Chapter 14

Bioinformatics Tools for the Discovery of New Nonribosomal Peptides

Valérie Leclère, Tilmann Weber, Philippe Jacques, and Maude Pupin

Abstract

This chapter helps in the use of bioinformatics tools relevant to the discovery of new nonribosomal peptides (NRPs) produced by microorganisms. The strategy described can be applied to draft or fully assembled genome sequences. It relies on the identification of the synthetase genes and the deciphering of the domain architecture of the nonribosomal peptide synthetases (NRPSs). In the next step, candidate peptides synthesized by these NRPSs are predicted in silico, considering the specificity of incorporated monomers together with their isomery. To assess their novelty, the two-dimensional structure of the peptides can be compared with the structural patterns of all known NRPs. The presented workflow leads to an efficient and rapid screening of genomic data generated by high throughput technologies. The exploration of such sequenced genomes may lead to the discovery of new drugs (i.e., antibiotics against multi-resistant pathogens or anti-tumors).

Key words Drug discovery, Bioinformatics tools, Nonribosomal, Antibiotic, Peptide, Genome mining

1 Introduction

The nonribosomal peptides (NRPs) are a large, structurally diverse group of natural products [1]. They constitute a source of bioactive compounds of great biomedical importance, including antibiotics such as vancomycin or daptomycin, antitumoral compounds such as actinomycin D, or immunosuppressants such as cyclosporine. In the Norine database [2], which currently is the most comprehensive database of NRPs, only 31 % of the curated peptides are linear, the remainder having branched, cyclic, or more complex primary structures [3]. Moreover, the peptides contain collectively over 525 different monomers including non-proteogenic and modified amino acids, carbohydrates, or lipids. In consequence, there is a high demand to develop specific bioinformatics tools dedicated to these compounds because they cannot be analyzed as classical peptides.

Bradley S. Evans (ed.), *Nonribosomal Peptide and Polyketide Biosynthesis: Methods and Protocols*, Methods in Molecular Biology, vol. 1401, DOI 10.1007/978-1-4939-3375-4_14, © Springer Science+Business Media New York 2016

Nonribosomal peptide synthetases (NRPSs) that build up the NRPs are working as enzymatic assembly lines. They are megaenzymes displaying an extraordinary modular architecture containing catalytic domains, essentially adenylation domains (A) for the selection of the monomers, thiolation domains (T or PCP for Peptidyl Carrier Proteins) for the covalent tethering of activated monomers onto the NRPSs, condensation domains (C) for the peptide bond formation and thioesterase domains (Te) for the release of the peptide from the NRPS. These enzymatic domains harbor highly conserved core motifs also called signatures [4]. These consensus sequences are useful to identify proteins with unknown functions as being NRPSs, considering that similar sequences are relevant to similar functions.

In most cases, one given NRPS is expected to synthesize one peptide (or close variants). Several computational tools have been developed to analyze NRPS pathways in silico (for reviews, *see* refs. 5–7). They use the amino acid sequence of NRPSs to predict their specificity towards the amino acid building blocks they assemble into the NRP. The first predictive methods were based on the work of Stachelhaus et al. [8] and Challis et al. [9], who demonstrated for the first time that eight amino acids in the active site of the NRPS adenylation domain (A-domain) are crucial for determining the substrate specificity and can be used to predict and engineer A-domain specificities. Querying these datasets is, for example, implemented in the PKS/NRPS Web Server/Predictive Blast Server [10]. Alternative approaches use machine learning based methods like profile Hidden Markov Models (pHMMs) [11] to infer substrate specificities [12, 13], Support Vector Machines (SVMs) [14, 15] or Latent Semantic Indexing [16]. In addition to predicting the A-domain substrate specificity, the isomery of each monomer can also be determined by considering the presence of domains with epimerisation activity, to give better predictions of the structure [17–19].

In this chapter, we present the different Web-based bioinformatics tools currently used for an efficient screening of genomic data to discover new NRPs, and to compare them structurally to all known NRPs.

Two study cases will illustrate the use of the bioinformatics tools presented in this chapter. The first example is cichofactin, a lipopeptide produced by *Pseudomonas* that was identified from sequencing data by following the advised in silico workflow, followed by confirmation in wet-lab experiments [20]. The GenBank accession number of the *c*ichofactin biosynthetic gene cluster is *KJ513093*. The second example shows how it is possible to rapidly detect the potential of NRPS synthesis from a draft genome of *Pseudomonas* split in more than 1000 contigs. The strain is *Pseudomonas syringae pv. tabaci* str. ATCC 11528, with GenBank assembly ID GCA_000159835.2.

2 Materials

2.1 Software for the Analysis of NRPSs and Their Products

During the last decade, several programs have been developed to aid on the analysis of NRPSs and NRPs. Most of these tools are accessible via websites and provide user friendly interfaces offering the detection of NRPS encoding genes within genomic sequences or the analysis of the NRPS domain organization and specificities based on their protein sequences without specific bioinformatics skills (*see* **Note 1**). A summary and links to the main tools is given in Table 1.

For the example workflows described in this manuscript we refer to the most commonly used tools.

2.2 Bioinformatics Tools to Predict NRPSs from Genome Sequence

1. antiSMASH (http://antismash.secondarymetabolites.org)

   antiSMASH [21–23], the antibiotics and secondary metabolites analysis shell, is a comprehensive genome mining platform that is capable of identifying and analyzing many different types of secondary metabolites, including products of NRPS. antiSMASH integrates several algorithms dedicated to the specific analysis of NRPS, e.g., the method of Minowa et al. [12] or NRPSpredictor [14, 15] and also provides direct links to different NRP and NRPS analysis websites like NORINE [2, 28] or NaPDoS [18].

   antiSMASH, which is licensed under the Affero GNU license, is freely accessible at http://antismash.secondarymetabolites.org. In addition, the software can be downloaded and installed on Linux, MS Windows, and Mac OS X.

   (a) Input data for antiSMASH

   antiSMASH can be used alternatively with protein sequences of NRPS in FASTA format, or preferably with genomic data. Genomic data can be provided as EMBL-formatted, or GenBank-formatted files including annotation (highly preferred method), but also multi-FASTA files containing only the DNA sequence are accepted.

   In the case of FASTA nucleotides uploaded to antiSMASH, genes are automatically predicted using the gene finding programs Glimmer for bacterial sequences [29], or GlimmerHMM for fungal sequences [30].

   Alternatively, antiSMASH can directly download sequences from NCBI RefSeq [31] or GenBank [32] databases if the user provides an accession number instead of uploading a file (*see* **Note 2** for remarks on sequence/assembly quality).

   (b) Output of antiSMASH

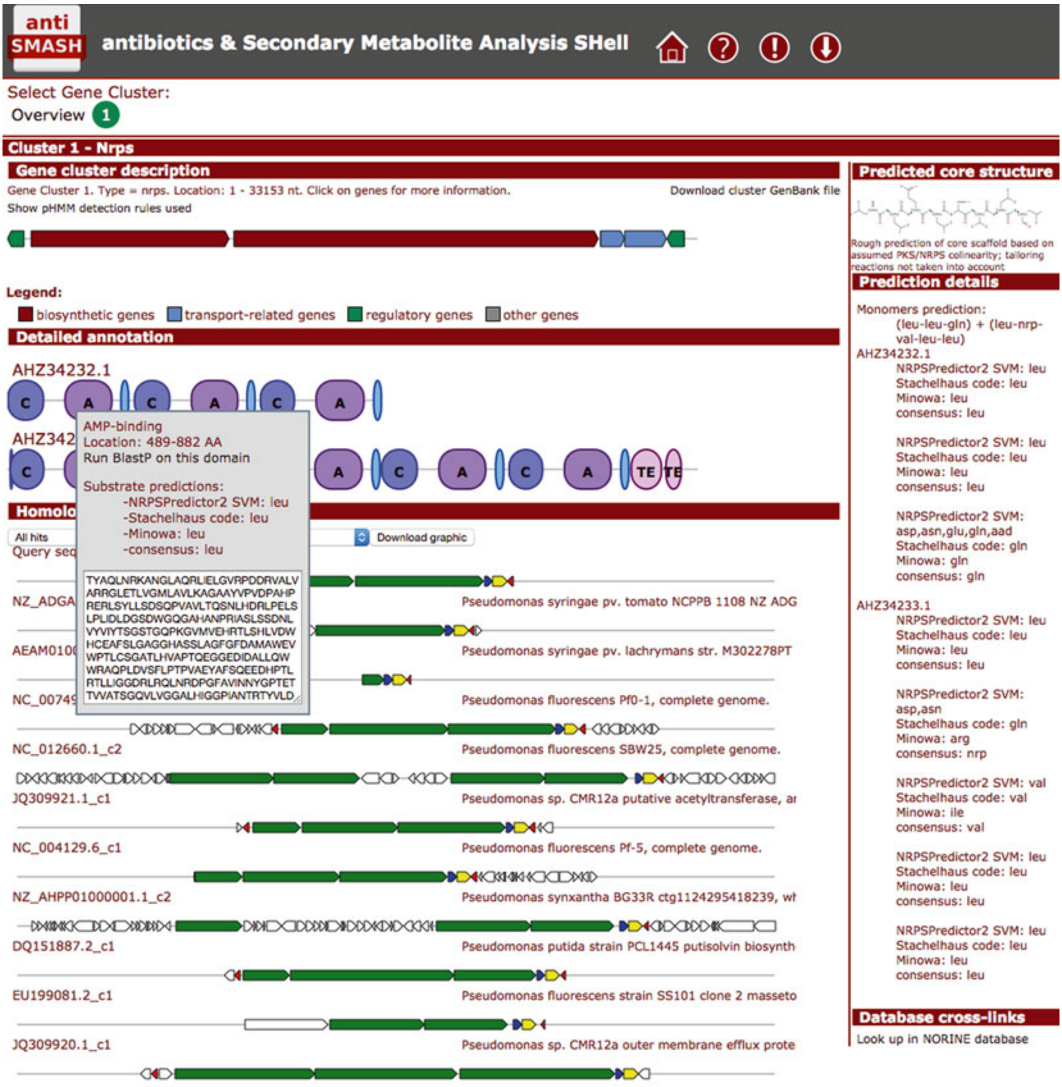
   antiSMASH returns a comprehensive genome mining analysis on the provided input data. If whole genome sequences are uploaded, antiSMASH will identify 44 different classes of antibiotics biosynthesis pathways. Similar

Table 1
Software dedicated to the analysis of nonribosomal peptide synthetases (Table adapted from Weber [6], with permission from Elsevier)

Program/database	URL	References
Software for the analysis of NRPS pathways		
antiSMASH	http://antismash.secondarymetabolites.org	[21–23]
ClustScan Professional	http://bioserv.pbf.hr/cms/index.php?page=clustscan	[24]
NaPDos	http://napdos.ucsd.edu/	[18]
NP.searcher	http://dna.sherman.lsi.umich.edu/	[25]
NRPS-PKS/SBSPKS	http://www.nii.ac.in/~pksdb/sbspks/master.html	[26, 27]
PKS/NRPS Web Server/Predictive Blast Server	http://nrps.igs.umaryland.edu/nrps/	[10]
Tools to predict NRPS substrate specificities		
LSI based A-domain function predictor	http://bioserv7.bioinfo.pbf.hr/LSIpredictor/AdomainPrediction.jsp	[16]
PKS/NRPS Web Server/Predictive Blast Server	http://nrps.igs.umaryland.edu/nrps/	[10]
NRPSpredictor/NRPSpredictor2	http://nrps.informatik.uni-tuebingen.de	[14, 15]
NRPSSP	http://www.nrpssp.com/	[13]

clusters or conserved operons encoding the biosynthesis of conserved compounds will be automatically identified in a database containing all currently known gene clusters and building-block biosynthesis pathways. For NRPS and PKS pathways, in addition, a comprehensive analysis of the domain organization and predictions of substrate specificity is performed and leads to putative products (Fig. 1).

Most results can directly be accessed on an interactive website (see paragraph).



Gene Cluster 1. Type = nrps. Location: 1 - 33153 nt. Click on genes for more information.

Download cluster GenBank file

Show pHMM detection rules used

AMP-binding
Location: 489-882 AA
Run BlastP on this domain

Substrate predictions:
-NRPSPredictor2 SVM: leu
-Stachelhaus code: leu
-Minowa: leu
-consensus: leu

```
TYAQLNRKANGLAQRLIELGVRPDDRVALV
ARRGLETLVGMLAVLKAGAAYVPVDPAHP
RERLSYLLSDSQPVAVLTQSNLHDRLPELS
LPLIDLDGSDWGQGAHANPRIASLSSDNL
VYVIYTSGSTGQPKGVMVEHRTLSHLVDW
HCEAFSLGAGGHASSLAGFGFDAMAWEV
WPTLCSGATLHVAPTQEGGEDIDALLQW
WRAQPLDVSFLPTPVAEYAFSQEEDHPTL
RTLLIGGDRLRQLNRDPGFAVINNYGPTET
TVVATSGQVLVGGALHIGGPIANTRTYVLD
```

Pseudomonas syringae pv. tomato NCPPB 1108 NZ ADG

Pseudomonas syringae pv. lachrymans str. M302278PT

Pseudomonas fluorescens Pf0-1, complete genome.

NC_012660.1_c2 Pseudomonas fluorescens SBW25, complete genome.

JQ309921.1_c1 Pseudomonas sp. CMR12a putative acetyltransferase, ar

NC_004129.6_c1 Pseudomonas fluorescens Pf-5, complete genome.

NZ_AHPP01000001.1_c2 Pseudomonas synxantha BG33R ctg1124295418239, wh

DQ151887.2_c1 Pseudomonas putida strain PCL1445 putisolvin biosynth

EU199081.2_c1 Pseudomonas fluorescens strain SS101 clone 2 masseto

JQ309920.1_c1 Pseudomonas sp. CMR12a outer membrane efflux prote

Rough prediction of core scaffold based on assumed PKS/NRPS colinearity; tailoring reactions not taken into account

Monomers prediction:
(leu-leu-gln) + (leu-nrp-val-leu-leu)
AHZ34232.1
NRPSPredictor2 SVM: leu
Stachelhaus code: leu
Minowa: leu
consensus: leu

NRPSPredictor2 SVM: leu
Stachelhaus code: leu
Minowa: leu
consensus: leu

NRPSPredictor2 SVM: asp,asn,glu,gln,aad
Stachelhaus code: gln
Minowa: gln
consensus: gln

AHZ34233.1
NRPSPredictor2 SVM: leu
Stachelhaus code: leu
Minowa: leu
consensus: leu

NRPSPredictor2 SVM: asp,asn
Stachelhaus code: gln
Minowa: arg
consensus: nrp

NRPSPredictor2 SVM: val
Stachelhaus code: val
Minowa: ile
consensus: val

NRPSPredictor2 SVM: leu
Stachelhaus code: leu
Minowa: leu
consensus: leu

NRPSPredictor2 SVM: leu
Stachelhaus code: leu
Minowa: leu
consensus: leu

Look up in NORINE database

Fig. 1 Example output of antiSMASH analysis. The query was performed using the accession number of the cichofactin synthetase KJ513093, while the “non ribosomal peptides” option was selected. A pop-up window reporting the specificity of the first A-domain appeared after clicking on the considered domain. The predicted core structure of the peptide is shown on the *right column*

Alternatively, the complete antiSMASH analysis (including the HTML pages and annotated sequence files) can be downloaded using the download option "Download all results" for offline-use on the local computer (*see* **Note 3**).

2. NRPSpredictor 2 (http://nrps.informatik.uni-tuebingen.de)
   NRPSpredictor [14, 15] is a Web-tool that allows the prediction of A-domain specificities. NRPSpredictor uses a Support-Vector-Machine based algorithm to classify the A-domain on four hierarchical levels (Table 2) ranging from gross physicochemical properties of the substrates down to single amino acid substrates. It uses an applicability domain model to assess the quality of the prediction (corresponding to a statistical validation). In addition, the query A-domain sequences are compared against an A-domain signature database based on the work of Stachelhaus et al. [8] and Challis et al. [9].

Table 2
Prediction levels of NRPSpredictor2

No.	Members	Description
Three clusters		
1	Arg, Asp, Glu, His, Asn, Lys, Gln, Orn, Aad	Hydrophilic
2	Gly, Ala, Val, Leu, Ile, Abu, Iva, Ser, Thr, Hpg, Dhpg, Cys, Pro, Pip	Hydrophobic-aliphatic
3	Phe, Tyr, Trp, Dhb, Phg, Bht	Hydrophobic-aromatic
Large clusters		
1	Gly, Ala, Val, Leu, Ile, Abu, Iva	Apolar, aliphatic side chains
2	Ser, Thr, Dhpg, Hpg	Aliphatic chain or phenyl group with -OH
3	Phe, Trp, Phg, Tyr, Bht	Aromatic side chain
4	Asp, Asn, Glu, Gln, Aad	Aliphatic side chain with H-bond donor
5	Cys	Cysteine
6	Orn, Lys, Arg	Long positively charged side chain
7	Pro, Pip	Cyclic aliphatic chain with polar -NH_2 group
8	Dhb, Sal	Hydroxy-benzoic acid derivatives
Small clusters		
1	Gly, Ala	Tiny size, hydrophilic, transition to aliphatic
2	Val, Leu, Ile, Abu, Iva	Aliphatic, branched hydrophobic side chain
3	Ser	Serine specific
4	Thr	Threonine specific
5	Dhpg, Hpg	Polar uncharged hydroxy phenyl
6	Phe, Trp	Apolar aromatic ring
7	Tyr, Bht	Polar aromatic ring
8	Asp, Asn	Asp-Asn hydrogen acceptor
9	Glu, Gln	Glu-Gln hydrogen bond acceptor
10	Aad	2-Amino-adipic acid
11	Orn	Orn and hydroxy-Orn specific
12	Arg	Arg specific
13	Pro	Pro-specific
14	Dhb, Sal	Hydroxy-benzoic acid derivatives

A command line version of NRPSpredictor2 is available for download at the NRPSpredictor homepage.

(a) Input data for NRPSpredictor

NRPSpredictor can be used with amino acid sequences of NRPS, which can either be pasted to a query box or uploaded in FASTA format. Alternatively, if the user does not want to upload the whole amino acid sequence due to patenting/privacy issues, one can provide manually extracted 34-aa signature sequences. The format to submit these signature sequences is "vntsfdgsvfdgfilfggeihvygptestvyaty domain1", with a tabular character between the sequence and the domain name (*see* **Note 4**).

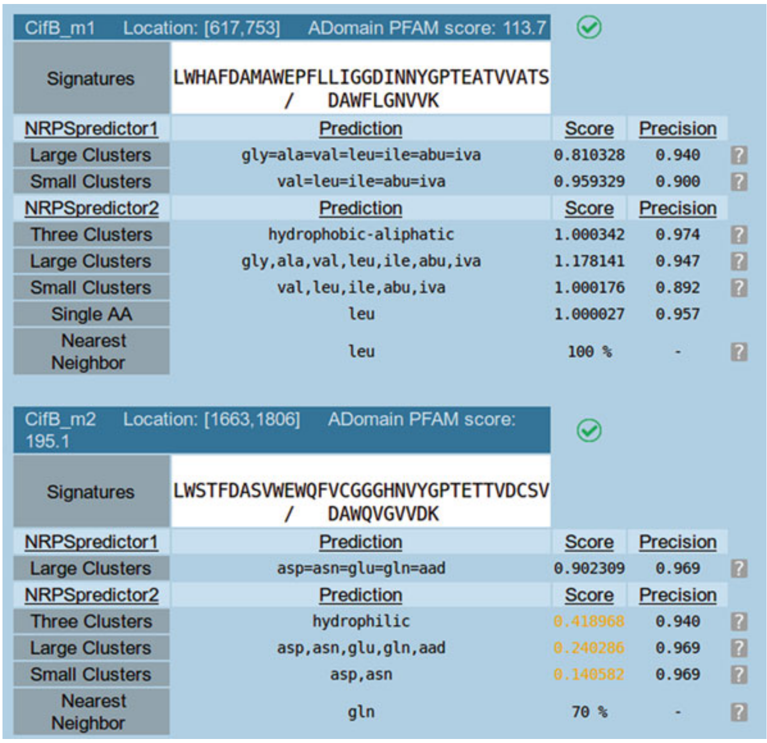
(b) Output of NRPSpredictor

The results of an NRPSpredictor analysis run are displayed as tables for each A-domain identified in the protein sequence(s). Each table contains the results of specificity prediction with the original NRPSpredictor1 algorithm, the improved NRPSpredictor2 algorithm and matches against the A-domain signatures of Stachelhaus et al. [8] and Challis et al. [9], which are determined by a Nearest Neighbour analysis.

NRPSpredictor2 returns prediction on four different hierarchical levels (Table 2, Fig. 2). The first level "Three clusters" makes a prediction on the general physicochemical properties of the incorporated amino acid, i.e., "hydrophilic, hydrophobic-aliphatic, hydrophobic-aromatic." The second level narrows down the prediction to large families/clusters of amino acids with similar properties. The third level tries to associate the A-domain specificity to smaller, more differentiated families. In the fourth level, a prediction is made based on matches against profiles only containing single amino acids. For all levels, a score and precision (*see* **Note 5**) are displayed. Additionally, a statistical support is notified by a green check✓ (supported) or red cross✕ (not supported). This support is the result of an applicability domain [33] analysis, which—simplified—provides information on how reliable the results are based on the similarity between the query sequence and the sequences used to train the SVM model.

3. NaPDoS (http://napdos.ucsd.edu/)

NaPDoS [18], the *Na*tural *P*roducts *Do*main *S*eeker, is a Web-tool that identifies and classifies condensation (C) domains of NRPS using a phylogenetic approach. With NaPDos it is easily possible to screen (meta)genomic DNA sequences and also protein sequences for the presence of C domains and classify them according to their catalytic function (starter-C-domain, $^{L}C_{L}$, $^{D}C_{L}$, Dual-condensation-epimerization, heterocyclization) and assign them to the closest known pathway.



CifB_m1 Location: [617,753] ADomain PFAM score: 113.7 ✓

Signatures	LWHAFDAMAWEPFLLIGGDINNYGPTEATVVATS / DAWFLGNVVK			
NRPSpredictor1	Prediction	Score	Precision	
Large Clusters	gly=ala=val=leu=ile=abu=iva	0.810328	0.940	?
Small Clusters	val=leu=ile=abu=iva	0.959329	0.900	?
NRPSpredictor2	Prediction	Score	Precision	
Three Clusters	hydrophobic-aliphatic	1.000342	0.974	?
Large Clusters	gly,ala,val,leu,ile,abu,iva	1.178141	0.947	?
Small Clusters	val,leu,ile,abu,iva	1.000176	0.892	?
Single AA	leu	1.000027	0.957	
Nearest Neighbor	leu	100 %	-	?

CifB_m2 Location: [1663,1806] ADomain PFAM score: 195.1 ✓

Signatures	LWSTFDASVWEWQFVCGGGHNVYGPTETTVDCSV / DAWQVGVVDK			
NRPSpredictor1	Prediction	Score	Precision	
Large Clusters	asp=asn=glu=gln=aad	0.902309	0.969	?
NRPSpredictor2	Prediction	Score	Precision	
Three Clusters	hydrophilic	0.418968	0.940	?
Large Clusters	asp,asn,glu,gln,aad	0.240286	0.969	?
Small Clusters	asp,asn	0.140582	0.969	?
Nearest Neighbor	gln	70 %	-	?

Fig. 2 Prediction of A-domain specificity using NRPSpredictor2. The analysis was performed with the N-terminal part of the cichofactin synthetase B, including two modules

In the first step of the NaPDoS analysis, the NRPS condensation domains are identified by a BlastP or BlastX, respectively for protein or coding DNA input, search against a hand-curated database containing 648 KS and C-domains or a combination of profile-HMM based domain detection and BlastP searches (*see* **Note 6**). After trimming, they are inserted into a manually curated alignment of NRPS C-domains, which is used to reconstruct a phylogenetic tree.

(a) Input data for NaPDoS

NaPDoS accepts three types of input: protein sequences, coding DNA sequences or DNA sequences of (meta) genomes/contigs. All sequences have either to be copy/pasted to the query box or uploaded as FASTA files.

In addition, there is the possibility to directly submit NaPDoS jobs from the info-box of C- within antiSMASH.

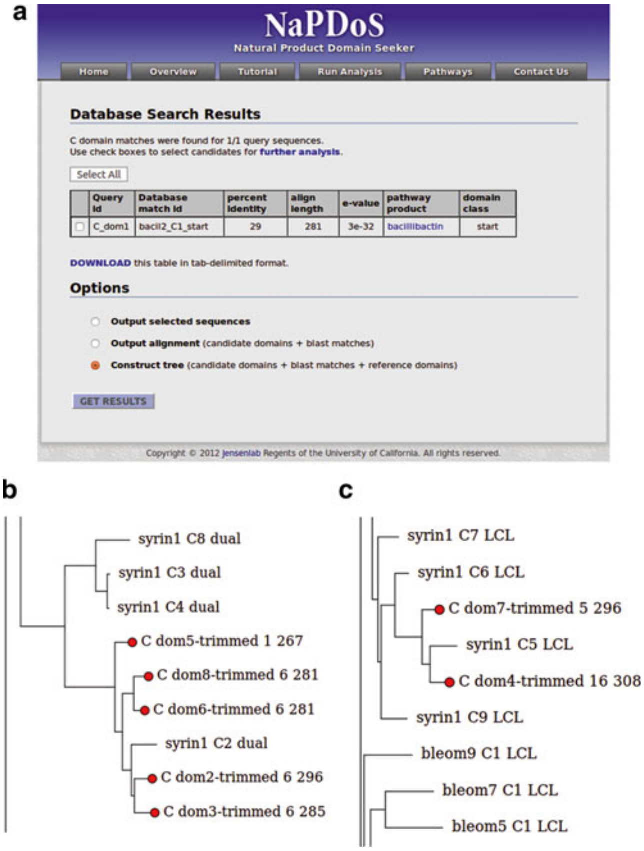
(b) Output of NaPDoS

The results of the domain identification step are displayed in a table at the first NaPDoS results page. This table contains information about the best hit(s), the identity of the query sequence with the hit, the length of the alignment, an *e*-value and an assignment to the next known pathway in the NaPDoS database. In addition, the classification of the domain (see above or Tutorial page) is shown. The table can also be downloaded as tabulator delimited text file.

Based on these results, the user can (1) select to download the identified and trimmed domain sequences in FASTA format, (2) select to download the alignment of the domain sequences with the best hit(s) of the NaPDoS database in PileUp-format, or use the alignment to calculate a maximum-likelihood tree, which can be downloaded as a SVG graphics file or as Newick formatted text (Fig. 3).

2.3 Bioinformatics tools to analyze NRPs

Norine (http://bioinfo.lifl.fr/norine/) [3, 28] is a unique resource dedicated to nonribosomal peptides that includes a database together with computational tools for data analysis. Norine currently contains more than 1100 peptides, coming from the



Database Search Results

C domain matches were found for 1/1 query sequences.
Use check boxes to select candidates for further analysis.

	Query id	Database match id	percent identity	align length	e-value	pathway product	domain class
□	C_dom1	bacil2_C1_start	29	281	3e-32	bacillibactin	start

DOWNLOAD this table in tab-delimited format.

Options

○ Output selected sequences
○ Output alignment (candidate domains + blast matches)
● Construct tree (candidate domains + blast matches + reference domains)

Copyright © 2012 Jensenlab Regents of the University of California. All rights reserved.

Fig. 3 C-domain subtypes predicted using NaPDoS. The complete sequence of cichofactin synthetase (including CifA and CifB) was analyzed. (**a**) The results presented as a table show that the first C-domain is predicted to be a C-starter. (**b-c**) Results presented as a phylogenetic tree for the seven remaining C-domains. They can be considered as $^{L}C_{L}$ (*see* part c) or dual-C/E (*see* part b) depending on the clade they are nested in

scientific literature or submitted directly by researchers. It provides detailed annotations such as structure, activity, producing organisms or bibliographical references. Norine also offers visualizer and editor for monomeric structures, as well as tools to search for monomeric structures [34, 35], which is a unique feature.

Norine can be queried either by annotations (through "general search" tab) or by structural information (through "structure search" tab) of the peptides. Searches among annotations are useful to obtain information about all peptides harboring a given activity or produced by a given organism, for example.

Structure-based search allows to find peptides containing given monomers with or without considering their 2D structure. In Norine, a specific format is used to represent the NRPs. It is based on the monomeric structure that is the monomers incorporated by the synthetases with the chemical bonds between them. The codes designating the more than 500 different monomers are mainly based on the IUPAC nomenclature (*see* **Note** 7 for more details). An editing applet is provided so that this structure can be easily drawn and the correct graph format is generated automatically (Fig. 4). Whatever the query (on annotations or structures), Norine returns a peptide list and supports several displays.

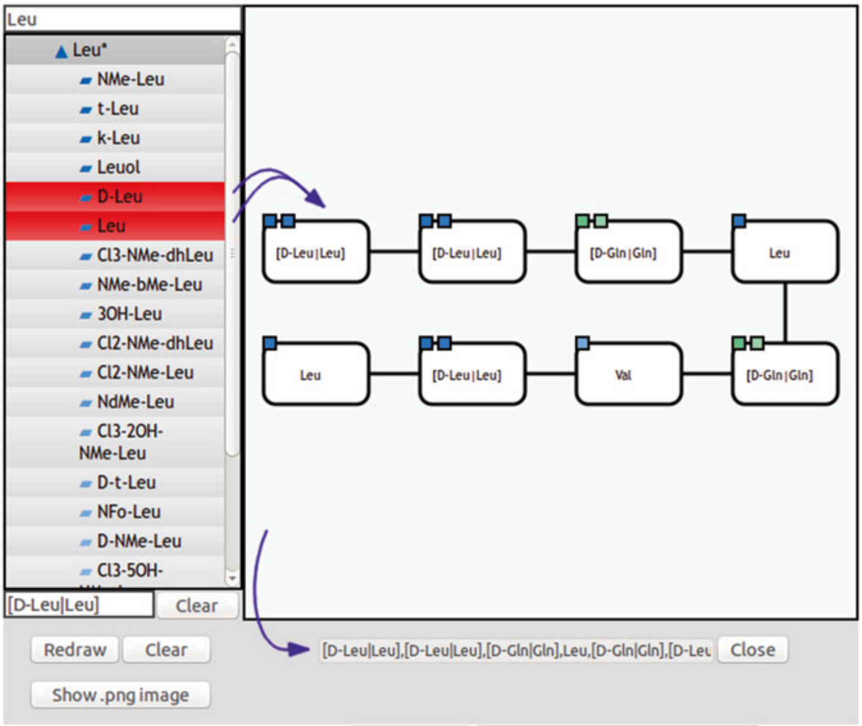
3 Methods

3.1 From Genome Sequence to NRPSs: Identification and Analysis of NRP Gene Clusters with antiSMASH

The following protocol is for antiSMASH version 2.

3.1.1 Submitting Analysis Jobs to antiSMASH

1. Open the antiSMASH start page http://antismash.secondarymetabolites.org in a modern Web-browser (e.g., Mozilla Firefox or Google Chrome).
2. Upload your own genomic data using the "Choose file" button or enter NCBI GenBank/RefSeq accession number in the "NCBI ACC#" field. The extension of the file must be ". fasta" for FASTA/multi-FASTA format, ".gbk" for GenBank format, or ".embl" for EMBL format. It is highly recommended to provide an e-mail address in the e-mail field to receive a notice and a direct link to the antiSMASH result when the computation is finished. If an e-mail address is not provided, it is important to bookmark the job status page, as this is the only way to access the results.
3. If the sequence to analyze is of fungal origin, select the "DNA of eukaryotic origin" checkbox; leave it empty if the sequence is of bacterial origin. A correct selection of this option is



Leu

[D-Leu|Leu]

[D-Leu|Leu],[D-Leu|Leu],[D-Gln|Gln],Leu,[D-Gln|Gln],[D-Leu

Fig. 4 Norine query using the "structure search" tool. The editor enables to draw candidate peptides. The monomers can be chosen from the menu at the left side of the editor and created by a simple click on its main window. The links between monomers are created by clicking on a monomer, then clicking on a different one that should be connected to the first. The graph representation shown under the editor is updated with every modification of the peptide

crucial, as it determines the gene prediction algorithm and also the model selection for some of the specificity predictors.

4. While the pre-selected options are suitable for most analysis requests, specific analysis options can be selected/deselected with the checkboxes.
5. Start antiSMASH job by pressing the "Submit" button at the end of the page. A new webpage showing the job status is displayed. Depending on the server load and the size of the genome, the antiSMASH analysis normally takes 2–4 h to complete.

3.1.2 Browsing the antiSMASH Results

The most important results of antiSMASH are displayed on an interactive webpage.

1. Once the computation is finished, you are redirected to an overview page with a list of all identified gene clusters of the analyzed genome. You can come back to this list by clicking on "Overview".
2. In the top panel "Select Gene Cluster" each identified gene cluster is represented as a colored circle. Each hit is also

referred in the table, including the compound type and its coordinates in the input sequence.

3. Clicking on the colored circles, or the colored "Cluster nb" (nb is the number of the cluster) labels displays the details of the analysis of the different gene clusters (Fig. 1).

3.1.3 Navigating the antiSMASH Cluster Result Page

1. In the top panel, an overview of the identified gene cluster and its coordinates are displayed. By clicking on "Show pHMM detection rules used" you can get a list of profile-HMMs that were used to identify the consulted gene cluster and assign its class.
2. Arrows represent each gene of a cluster. By clicking on these arrows, additional information is displayed, e.g., annotation from the EMBL/GenBank file; classification of the enzymes into SMCOGs (Secondary metabolite—clusters of orthologous genes), conserved protein domains/hits against the antiSMASH HMM profile database, direct links to NCBI blast and—if the sequence was downloaded from NCBI—the NCBI genome browser. Finally, there is a text box displaying the amino acid sequence of the gene product, which can be used to easily copy/paste the sequence to third party Web-servers.
3. For pathways containing PKS or NRPS genes, a detailed analysis of the domain organization is provided in the second panel. Additional information on the identified domains can be displayed by clicking on the respective cartoons. For NRPS A-domains, the results of the amino acid specificity prediction are included. For the C-domains, the precise catalytic function is provided (starter-C-domain, $^{L}C_{L}$, $^{D}C_{L}$, Dual-condensation-epimerization), together with a direct link to NaPDoS for more detailed analysis.
4. If the antiSMASH search was performed by submitting a whole genome/whole gene cluster sequence, homologous gene clusters of other organisms are identified by an integrated MultiGeneBlast [36] analysis against a custom database containing conserved operons involved in the biosynthesis of common secondary metabolite building-blocks, e.g., biosynthetic genes to produce non-proteinogenic amino acids like β-hydroxyphenylglycine. The results are displayed graphically in the bottom panel of the result page. Information about the hits can be obtained by clicking on the respective arrows representing similar genes. The color-coding is consistent for all results, i.e., similar genes always have the same color.
5. On the right side of the antiSMASH result page, a predicted core structure is displayed for PKS and NRPS gene clusters. Very important: The structure displayed is only a rough prediction of the core structure of the secondary metabolite and must not be confused with the structure of the final product of the pathway (*see* also **Note 8**).

6. Below the structure prediction panel, details of the substrate prediction are displayed for the individual genes/enzymes and detection methods (e.g., NRPSPredictor2 SVM, Stachelhaus code and Minowa) for each monomer if available; at the bottom of the panel a direct link to NORINE (*see* Subheading 2.3 or 3.4) is provided.

3.1.4 Downloading antiSMASH Results

1. antiSMASH-results on the Web-server antismash secondary metabolites.org are deleted after 4 weeks. Therefore, it is recommended to download the antiSMASH results to a local computer. antiSMASH currently provides four types of downloads, which can be accessed by pressing the <<down arrow>> button on top of the results page:
2. *Download all results*: Downloads a ZIP compressed file containing all analyses (including the files described below) and output files. To use this archive, uncompress it at a convenient directory. The most important files for further analysis are:
   (a) index.html: This file contains the antiSMASH result page, which can be viewed with a Web-browser (*see* **Note 3**) and contains exactly the same information than the online result page described above.
   (b) <accession-number>.final.embl and <accession-number>.final.gbk: These are EMBL/GenBank formatted summary files containing the detailed antiSMASH annotation (see below)
   (c) <accession-number>.clusterXXX.gbk: Annotated files of the identified gene clusters in GenBank format ("XXX" in the file name indicates cluster number.).
3. *Download XLS overview file*: Downloads a table with an overview on all identified gene clusters and the involved genes (accession numbers) in MS Excel format.
4. *Download EMBL summary file/Download GenBank summary file*: Download the sequence file (in EMBL or GenBank format) containing all the antiSMASH annotations. This file can be used with most programs allowing import of annotated sequence data, for example Artemis [37] or standard bioinformatics libraries like Bioperl [38] or BioPython [39].

3.2 Additional Analysis of NRPS Adenylation Domain Specificities with NRPSpredictor2

While antiSMASH provides a comprehensive analysis of secondary metabolite gene clusters on whole genome level, it sometimes may be of interest just to predict the monomers incorporated by NRPS proteins. In such cases, it is convenient to use the website of NRPSpredictor.

An overview on all options can be retrieved by clicking the "Help" button on top of the page.

3.2.1 Submitting NRPSpredictor2 Jobs

1. Open the NRPSpredictor webpage http://nrps.informatik.uni-tuebingen.de in the Web-browser.
2. Either copy and paste the FASTA-formatted amino acid sequence of the NRPS into the text box, or upload a FASTA-file via the "Choose file" button. Alternatively, paste the manually extracted Signature sequences into the search field/upload a file and select the "Signatures" checkbox as "Filetype".
3. As a standard, NRPSpredictor uses the bacterial models for prediction. If the sequence to analyze is of fungal origin, select "fungal" in the "Type of predictor" checkbox.
4. Start the NRPSpredictor calculation by pressing the "Submit" button.

3.2.2 Browsing NRPSpredictor2 Results

1. A table containing all the prediction results is displayed for every A-domain that was identified in the query sequence (*see* Subheading 2.2 for details on the output) (Fig. 2).
2. A text file containing a table of the results can be downloaded by clicking on the "Report file" link at the top of the NRPSpredictor2 results page.

3.3 Additional Analysis of NRPS Condensation Domains by Phylogenetic Classification with NaPDoS

In addition to the information provided by antiSMASH and NRPSpredictor, phylogenetic analyses of the NRPS can provide additional information on the function and the relation to known biosynthetic pathways.

A detailed description on all NaPDoS functions can be retrieved by clicking the "Tutorial" tab on the NaPDoS homepage.

3.3.1 Submitting NaPDoS Jobs

1. Open the NaPDoS webpage at http://napdos.ucsd.edu/ in the Web-browser.
2. Select the "Run Analysis" tab.
3. Select the domain type you want to analyze ("KS-domain" or "C-domain").
4. Select the query type depending on the data you want to analyze: choices are (1) "Predicted protein sequences (amino acid)" for amino acid queries; (2) "Predicted coding sequences or PCR products (DNA)" for DNA sequences of PKS or NRPS genes; (3) "Genome or metagenome contigs (DNA)" for DNA sequences of (meta)genomes.
5. Either copy and paste the query data into the "Query sequence" in protein or nucleotide FASTA format into text box, or select a file containing the data for upload.
6. Start NaPDos analysis (with standard parameters) by clicking the "SEEK" button, then "Submit Job" in the next page.

7. By clicking on "Advanced Settings" you have the possibility to influence the parameters used for domain identification, i.e., defining *e*-values, alignment lengths and number of displayed hits (*see* **Note 6**).

3.3.2 Browsing and Downloading the NaPDoS Results

1. Shortly after submitting the NaPDoS job, the results of the Database Search are displayed in a table. In "Genome or metagenome contigs (DNA)"-mode, first a table of the identified domains and their coordinates is presented. By clicking on "Get more info" a more detailed table is shown. If "Predicted protein sequences (amino acid)" or "Predicted coding sequences or PCR products (DNA)" was chosen as search mode, an equivalent detailed table containing information on the identified domains, similar pathways and a classification (*see* also Subheading 2) is displayed directly.
2. If a similar pathway was identified, information about this pathway can be obtained by clicking on the link in the table.
3. For further analyses, individual hits or all hits can be selected.
4. To download the table as tab-delimited text file use the "DOWNLOAD" link.
5. At the bottom of the page, options can be selected to download.
   (a) The identified NRPS C-domains in FASTA format.
   (b) The NaPDoS alignment containing the query sequences and the best hits against the NaPDoS database.
   (c) The phylogenetic trees as SVG graphics files or in Newick format for use in other phylogenetic software, for example Dendroscope [40].

3.4 Structure-based search against All NRPs Contained in Norine

This step can be performed either directly from the antiSMASH result pages by using the direct link to fingerprint search in Norine, or using the "structure search" tab of the Norine tool.

3.4.1 Submitting Structure Search to Norine

1. Open http://bioinfo.lifl.fr/norine in the Web-browser.
2. Select "structure search" tab on top of the page. The editor shown in Fig. 4 is open in the page.
3. Draw your peptide in monomer representation using the editor (Fig. 4).
   (a) Choose the monomers composing the query peptide. To search for a specific monomer, you can use the search box at the top of the menu at the left side of the editor. Clicking on one or several monomers will select them. It clusters them at the same position of the peptide, enabling, for example, searching for both isomers or for the monomer list obtained by analysing a NRPS. The list of all chosen

monomers is shown in the box at the bottom of the menu. Once the monomer(s) are chosen, click on the main window of the editor to place it.

(b) Add bonds between the monomers. To connect a monomer to another one, click on it, then go to the monomer you want to connect to. To create a double link, when the monomers are linked by two different chemical bonds, repeat the process. Clicking on the link will erase it.

4. The designed peptide is automatically translated in a graph representation (*see* **Note 7** for a detailed description) in the text box under the editor. Once the peptide is complete, click on "Submit" button. Two searches are performed in one submission. The structure-based search [34] outputs all the peptides containing exactly the complete query structure, called pattern. One pattern can be composed of one or more fragments. The "monomer composition fingerprint search" [35] is a matching performed by calculating a distance between the query and the peptides represented as fingerprints.

3.4.2 Browsing and Downloading Norine Structure Search Results

1. As a result, a table containing the peptides matching the query structure is displayed.
2. The table contains the names of the peptides, a calculated distance-metric between the peptide and the query, the number of matched fragments and their list. The results can be sorted by clicking on the arrows located in the column names.
3. The graphical representation allows filtering of the results by one of the following criteria at a time: putative or curated peptides, 2D structure types (linear, with one or two cycles, etc.), categories (peptide, peptaibol, lipopeptide, …), activities, number of monomers. Observation of the common features of the peptides similar to a query can help predicting some of its features such as its category, its 2D structure type or its activity.
4. The table view enables results manipulation and refinement by customizing the displayed columns.
5. The peptide lists can be downloaded in text (CSV that can be opened in any spreadsheet application), XML, or JSON formats.
6. Clicking on one line of the table in the distance or fragment count column, gives access to the comparison of the query and the concerned peptide.
7. Clicking on a peptide name gives access to the description of this peptide.

3.4.3 Browsing and Downloading Norine Peptide Description

1. Detailed annotations are provided in the description page of the peptide, about including general features, its structure, the producing organisms, relevant articles concerning their structure, links to other databases if available: producing synthe-

tases in UniProt [41], 3D structure in PDB [42], chemical structure in PubChem [43].

2. Annotations can be downloaded in XML or JSON format.

3.5 Study Cases

3.5.1 Studying the Cichofactin Biosynthetic Gene Cluster to Predict its Produced Peptide

In this study case, we assume, that nothing is known about the genes/functions encoded.

1. *Analyzing the cichofactin NRPS domain organization and predicting the produced peptide with antiSMASH*

   If a comprehensive analysis of an unknown biosynthetic gene is required, antiSMASH provides a huge collection of different tools for this task. In this study case, we use the pre-annotated sequence of the cichofactin biosynthetic gene cluster directly downloaded into antiSMASH.
2. Start an antiSMASH analysis at http://antismash.secondarymetabolites.org by entering your e-mail address in the "Email-address" field and the accession number "*KJ513093*" (=cichofactin biosynthetic gene cluster) in the search field and clicking on the submit button.
3. After the analysis is completed, the antiSMASH cluster overview page is displayed. As our demo sequence only contains one cluster, there is only one entry in the table. Click on the hit "Cluster 1".
4. The result page of the antiSMASH analysis is displayed (Fig. 1).
5. To get information on the identified gene click on arrows in the "Gene cluster description" panel. From the information window, you can directly forward the sequence to NCBI Blast or start the NCBI genome viewer (the latter only works for sequences downloaded from GenBank).
6. In the "Detailed annotation" panel, the domain organization of the cichofactin NRPSs is displayed. The cluster is composed of two NRPS coding genes. The first one, called AHZ34232, encodes three modules, and the second one, called AHZ34233, encodes five modules (numbered from 4 to 8).
7. To get information on the catalytic function of the C-domains, click on the respective C-domain cartoon. A pop-up window with additional information opens. From this pop-up, you can also directly submit the extracted domain sequence to NaPDos using the "Analyze with NaPDoS" link. The predictions are, in nearly all cases, the same between antiSMASH and NaPDoS, but NaPDoS additionally provides a phylogenetic analysis of the domain. The first gene starts with a "Condensation_Starter" that may incorporate a lipid moiety at the beginning of the produced NRP. The two other C-domains of this gene are annotated as "Condensation_Dual" which means that they epimerize the amino acid they add to the growing peptide

(selected by the previous module). The second gene ends by two thioesterase (TE) domains. This feature is mainly observed in NRPSs producing lipopeptides. In this protein, the modules 4, 6 and 8 contains "Condensation_Dual" C-domains. The two genes form a complete and coherent synthetase expected to produce a lipopeptide composed of eight amino acids with a the first, second, third, fifth and seventh ones in D-configuration.

8. At the right side of the result page, you find the prediction of the core structure and a text report on the various specificity predictions. The amino acid composition of the predicted peptide is reported as (leu-leu-gln) + (leu-nrp-val-leu-leu), where nrp stands for undetermined amino acid that is assigned when the three integrated predictors (Minowa, NRPSpredictor, Stachelhaus code) disagree.

3.5.2 Confirming the A-Domain Specificities of the Cichofactin NRPS Proteins with NRPSPredictor

1. Download the amino acid sequences of the cichofactin NRPS CifA and CifB from NCBI.
   (a) Open http://www.ncbi.nlm.nih.gov/ in your Web-browser.
   (b) Select "Protein" at the selection box on top of the screen.
   (c) Search for "cichofactin nrps"; you should get two results, "CifA" and "CifB".
   (d) Select both sequences by clicking on the checkbox.
   (e) Press the "Display Settings" Link, which is displayed on top of the results list; select "FASTA(text)" and confirm by clicking "Apply".
2. Open a second browser window/tab and go to http://nrps.informatik.uni-tuebingen.de
3. Copy the sequences into the text NRPSpredictor "Sequence to analyse" text-box.
4. Start NRPSpredictor analysis with "Submit Job".
5. The results of NRPSpredictor are displayed as separate tables for each identified A-Domain (*see* Fig. 2 for the two first A-domains of CifB). The outputs are more detailed than the ones obtained in antiSMASH and also include statistic information. We can notice that a low score (displayed in orange) is obtained for the second A-domain of the CifB protein. The "small cluster" hit is "asp,asn" and has a score of only 0.140582 (the lowest is 0), while the "nearest neighbour" (equivalent to "Stachelhaus code"-results in antiSMASH) is "gln" and has a score of 70 % (best is 100 %). Those scores are congruent with the fact that the three predictors give different results in antiSMASH, so with the "nrp" prediction for the fifth amino acid of the peptide.

3.5.3 Comparing the Predicted Monomeric Structure to Known NRPs of Norine

The features described in this chapter correspond to the Norine version of November 2015.

1. A direct link to Norine is provided at the right end of the antiSMASH results page, in the frame "Database cross-links". This link automatically generates a pattern search of the two peptide fragments produced by the two studied NRPS proteins, including the potential derivatives of the monomers (for example their D-configuration). A similarity between the fingerprints is also calculated.
2. Click on the graphical output icon to obtain an overview on the NRPs that contain the peptides most similar to the query by selecting them. In our study case, only lipopeptides are found confirming the prediction from the NRPS with the presence of a C-starter. As the peptides most similar to the query share a linear structure and a surfactant activity, those features can be assigned to the studied peptide.
3. In particular, the six members of the syringafactin family have a similarity higher than 0.7 with the combination of the two fragments. By clicking on the table output icon, interesting features of those peptides, such as the 2D structure, can be visualized together. They are all produced by the same strain, *Pseudomonas syringae* pv. tomato DC3000 [44] and are composed of nine monomers, including a lipid moiety that can vary. Their structures vary also in the seventh position where Val, Leu or Ile are observed. This variation occurs because the A-domain can select different amino acids, according to the substrates available in the culture media. Those amino acids are part of the "small cluster" in NRPSPredictor2 results.
4. The monomeric composition of cichofactin was determined experimentally by Pauwelyn et al. [20] as Fatty-acid_Leu_Leu_Gln_Leu_Gln_Val_Leu_Leu. So the unpredicted X was demonstrated to be a glutamine (Gln) pointing out a single difference with syringafactin, in which a threonine (Thr) is found at the same sixth position (the first monomer being the fatty acid). The bioinformatic study of the NRPS genes/proteins also considering the conformation of the amino acids allows predicting the following linear structure: Fatty-acid_D-Leu_D-Leu_D-Gln_Leu_D-Gln_Val_D-Leu_Leu.

3.5.4 Seeking for NRPS Genes in a Draft Genome of Pseudomonas

In this study case, we propose a workflow to infer the presence of NRPS genes in draft genome or metagenomic sequences. Those sequences have the particularities to be split in many fragments. They are contained in multi-FASTA files (each fragment is separated from the previous one with a line starting by > character). The studied sequences are from *Pseudomonas syringae pv. tabaci* str. ATCC 11528 [45], with GenBank assembly ID GCA_000159835.2.

1. *Screening for the presence of NRPS genes with NaPDoS*

   (a) Open NaPDoS (at http://napdos.ucsd.edu/), and select "Run Analysis" tab. Select "C domains" under Domain type and "Genome or metagenome contigs (DNA)" as "Query Type". Paste the DNA sequences of all contigs into the text box or upload the corresponding file in multi-FASTA format. Start analysis with "Seek" and then "Submit job" in the second page if the parameters are correct.

   (b) The result table provides the list of contigs containing C-domains in the "parent seq" column. Clicking on "Get more information", gives more details and the domains sorted by contig order. So it is easily possible to manually extract the contigs with at least one C-domain from the multi-FASTA file. In our study case, 21 C-domains were predicted, in 20 different contigs.

   (c) With a text editor software, search for the contig names and copy paste the corresponding sequence in another file. You can also extract two (or more) contigs before and after, and the ones between two close contigs. For example, NaPDoS finds C-domains in the following contigs (the numbers refer to the order in the genome): ACHU02000685, ACHU02000688, ACHU02000691. So you can extract all contigs between ACHU02000683 and ACHU02000693. This gives 59 contigs in our study case.

2. *Submission of the most interesting contigs to antiSMASH*

   The extracted sequences can then be submitted to antiSMASH.

   (a) In the submission form, give your e-mail and the fragments in multi-Fasta file. Open the "Restrict which of the 24 supported secondary metabolite types to detect" parameter to select only "nonribosomal peptides" as you extract contigs related to C-domains. This selection reduces the calculation time.

   (b) In the study case, 19 clusters (2 of the type "Other") are annotated among the 59 extracted.

   (c) Those contigs only consist of truncated NRPS genes with two to four domains. Cluster 1 and 11 also contain other genes such as a dioxygenase or an ABC transporter. The observation of the "Homologous gene clusters" given help in identifying the possible category of the peptide.

   (d) Genes in cluster 1 align with clusters of other *pseudomonas* strains annotated as pyoverdin producers so we can presume that the studied strain also produce a pyoverdin. But we cannot predict its structure or monomeric composition as the NRPS genes are incomplete. The number of genes in cluster 11 is too small to predict a peptide category.

(e) In clusters 14, 16, 17, and 19, dual C/E condensation domains are predicted. Those domains are, until now, only observed in lipopeptides. So we can infer that the studied strain also produces a lipopeptide.

4 Notes

1. The use of these Web services is intended for the analysis of a limit number of genomes or protein sequences. If you plan do large-scale analyses, for example analyze all NRPS sequences from GenBank or analyze large metagenomic datasets, please contact the providers/authors before submitting the jobs via the website or scripts, as there may be better ways to run such big analyses independent of the Web service.

2. The quality of the input data has very important effect on the quality of the antiSMASH predictions (and generally of all predictive tools): As antiSMASH uses rules to detect the biosynthetic gene clusters require the presence of conserved domains (e.g., a NRPS is identified by detecting condensation, adenylation and PCP domains encoded by in the same gene), good predictions can only be made if the input data has a sufficient length. When using antiSMASH with draft sequence data, sequence scaffolds should always be preferred over analyzing the contig data. The use of antiSMASH with metagenomic datasets is technically possible, but due to the mostly very short length of the assembled sequences, most gene clusters will be missed as the conserved domains used for cluster and biosynthetic type identification are encoded on different contigs. Therefore it is disencouraged to use antiSMASH for this kind of analyses.

3. antiSMASH can be used with any major up-to date Web-browsers, like for example Mozilla Firefox or Google Chrome. However, for browsing downloaded antiSMASH results, it is recommended to use Mozilla Firefox, as very strict Javascript security settings in Google Chrome prevent the interactive display of the antiSMASH (Sub)Clusterblast results for locally saved HTML pages.

4. The signature data should be prepared in a text editor and uploaded as a file or transferred into the text box by copy/paste as pressing the [tab] key in most Web-browsers switches to the next element of the HTML page instead of inserting a "tabulator" character to the text in the text box.

5. The precision (=number of true positive hits/(number of true positives + false positives) on the validation test set of NRPS sequences) is a measurement of the prediction quality of the hit-SVM model.

6. Depending on the type of input, NaPDoS uses a blast or HMM based approach to identify the C-domains. Thus, the results obtained may differ, i.e., there may be domains identified with one method while the standard thresholds for the domain identification criteria prevent the detection with the other method. It therefore is highly recommended to try out different parameters. For some classes of NRPS (e.g., glycopeptides) it is worth to select less stringent domain identification criteria (i.e., shorter minimal alignment size, increased *e*-value) in the NaPDoS "Advanced Settings" options of the "Run Analysis" page.
7. Description of the formatted strings adopted by Norine for the monomeric structures: the monomers composing the peptide are separated by commas. The monomers are symbolized by three letter codes for amino acids and carbohydrates, with their substituent groups or configuration if needed; lipid numbers for long carbon chains such as fatty acids; commonly used short names for other compounds. For pattern search, wildcard is allowed: X stands for any monomer, Code* for the derivatives of the 'Code' monomer, names of the clusters defined in Norine (monomers are clustered based on their chemical properties, for example "fatty acid") (for a complete list consult the "monomers list"), personal set of alternatives represented between brackets by a monomer list separated by "|" (pipe character). The links between monomers are cited after the monomer sequence, isolated by "@". The monomers are ranked from 0 in the order they appear in the sequence and their respective neighborhoods are displayed.
8. Currently, it is not yet possible to predict an accurate chemical structure of the final product just based on sequence analysis. The structure displayed in antiSMASH is solely based on the prediction of the substrate specificities of PKS and NRPS; in the current version antiSMASH2, it does not take into account information about the stereochemistry, any post-PKS/post-NRPS modifications or unusual features of the PKS/NRPS enzymes. As such modifications are very common, the displayed structure must only be regarded as a hint on the final product of the biosynthetic pathway.

Acknowledgements

TW is supported by a grant of the Novo Nordisk Foundation. The work on Norine is supported by PPF Bioinformatique of Lille 1 University and Bilille.

We thank Hyun Uk Kim for testing the workflows and careful reading of the chapter.

References

1. Sieber SA, Marahiel MA (2003) Learning from nature's drug factories: nonribosomal synthesis of macrocyclic peptides. J Bacteriol 185:7036–7043
2. Caboche S, Leclère V, Pupin M et al (2010) Diversity of monomers in nonribosomal peptides: towards the prediction of origin and biological activity. J Bacteriol 192:5143–5150
3. Caboche S, Pupin M, Leclère V et al (2008) NORINE: a database of nonribosomal peptides. Nucleic Acids Res 36:D326–D331
4. Conti E, Stachelhaus T, Marahiel MA et al (1997) Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. EMBO J 16:4174–4183
5. Fedorova ND, Moktali V, Medema MH (2012) Bioinformatics approaches and software for detection of secondary metabolic gene clusters. Methods Mol Biol. 944:23–45
6. Weber T (2014) In silico tools for the analysis of antibiotic biosynthetic pathways. Int J Med Microbiol. 304:230–235
7. Boddy CN (2014) Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides. J Ind Microbiol Biotechnol. 41(2):443–50
8. Stachelhaus T, Mootz HD, Marahiel MA (1999) The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chem Biol 6:493–505
9. Challis GL, Ravel J, Townsend CA (2000) Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. Chem Biol 7: 211–224
10. Bachmann BO, Ravel J (2009) Chapter 8 Methods for in silico prediction of microbial polyketide and nonribosomal peptide biosynthetic pathways from DNA sequence data. In: Hopwood DA (ed) Methods in enzymology. Academic, New York, pp 181–217
11. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763
12. Minowa Y, Araki M, Kanehisa M (2007) Comprehensive analysis of distinctive polyketide and nonribosomal peptide structural motifs encoded in microbial genomes. J Mol Biol 368:1500–1517
13. Prieto C, García-Estrada C, Lorenzana D et al (2012) NRPSsp: non-ribosomal peptide synthase substrate predictor. Bioinformatics 28: 426–427
14. Röttig M, Medema MH, Blin K et al (2011) NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res 39:W362–W367
15. Rausch C, Weber T, Kohlbacher O et al (2005) Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). Nucleic Acids Res 33:5799–5808
16. Baranašić D, Zucko J, Diminic J et al (2014) Predicting substrate specificity of adenylation domains of nonribosomal peptide synthetases and other protein properties by latent semantic indexing. J Ind Microbiol Biotechnol. 41:461–467
17. Rausch C, Hoof I, Weber T et al (2007) Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. BMC Evol Biol 7:78
18. Ziemert N, Podell S, Penn K et al (2012) The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One 7:e34064
19. Caradec T, Pupin M, Vanvlassenbroeck A et al (2014) Prediction of monomer isomery in Florine: a workflow dedicated to nonribosomal peptide discovery. PLoS One 9:e85667
20. Pauwelyn E, Huang C-J, Ongena M et al (2013) New linear lipopeptides produced by *Pseudomonas cichorii* SF1-54 are involved in virulence, swarming motility, and biofilm formation. Mol Plant Microbe Interact 26:585–598
21. Blin K, Medema MH, Kazempour D et al (2013) antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res 41:W204–W212
22. Medema MH, Blin K, Cimermancic P et al (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
23. Weber T, Blin K, Duddela S et al (2015) antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucl Acids Res 43:W237–W243. doi:10.1093/nar/gkv437
24. Starcevic A, Zucko J, Simunkovic J et al (2008) ClustScan: an integrated program package for the semi-automatic annotation of modular biosynthetic gene clusters and in silico prediction of novel chemical structures. Nucl Acids Res 36:6882–6892. doi:10.1093/nar/gkn685
25. Li MH, Ung PM, Zajkowski J et al (2009) Automated genome mining for natural products. BMC Bioinformatics 10:185. doi:10.1186/1471-2105-10-185
26. Anand S, Prasad MVR, Yadav G et al (2010) SBSPKS: structure based sequence analysis of

polyketide synthases. Nucl Acids Res 38:W487–W496. doi:10.1093/nar/gkq340

27. Ansari MZ, Yadav G, Gokhale RS, Mohanty D (2004) NRPS-PKS: a knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucl Acids Res 32:W405–W413. doi:10.1093/nar/gkh359
28. Flissi A, Dufresne Y, Michalik J, et al (2016) Norine, the knowledgebase dedicated to non-ribosomal peptides, is now open to crowd-sourcing. Nucl Acids Res (in press)
29. Delcher AL, Harmon D, Kasif S et al (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Res 27: 4636–4641
30. Majoros WH, Pertea M, Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20:2878–2879
31. Pruitt KD, Tatusova T, Brown GR et al (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. Nucleic Acids Res 40:D130–D135
32. Benson DA, Clark K, Karsch-Mizrachi I et al (2014) GenBank. Nucleic Acids Res 42: D32–D37
33. Schölkopf B, Platt JC, Shawe-Taylor J et al (2001) Estimating the support of a high-dimensional distribution. Neural Comput 13:1443–1471
34. Caboche S, Pupin M, Leclère V et al (2009) Structural pattern matching of nonribosomal peptides. BMC Struct Biol 9:15
35. Abdo A, Caboche S, Leclère V et al (2012) A new fingerprint to predict nonribosomal peptides activity. Journal of computer-aided molecular design 26:1187–1194
36. Medema MH, Takano E, Breitling R (2013) Detecting sequence homology at the gene cluster level with MultiGeneBlast. Mol Biol Evol 30:1218–1223
37. Rutherford K, Parkhill J, Crook J et al (2000) Artemis: sequence visualization and annotation. Bioinformatics 16:944–945
38. Stajich JE, Block D, Boulez K et al (2002) The Bioperl toolkit: perl modules for the life sciences. Genome Res 12:1611–1618
39. Cock PJA, Antao T, Chang JT et al (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423
40. Huson DH, Richter DC, Rausch C et al (2007) Dendroscope: an interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:460. doi:10.1186/1471-2105-8-460
41. The UniProt Consortium (2013) Update on activities at the Universal Protein Resource (UniProt) in 2013. Nucleic Acids Res 41: D43–D47
42. Berman HM, Kleywegt GJ, Nakamura H et al (2013) The future of the protein data bank. Biopolymers 99:218–222
43. Bolton EE, Wang Y, Thiessen PA et al (2008) PubChem: integrated platform of small molecules and biological activities. In: Wheeler RA, Spellmeyer DC (eds) Annual reports in computational chemistry. Elsevier, Amsterdam, pp 217–241
44. Berti AD, Greve NJ, Christensen QH et al (2007) Identification of a biosynthetic gene cluster and the six associated lipopeptides involved in swarming motility of Pseudomonas syringae pv. tomato DC3000. J Bacteriol 189:6312–6323
45. Studholme DJ, Ibanez SG, MacLean D et al (2009) A draft genome sequence and functional screen reveals the repertoire of type III secreted proteins of Pseudomonas syringae pathovar tabaci 11528. BMC Genomics 10:395