. 2004) for eukaryotic inputs). In a next step, protein sequences for the annotated genes are compared to a manually curated database of secondary metabolite specific signature profiles using HMMer (Eddy 2011). The signature profiles are taken from the PFAM database (Punta et al. 2012) or generated from seed alignments specifically for antiSMASH. A set of secondary metabolite cluster rules defines which biosynthetic domains need to be present for biosynthesis to occur and the overall size range of the cluster. Cluster rules can range from simple rules (phenazine clusters are identified by a single hit against the phzB domain profile) via combination rules (type I PKS clusters are identified by a hit against the KS profile in proximity to a hit against the AT profile) to complex rules (oligosaccharide clusters are identified by hitting at least three different profiles out of a set of six glycosyltransferase-like profiles). Secondary metabolite clusters are annotated in the output when all the protein domains are identified within the defined distance range on the input genome.

Once secondary metabolite clusters are identified, cluster-specific analyses are run for NRPS and PKS type I clusters. For both these types, the domain structure of their modular biosynthetic megaenzymes is analysed. Substrates of PKS AT domains are predicted using an active site signature sequence consisting of 24 amino acids (Yadav et al. 2003) and a profile Hidden Markov Model (pHMM) based approach (Minowa

et al. 2007). Substrates of NRPS A domains are predicted using both a pHMM based approach and a SVM based approach (Röttig et al. 2011). Using the individual domain predictions from the different tools, a consensus prediction is generated for every AT/A domain. From these consensus predictions and the order of the domains in the synthase, the core polyketide/polypeptide sequence is predicted and visualised in a picture file.

A secondary metabolite cluster does not only consist of the genes encoding for the core biosynthetic enzymes. In order to aid in the annotation of the accessory genes surrounding the signature genes, antiSMASH contains four additional analysis steps. ClusterBlast uses a manually curated set of all secondary metabolite gene clusters identified from the NCBI *nt* database. Using NCBI BLAST$^+$ (Camacho et al. 2009), secondary metabolite clusters identified in the user's input sequence are compared to the gene clusters in our database. The ten most similar clusters from the database are reported in the results.

The same database was used to construct secondary metabolite clusters of orthologous groups (smCOGs), an evolutionary classification system in the spirit of the established cluster of orthologous groups (COG) classification system (Tatusov et al. 2003). Additionally, a whole-genome search against the PFAM database and all bacterial/fungal genomes from the GenBank database is offered to improve annotations at the cost of additional run-time.

4.2 antiSMASH 2

After the publication of antiSMASH 1, the development methodology was changed from the rapid prototyping approach taken with the first version to a more formalised development process better suited for future sofware maintenance. While the goal still was to add new features such as novel predictors or secondary metabolite types, it was also important to ensure that new code did not interfere with existing predictors. Additionally, any issues identified with the running web server instance needed to be resolved, again without breaking other functionality.

As a result of the rapid prototyping approach used to develop antiSMASH 1, different prediction modules were tightly coupled, responsibilities in the code were not clearly separated, and parts of the functionality were duplicated (Figure 1). Adding new modules required changes at many locations of the source code, making it relatively

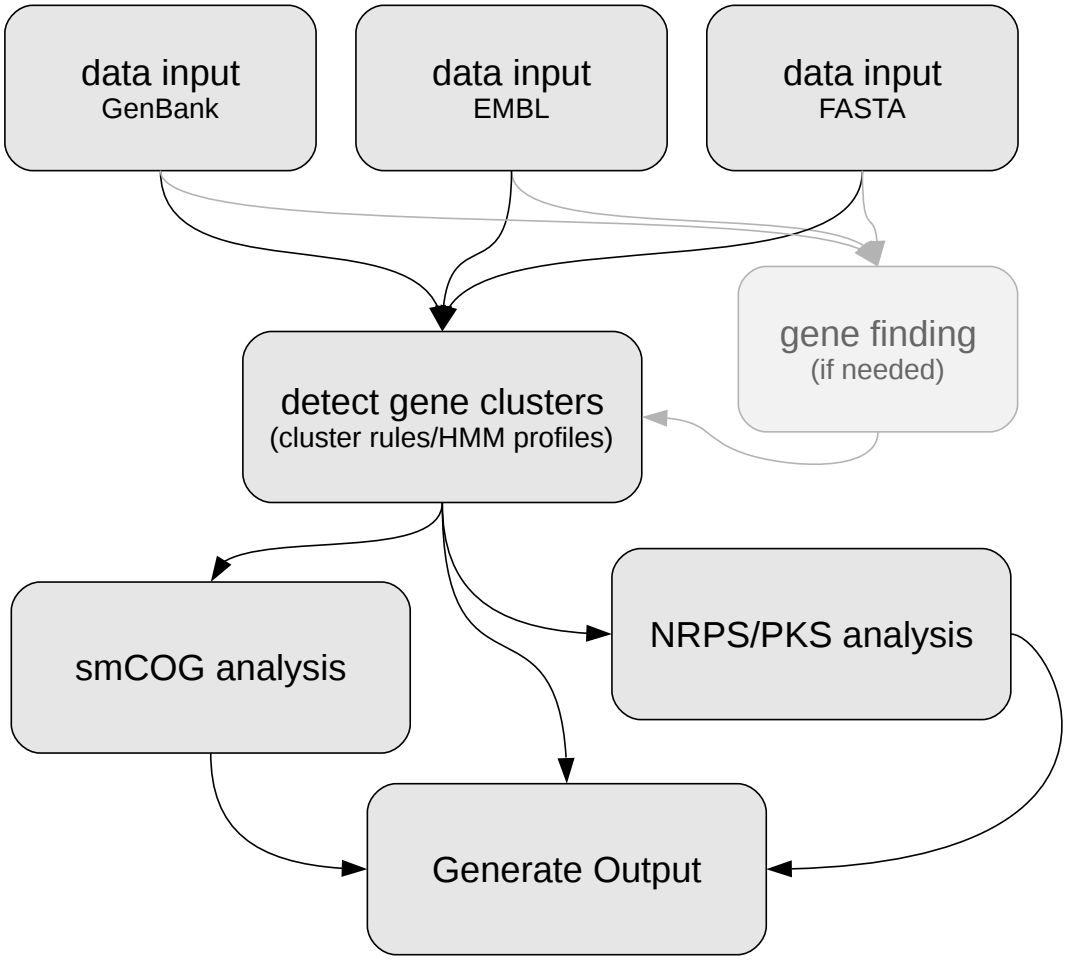


Figure 1: antiSMASH 1 architecture. Boxes show different tasks, arrows show dependencies between the tasks

hard to maintain and extend. To improve the state of the software in a controlled manner – commonly called refactoring (Fowler 1999) – the dependencies were isolated following the recommendations set by Feathers (2004). The individual parts were then brought under unit test (Martin 2008) coverage. Following the so-called DRY principle: *"Every piece of knowledge must have a single, unambiguous, authoritative representation within a system*" (Thomas and Hunt 1999, page 27), areas of duplicated functionality were abstracted out and collected in a central utility library. The input modules were switched to using parsers provided by the BioPython project (Cock et al. 2009). Instead of hard-coding lists of signature profiles and cluster rules, plain text files (Thomas and Hunt 1999, page 73) were used to dynamically load the rules to identify secondary metabolite clusters. Predictor modules specific for a sec-

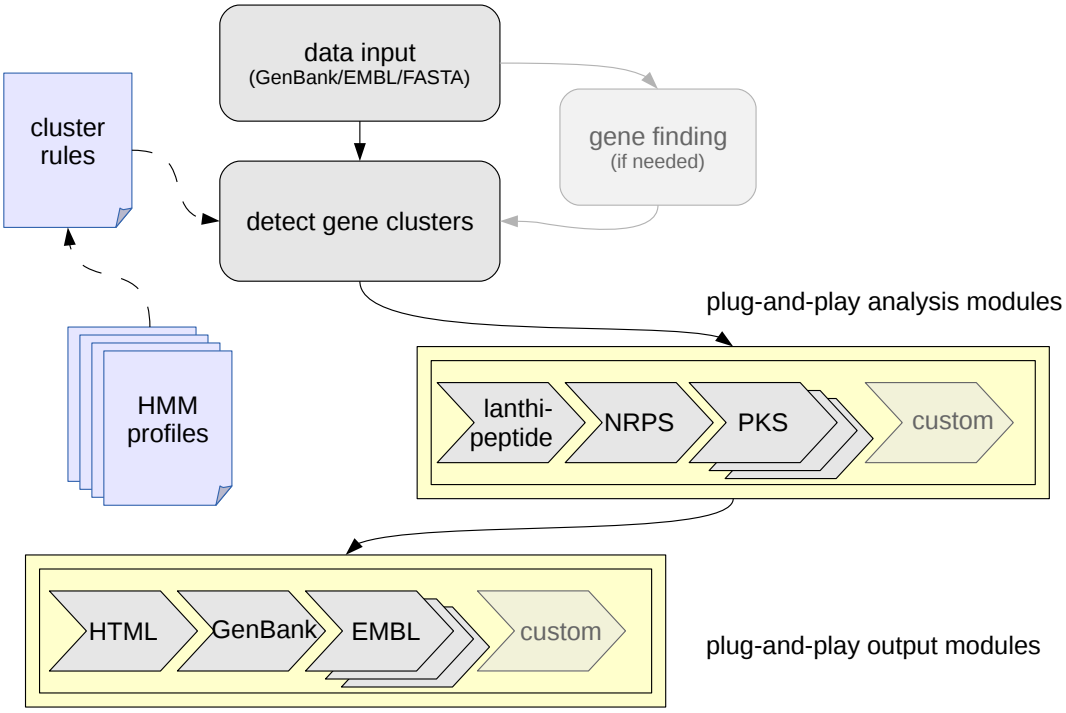


Figure 2: antiSMASH 2 architecture. Rounded boxes are predefined tasks, yellow frames are modular tasks, box arrows are individual modules of modular tasks, and folded-corner boxes are external files that can be modified by the antiSMASH user.

ondary metabolite cluster type work independently, all using the same application programming interface (API). As the last step, all output modules work on the same data, again producing their output independent of each other. The resulting architecture (Figure 2) is much less complex and easier to extend.

In addition to these significant changes to the antiSMASH software back-end components, a number of user-visible changes were introduced to antiSMASH 2. In addition to the 18 secondary metabolite cluster types identified by antiSMASH 1 (polyketides, nonribosomal peptides, terpenes, aminoglycosides, aminocoumarins, indolocarbazoles, lantibiotics, bacteriocins, nucleosides, β-lactams, ectoines, butyrolactones, siderophores, melanins and others), antiSMASH 2 supports six additional cluster types: oligosaccharides, thiopeptides, phenazines, furans, homoserine lactones and phosphonates. Furthermore, improved detection profiles for many of the existing cluster types were added (for a detailed list, see Blin et al. 2013b, Table S1). All these 24 cluster types are identified using the secondary metabolite cluster rules from Blin et al. (2013b, Table S2). Extending the method used by the ClusterBlast algorithm,

a new method called SubClusterBlast was introduced, making it possible to identify operons related in the biosyntheis of precursor moieties. SubClusterBlast contains a database of 126 sub-clusters related to precursor biosynthesis from gene clusters encoding known compounds. A first implementation of a cluster-specific prediction module for lanthipeptides was added, predicting the lanthipeptide class based on the sequence of the precursor peptide.
Apart from these scientific additions, the interface was improved to provide a better user experience. Support for draft genome input was added, both for NCBI downloads and uploads from the user's browser. For prokaryotic sequences alone, this change increases the number of available genomes from 2570 to 8898. In coordination with feature requensts from antiSMASH users, several changes to the antiSMASH web interface were made in an agile development approach (Beck 2000). Most notably, an overview page gives a summary of the identified secondary metabolite clusters. Cluster navigation buttons are now colour coded by secondary metabolite type, allowing quick access to secondary metabolite clusters of interest. Most graphics are now vector-based, so they can be used to create publication quality illustrations, even for poster formats. Major improvements were made to the page loading times. On the arbitrary example of *Streptomyces tsukubaensis* NRRL18488 (Genbank accession no. AJSZ01000001), loading the result page with 35 detected clusters took over 40 seconds in antiSMASH 1, due to the large size of the result page. In antiSMASH 2, the visualisation component was redesigned and optimized for loading speed. Loading the result page for the example *S. tsukubaensis* NRRL18488 took less than 2 seconds, even though 37 secondary metabolite clusters were identified. Thanks to the redesign of the visualisation, antiSMASH results now can also be browsed from smart phone and tablet browsers.

4.3 NRPSPredictor2

After providing an accessible user interface for secondary metabolite analysis tools in the form of antiSMASH, focus was shifted towards improving individual prediction algorithms. Six years after the original release of the NRPSPredictor software (Rausch et al. 2005), an updated version of NRPSPredictor was released, taking into account the novel A domains with elucidated substrate specificities and an improved prediction algorithm. To the 397 labeled domains taken from the original release, 79

labeled bacterial and 100 labeled fungal A domains were added. To further improve the predictions, 4282 unlabeled bacterial and 814 unlabeled fungal A domains were also added to the training set (for a detailed list, see Röttig et al. 2011, supplemental material file S1). To train classical SVMs, only labeled training data may be used. As labeled training data for A domains are scarce, especially for fungal A domains, transductive support vector machines (TSVMs) (Joachims 1999) were integrated to also utilise unlabeled training data to train more robust classifiers.
In order to provide the best possible prediction with the available training data, NRPSPredictor2 predicts A domain substrate specificities in four different detail levels for bacterial A domains and one detail level for fungal A domains. The detail levels available for bacterial sequences are predictions of the gross physio-chemical properties (hydrophobic-aromatic, hydrophobic-aliphatic and hydrophilic), large clusters of amino acids with similar physiochemical properties and sizes, small clusters of closely related amino acids, and single amino acids (see Röttig et al. (2011, Table 1) for a detailed list). Due to the lack of sufficient training data, the fungal predictor only predicts the gross physio-chemical properties. For substrates where less than five A domain sequences are known, no single amino acid SVM model was constructed. In order to also cover these substrates, NRPSPredictor2 utilises a nearest neighbour rule to predict the substrate specificity based on the most similar active site signature, based on the Stachelhaus code (Stachelhaus et al. 1999).
By using different predictive models for bacterial and fungal sequences, NRPSPredictor2 is the first NRPS prediction tool to account for the different active site residues, yielding in a more accurate overall prediction. NRPSPredictor2 was fully integrated into the NRPS prediction module of antiSMASH.

4.4 Lanthipeptide Prediction

When trying to detect ribosomally synthetised and post-translationally modified peptides (RiPPs), a common approach is to run culture extracts from the organism of interest through HPLC analysis and comparing the identified mass peaks with product predictions from an antiSMASH analysis for said organism. Having an accurate mass available makes it easier to identify compounds in an HPLC screening. In the initial antiSMASH 2.0 release, the lanthipeptide prediction only took the precursor peptide into account and thus did not model the biosynthesis performed by the whole

biosynthetic gene cluster accurately. A novel prediction algorithm was implemented in antiSMASH 2.2 to improve the prediction accuracy for lanthipeptide products from identified clusters. Unlike the old version, the new algorithm specifically considers the whole lanthipeptide biosynthetic gene cluster composition. This allows the algorithm to distinguish between class I, II, III and IV lanthipeptides by checking for the presence of LanB & LanC, LanM, LanKC, and LanL enzymes, respectively. Based on the identified class, the applicable cleavage site profile is selected.

After predicting the cleavage site, both the monoisotopic mass and the average molecular weight are calculated under the assumption that all Ser and Thr residues are dehydrated. A lack of dehydration is observed frequently. Unfortunately no mechanism behind this has been described so far. To still allow for an easy identification of these partially still hydrated peptides, a list of alternative weights is calulated to cover the possible hydration of all Ser and Thr not participating in a lanthionine bridge with a Cys residue.

The presence of further tailoring enzymes determines the post-translational tailoring reactions. The presence of a LanD-type flavin-dependent decarboxylase indicates the formation of a C-terminal *S*-[(*Z*)-2-aminovinyl]-D-cysteine (AviCys) or *S*-[(*Z*)-2-aminovinyl]-3-methyl-D-cysteine (AviMeCys) residue. LanH-type halogenases chlorinate amino acid side chains. LanO-type cytochrome P450 oxygenases regiospecifically oxidise non-activated hydrocarbons. EciO-type short chain dehydrogenases catalyse the final step in the conversion of an N-terminal Dha residue into lactate. All these tailoring reactions affect the predicted molecular mass and are also identified explicitly in the lanthipeptide cluster details page. By considering the whole gene cluster instead of basing the prediction on the precursor peptide alone, it is now possible to correctly predict 89 % of the lanthipeptide benchmark dataset (see Blin et al. 2013a, Table 4).

4.5 Conclusions

antiSMASH is the first Open Source Software pipeline to assist natural product researchers in the analysis of a wide range of secondary metabolite gene cluster types. Not only does it provide an accessible and easy to use web interface for prediction algorithms that have been published previously (e.g. Rausch et al. 2005; Minowa et al. 2007; Weber et al. 2009; Stachelhaus et al. 1999), it also adds a number of

novel mechanisms to further annotate secondary metabolite clusters, like the smCOG analysis, ClusterBlast & SubClusterBlast, and the lanthipeptide prediction algorithm. Since the release of the antiSMASH 1 publication (Medema et al. 2011) in July 2011, the publicly available web server at `http://antismash,secondarymetabolites.org` has processed about 50 000 analysis jobs. The standalone version of antiSMASH has been downloaded over 6000 times.

In the past years, antiSMASH has established itself as the standard tool natural product scientists run on newly sequenced genomes, to date over 120 genome anouncement papers have used antiSMASH to identify secondary metabolite clusters. The standalone version has been integrated into several other toolchains, both proprietary and published (e.g. Conway and Boddy 2013). Even in the competitive field of NRPS/PKS analysis tools, antiSMASH is considered "*the most comprehensive tool currently available*" (Boddy 2013). An important factor for the success of antiSMASH is the public web server. It is both easy to use and free, without forcing users to sign up for an account or getting a formal license, thus lowering the barrier of entry. Power users who have more requirements in terms of performance or who cannot send their genome data to a third party can use the standalone version, again without having to obtain an explicit license. The source code for antiSMASH is publicly available as well, enabling both peer review of the implementation details as well as an easy creation of fixes for any potential bugs identified in the software.

With the new software architecture introduced with antiSMASH 2, the software has become easy to extend, making it possible to quickly add new prediction modules as new research fields gain influence and more data becomes available. The lanthipeptide prediction module is an example, more specific predictors for the currently trending field of RiPPs (Arnison et al. 2013) can be added without much effort.

List of Figures

List of Tables

References

Ansari, M. Z., Yadav, G., Gokhale, R. S., and Mohanty, D. (2004). "NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases". *Nucleic Acids Research* 32 (suppl 2), W405–W413.

Arnison, P. G., Bibb, M. J., Bierbaum, G., Bowers, A. A., Bugni, T. S., Bulaj, G., Camarero, J. A., Campopiano, D. J., et al. (2013). "Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature". *Natural Product Reports* 30 (1), pp. 108–160.

Beck, K. (2000). *Extreme Programming, Das Manifest*. Addison Wesley München.

Bell, G., Hey, T., and Szalay, A. (2009). "Beyond the data deluge". *Science* 323 (5919), pp. 1297–1298.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2013). "GenBank". *Nucleic Acids Research* 41 (D1), pp. D36–D42.

Bentley, S., Chater, K., Cerdeno-Tarraga, A.-M., Challis, G. L., Thomson, N., James, K., Harris, D., Quail, M., et al. (2002). "Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3 (2)". *Nature* 417 (6885), pp. 141–147.

Bevitt, D. J., Cortés, J., Haydock, S. F., and Leadley, P. F. (1992). "6-Deoxyerythronolide-B synthase 2 from *Saccharopolyspora erythraea*". *European Journal of Biochemistry* 204 (1), pp. 39–49.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., et al. (1997). "The Complete Genome Sequence of *Escherichia coli* K-12". *Science* 277 (5331), pp. 1453–1462.

Blin, K., Kazempour, D., Wohlleben, W., and Weber, T. (2013a). "Improved lanthipeptide detection and prediction in antiSMASH". *submitted*.

Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. A., Breitling, R., Takano, E., and Weber, T. (2013b). "antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers". *Nucleic Acids Research* 41 (W1), W204–W212.

Boddy, C. N. (2013). "Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides". *Journal of Industrial Microbiology & Biotechnology*, pp. 1–8.

Caboche, S., Pupin, M., Leclère, V., Fontaine, A., Jacques, P., and Kucherov, G. (2008). "NORINE: a database of nonribosomal peptides". *Nucleic Acids Research* 36 (suppl 1), pp. D326–D331.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., and Madden, T. (2009). "BLAST+: architecture and applications". *BMC Bioinformatics* 10 (1), p. 421.

Challis, G. L., Ravel, J., and Townsend, C. A. (2000). "Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains". *Chemistry & Biology* 7 (3), pp. 211–224.

Chevreux, B. (2005). "MIRA: an automated genome and EST assembler". PhD thesis. Ruprecht-Karls University, Heidelberg, Germany.

Chinault, A. C. and Carbon, J. (1979). "Overlap hybridization screening: Isolation and characterization of overlapping DNA fragments surrounding the leu2 gene on yeast chromosome III". *Gene* 5 (2), pp. 111–126.

Claisen, L. (1887). "Ueber die Einführung von Säureradicalen in Ketone". *Berichte der deutschen chemischen Gesellschaft* 20 (1), pp. 655–657.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., and Hoon, M. J. L. de (2009). "Biopython: freely available Python tools for computational molecular biology and bioinformatics." *Bioinformatics* 25 (11), pp. 1422–3.

Conti, E., Stachelhaus, T., Marahiel, M. A., and Brick, P. (1997). "Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S". *The EMBO journal* 16 (14), pp. 4174–4183.

Conway, K. R. and Boddy, C. N. (2013). "ClusterMine360: a database of microbial PKS/NRPS biosynthesis". *Nucleic Acids Research* 41 (D1), pp. D402–D407.

Dayhoff, M. O. and Eck, R. V. (1969). *Atlas of Protein Sequence and Structure: 1967-68; Margaret O. Dayhoff, Richard V. Eck*. National Biomedical Research Foundation.

Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). "Identifying bacterial genes and endosymbiont DNA with Glimmer". *Bioinformatics* 23 (6), pp. 673–679.

Demain, A. L. (1999). "Pharmaceutically active secondary metabolites of microorganisms". *Applied Microbiology and Biotechnology* 52 (4), pp. 455–463.

Donadio, S., Staver, M. J., McAlpine, J. B., Swanson, S. J., and Katz, L. (1991). "Modular organization of genes required for complex polyketide biosynthesis". *Science* 252 (5006), pp. 675–679.

Eddy, S. R. (2011). "Accelerated Profile HMM Searches". *PLoS Computational Biology* 7 (10), e1002195.

Feathers, M. (2004). *Working effectively with legacy code*. Prentice Hall Professional.

Ferrer, J.-L., Jez, J. M., Bowman, M. E., Dixon, R. A., and Noel, J. P. (1999). "Structure of chalcone synthase and the molecular basis of plant polyketide biosynthesis". *Nature Structural & Molecular Biology* 6 (8), pp. 775–784.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J.-F., et al. (1995). "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd". *Science* 269 (5223), pp. 496–512.

Fleming, A. (1929). "On the antibacterial action of cultures of *Penicillium*, with special reference to their use in the isolation of *B. influenzae*." *British Journal of Experimental Pathology* 10, pp. 226–236.

Foulston, L. C. and Bibb, M. J. (2010). "Microbisporicin gene cluster reveals unusual features of lantibiotic biosynthesis in actinomycetes." *Proceedings of the National Academy of Sciences* 107 (30), pp. 13461–6.

Fowler, M. (1999). *Refactoring: improving the design of existing code*. Addison-Wesley Professional.

Heidrich, C., Pag, U., Josten, M., Metzger, J., Jack, R. W., Bierbaum, G., Jung, G., and Sahl, H.-G. (1998). "Isolation, characterization, and heterologous expression of the novel lantibiotic epicidin 280 and analysis of its biosynthetic gene cluster". *Applied and Environmental Microbiology* 64 (9), pp. 3140–3146.

Hesper, B. and Hogeweg, P. (1970). "Bioinformatica: een werkconcept". *Kameleon* 1 (6), pp. 28–29.

Hogeweg, P. (2011). "The Roots of Bioinformatics in Theoretical Biology". *PLoS Computational Biology* 7 (3), e1002021.

Huerta, M., Downing, G., Haseltine, F., Seto, B., and Liu, Y. (2000). "NIH working definition of bioinformatics and computational biology". *US National Institute of Health*.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M., et al. (2003). "Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*". *Nature Biotechnology* 21 (5), pp. 526–531.

Joachims, T. (1999). "Transductive inference for text classification using support vector machines". In: *Proceedings of the Sixteenth International Conference on Machine Learning*. Vol. 99, pp. 200–209.

Kaletta, C. and Entian, K.-D. (1989). "Nisin, a peptide antibiotic: cloning and sequencing of the nisA gene and posttranslational processing of its peptide product." *Journal of Bacteriology* 171 (3), pp. 1597–1601.

Kamra, P., Gokhale, R. S., and Mohanty, D. (2005). "SEARCHGTr: a program for analysis of glycosyltransferases involved in glycosylation of secondary metabolites". *Nucleic Acids Research* 33 (suppl 2), W220–W225.

Knerr, P. J. and Donk, W. A. van der (2012). "Discovery, Biosynthesis, and Engineering of Lantipeptides." *Annual Review of Biochemistry* 81 (February), pp. 1–27.

Kuipers, O. P., Beerthuyzen, M. M., Siezen, R. J., and Vos, W. M. (1993). "Characterization of the nisin gene cluster nisABTCIPR of *Lactococcus lactis*". *European Journal of Biochemistry* 216 (1), pp. 281–291.

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). "TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders". *Bioinformatics* 20 (16), pp. 2878–2879.

Malpartida, F. and Hopwood, D. A. (1984). "Molecular cloning of the whole biosynthetic pathway of a *Streptomyces* antibiotic and its expression in a heterologous host". *Nature* 309 (5967), pp. 462–464.

Martin, R. C. (2008). *Clean code: a handbook of agile software craftsmanship*. Pearson Education.

Medema, M. H., Blin, K., Cimermancic, P., Jager, V. de, 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 Research* 39 (Web Server issue), W339–W346.

Minowa, Y., Araki, M., and Kanehisa, M. (2007). "Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes." *Journal of Molecular Biology* 368 (5), pp. 1500–17.

Needleman, S. B. and Wunsch, C. D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology* 48 (3), pp. 443–453.

Nussinov, R. and Jacobson, A. B. (1980). "Fast algorithm for predicting the secondary structure of single-stranded RNA". *Proceedings of the National Academy of Sciences* 77 (11), pp. 6309–6313.

Ohnishi, Y., Ishikawa, J., Hara, H., Suzuki, H., Ikenoya, M., Ikeda, H., Yamashita, A., Hattori, M., and Horinouchi, S. (2008). "Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350". *Journal of Bacteriology* 190 (11), pp. 4050–4060.

Pelzer, S., Süssmuth, R., Heckmann, D., Recktenwald, J., Huber, P., Jung, G., and Wohlleben, W. (1999). "Identification and analysis of the balhimycin biosynthetic gene cluster and its use for manipulating glycopeptide biosynthesis in *Amycolatopsis mediterranei* DSM5908". *Antimicrobial Agents and Chemotherapy* 43 (7), pp. 1565–1573.

Piel, J. (2010). "Biosynthesis of polyketides by trans-AT polyketide synthases". *Natural Product Reports* 27 (7), pp. 996–1047.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., et al. (2012). "The Pfam protein families database". *Nucleic Acids Research* 40 (D1), pp. D290–D301.

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W., and Huson, D. H. (2005). "Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs)." *Nucleic Acids Research* 33 (18), pp. 5799–5808.

Richardson, E. J. and Watson, M. (2013). "The automatic annotation of bacterial genomes". *Briefings in Bioinformatics* 14 (1), pp. 1–12.

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 research* 39 (Web Server issue), W362–W367.

Schatz, A., Bugie, E., and Waksman, S. A. (Jan. 1944). "Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria." *Experimental Biology and Medicine* 55 (1), pp. 66–69.

Schwarzer, D., Finking, R., and Marahiel, M. A. (2003). "Nonribosomal peptides: from genes to products". *Natural Product Reports* 20 (3), p. 275.

Sit, C. S., Yoganathan, S., and Vederas, J. C. (2011). "Biosynthesis of Aminovinyl-Cysteine-Containing Peptides and Its Application in the Production of Potential Drug Candidates". *Accounts of Chemical Research* 44 (4), pp. 261–268.

Stachelhaus, T., Mootz, H. D., and Marahiel, M. A. (1999). "The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases." *Chemistry & Biology* 6 (8), pp. 493–505.

Staunton, J. and Weissman, K. J. (2001). "Polyketide biosynthesis: a millennium review". *Natural Product Reports* 18, pp. 380–416.

Tang, Y., Tsai, S.-C., and Khosla, C. (2003). "Polyketide chain length control by chain length factor". *Journal of the American Chemical Society* 125 (42), pp. 12708–12709.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., et al. (2003). "The COG database: an updated version includes eukaryotes". *BMC Bioinformatics* 4 (1), p. 41.

Thomas, D. and Hunt, A. (1999). *The Pragmatic Programmer: From Journeyman to Master*. Addison-Wesley Professional.

Walsh, C. T. and Fischbach, M. A. (2010). "Natural products version 2.0: connecting genes to molecules". *Journal of the American Chemical Society* 132 (8), pp. 2469–2493.

Weber, T., Laiple, K. J., Pross, E. K., Textor, A., Grond, S., Welzel, K., Pelzer, S., Vente, A., and Wohlleben, W. (2008). "Molecular analysis of the kirromycin biosynthetic gene cluster revealed β-alanine as precursor of the pyridone moiety". *Chemistry & Biology* 15 (2), pp. 175–188.

Weber, T., Rausch, C., Lopez, P., Hoof, I., Gaykova, V., Huson, D. H., and Wohlleben, W. (2009). "CLUSEAN: A computer-based framework for the automated analysis of bacterial secondary metabolite biosynthetic gene clusters". *Journal of Biotechnology* 140 (1-2), pp. 13–17.

Wetterstrand, K. A. (July 2013). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. http://www.genome.gov/sequencingcosts/. Accessed: 2013-09-29.

Yadav, G., Gokhale, R. S., and Mohanty, D. (2003). "Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases". *Journal of Molecular Biology* 328 (2), pp. 335–363.

Contributions

In publication 1 (Medema et al. 2011), I designed the antiSMASH web service job control and scheduling system, and the sandboxing system that isolates the different jobs from each other while running. Marnix H. Medema and myself implemented and tested the core antiSMASH program in equal parts. The Linux install scripts were also developed and tested by me.
In publication 2 (Blin et al. 2013b), I redesigned and reimplemented the antiSMASH web service component, and adjusted the job control, scheduling and sandboxing systems accordingly. I designed the new modular core program architecture, wrote the unit tests and reimplemented the sequence input modules. I converted the cluster identification logic to the new architecture. I redesigned the HTML output page, performing continuous benchmarks to identify performance bottlenecks, and reimplemented the JavaScript logic behind the dynamic elements of the output page. I set up the continuous integration system ensuring that the test routines were run on every change of the antiSMASH program. I also wrote most of the manuscript.
In publication 3 (Röttig et al. 2011), I performed the literature mining to identify novel A domains with biochemically characterized substrate specificity. I also assisted in testing the NRPSPredictor 2 program.
In publication 4 (Blin et al. 2013a), I developed the algorithm for predicting cleavage site, modifications and the resulting molecular mass of lanthipeptide cluster core peptides. I supervised Daniyal Kazempour during his initial attempt of cleavage site prediction. I wrote the program parts that predict the tailoring reactions based on other enzymes present in the cluster, updated the cleavage site prediction logic to infer the lanthipeptide class from the cluster layout and integrated the prediction tool as an antiSMASH module. I also worte the manuscript.