Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

# Recent development of computational resources for new antibiotics discovery

Hyun Uk Kim[1,2], Kai Blin[1], Sang Yup Lee[1,2] and Tilmann Weber[1]

Understanding a complex working mechanism of biosynthetic gene clusters (BGCs) encoding secondary metabolites is a key to discovery of new antibiotics. Computational resources continue to be developed in order to better process increasing volumes of genome and chemistry data, and thereby better understand BGCs. In this context, this review highlights recent advances in computational resources for secondary metabolites with emphasis on genome mining, compound identification and dereplication as well as databases. We also introduce an updated version of Secondary Metabolite Bioinformatics Portal (SMBP; http://www.secondarymetabolites.org), which we previously released as a curated gateway to all the computational tools and databases useful for discovery and engineering of secondary metabolites.

**Addresses**
[1] The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Dk 2800 Kgs. Lyngby, Denmark
[2] Department of Chemical and Biomolecular Engineering & BioInformatics Research Center, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea

Corresponding author: Weber, Tilmann (tiwe@biosustain.dtu.dk)

**Current Opinion in Microbiology** 2017, **39**:113–120

This review comes from a themed issue on **Antimicrobials**

Edited by **Olga Genilloud** and **Rolf Müller**

http://dx.doi.org/10.1016/j.mib.2017.10.027



## Introduction

Bacterial, fungal and plant secondary metabolites have served as major sources of antibiotics among many other medical and environmental applications. Secondary metabolite producers typically contain biosynthetic gene clusters (BGCs) in their genomes wherein multiple genes are located close to one another and encode secondary metabolites upon expression [1]. When the first whole genome sequences of secondary metabolite producers (i.e., *Streptomyces* species) became available, it became evident that these organisms biosynthesize a much greater number of secondary metabolites than initially expected based on compounds that had already been isolated from these model strains [2,3,4]. Thus, mining the genome data can deliver an excellent estimation of the genetic potential of a strain with respect to its biosynthesis of secondary metabolites. This initial discovery eventually led to the development of computational resources that facilitate the genome mining process [5]. Now that a massive volume of genomic data continue to be generated, use of computational genome mining tools has become essential to process the data and draw insights on secondary metabolites. Also, along with increasing volumes of chemical data (i.e., mass spectrometry (MS) data), genome mining process started to be more systematically supported with cheminformatic dereplication (i.e., identification and disregard of known compounds) tools. Databases can also be very useful in executing every process of new antibiotics discovery, which have been established with data of BGCs and antibiotics/secondary metabolites accumulated over the years. All these computational resources make up a pipeline for efficient discovery of new antibiotics (Figure 1).

In this context, this review highlights recent advances in computational resources developed past two years (since 2015) that facilitate new antibiotics discovery. This review therefore complements previous reviews on computational resources for secondary metabolite studies covering earlier years [6,7,8]. In particular, three types of computational resources are discussed herein, namely 'genome mining tools', 'cheminformatic compound identification and dereplication tools' and 'databases for BGCs and antibiotics/secondary metabolites' (Table 1). We also introduce an updated version of Secondary Metabolite Bioinformatics Portal (SMBP), which we previously released as a curated gateway to all the computational tools and databases useful for discovery and engineering of secondary metabolites [8].

## Genome mining tools

Genome mining tools scan an entire genome of interest, and look for specific sequence patterns that are indicative of a specific type of BGC (Figure 1). The two most comprehensive genome mining tools currently available are 'antibiotics and Secondary Metabolite Analysis SHell' (antiSMASH) [9,10,11,12,13••] and 'PRediction Informatics for Secondary Metabolomes' (PRISM) [14,15,16•]. In common, these two tools accept genome data as inputs (FASTA or GenBank), and generate a list of BGCs detected that are presented via web interface [17]. Both use pre-constructed Hidden Markov Models (HMMs) to detect BGCs, but HMMs were trained with different sets of BGC sequences. Plus, they deploy own additional

**Table 1**

**List of computational resources developed last two years (since 2015) for new antibiotics discovery. Only those accessible online and/or available as executable software tools are listed. Some resources belong to more than one category.**

| Name of computational resources | Comments | URL | Reference |
|---|---|---|---|
| SMBP (Secondary Metabolite Bioinformatics Portal) | A comprehensive portal for computational resources on secondary metabolites | http://www.secondarymetabolites.org | [8]; This Review |
| **Genome mining tools** | | | |
| *Comprehensive* | | | |
| antiSMASH (antibiotics and Secondary Metabolite Analysis SHell) [9,10,11,12,13••] | Version 4 released | https://antismash.secondarymetabolites.org/ | |
| ARTS (Antibiotic Resistant Target Seeker) | Prioritizes detected BGCs and identifies drug targets; Uses antiSMASH | https://arts.ziemertlab.com/ | [24] |
| GNP (Genomes-to-Natural Products) | Detects BGCs and additionally processes LC-MS/MS data for dereplication; Can also be considered as a dereplication tool | https://magarveylab.ca/gnp/ | [25•] |
| PhytoClust | Detects BGCs in plant genomes; Uses antiSMASH | http://phytoclust.weizmann.ac.il/ | [22] |
| plantiSMASH (antiSMASH for plants) | Available as a part of the antiSMASH web server | http://plantismash.secondarymetabolites.org/ | [21•] |
| PRISM (PRediction Informatics for Secondary Metabolomes) | Version 3 released | https://magarveylab.ca/prism | [14,15,16•] |
| SeMPI (Secondary Metabolite Prediction and identification) | Predicts structures of secondary metabolites biosynthesized by type I modular PKS; Uses antiSMASH and StreptomeDB 2.0; Can also be considered as a dereplication tool | http://www.pharmaceutical-bioinformatics.de/SeMPI/ | [23] |
| *BGC-specific* | | | |
| CASSIS (Cluster Assignment by Islands of Sites) | Detects BGCs in eukaryotic genomes | https://sbi.hki-jena.de/cassis/ | [18] |
| RiPPMiner | Predicts chemical structures of RiPPs | http://www.nii.ac.in/rippminer.html | [27] |
| RODEO (Rapid ORF Description and Evaluation Online) | Detects RiPP BGCs | http://www.ripprodeo.org/ | [20] |
| SANDPUMA (Specificity of AdenylatioN Domain Prediction Using Multiple Algorithms) | Predicts substrate specificities of adenylation domains of NRPS | https://bitbucket.org/chevrm/sandpuma | [19] |
| SBSPKSv2 (Structure based sequence analysis of PKS and NRPS) | Allows various chemical analyses for experimentally characterized PKS/NRPS BGCs; Version 2 released | http://www.nii.ac.in/sbspks2.html | [26] |
| **Cheminformatic compound identification and dereplication tools** | | | |
| DEREPLICATOR | Implements high-throughput identification of peptidic natural products | Standalone: http://cab.spbu.ru/software/dereplicator/ Web: http://gnps.ucsd.edu/ | [31] |
| GARLIC (Global alignment for natural products cheminformatics) | Relates a detected BGC to a known secondary metabolite | Web: http://www.magarveylab.ca/garlic Source code: https://github.com/magarveylab/garlic-release | [34••] |
| GNPS (Global Natural Products Social Molecular Networking) | Allows analysis, storage and sharing of raw MS/MS data | http://gnps.ucsd.edu/ | [30••] |
| GRAPE (Generalized retrobiosynthetic assembly prediction engine) | Predicts building block monomers of a given secondary metabolite | Source code: https://github.com/magarveylab/grape-release | [34••] |
| iSNAP (Informatic search strategy for natural products) | Implements highly accurate dereplication of NRPs | https://magarveylab.ca/analogue | [32,33] |