The secondary metabolite bioinformatics portal
Computational tools to facilitate synthetic biology of secondary metabolite production

Weber, Tilmann; Kim, Hyun Uk

Published in:
Synthetic and Systems Biotechnology

Link to article, DOI:
10.1016/j.synbio.2015.12.002

Publication date:
2016

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

Link back to DTU Orbit

Citation (APA):
Weber, T., & Kim, H. U. (2016). The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. Synthetic and Systems Biotechnology, 1(2), 69-79. https://doi.org/10.1016/j.synbio.2015.12.002
The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production

Tilmann Weber, Hyun Uk Kim

Abstract

Natural products are among the most important sources of lead molecules for drug discovery. With the development of affordable whole-genome sequencing technologies and other ‘omics tools, the field of natural products research is currently undergoing a shift in paradigms. While, for decades, mainly analytical and chemical methods gave access to this group of compounds, nowadays genomics-based methods offer complementary approaches to find, identify and characterize such molecules. This paradigm shift also resulted in a high demand for computational tools to assist researchers in their daily work. In this context, this review gives a summary of tools and databases that currently are available to mine, identify and characterize natural product biosynthesis pathways and their producers based on ‘omics data. A web portal called Secondary Metabolite Bioinformatics Portal (SMBP) is introduced to provide a one-stop catalog and links to these bioinformatics resources. In addition, an outlook is presented how the existing tools and those to be developed will influence synthetic biology approaches in the natural products field.

1. Introduction

Antimicrobial resistance is projected to be one of the major global challenges for maintaining our future health systems. According to the report commissioned by the Department of Health of the UK government, chaired by the economist Jim O’Neill, the global economic costs of antimicrobial resistance will result in more than 10 million annual deaths, leading to a loss of 2.0–3.5% of the world gross domestic product equivalent to 60–100 trillion USD by 2050 [e.g., references 1–3]. While this report may predict a worst-case scenario, it is clear that the problem of antimicrobial resistance has to be urgently addressed globally. As there will be no simple single solution, efforts have to be undertaken in various fields, for example in optimizing hygiene, access to clear water, vaccinations, increased efforts to prevent infections, or reduced use of antibiotics families that are used in human medicine and feedstock. Another important challenge will be to develop novel antimicrobial therapies and drugs.

Historically, natural products have been the major source of lead compounds for antimicrobial drugs, but also are used in other application fields, such as anti-cancer drugs, insecticides, anthelmintics, painkillers, flavors, cosmeceuticals and crop protection. Nevertheless, most big pharma companies have severely reduced their research efforts on natural products during the last 20 years due to high rediscovery rates of known molecules and a lack of innovative screening approaches. Therefore, it is surprising that still the majority of newly approved small-molecule drugs are natural products or their derivatives.

With the broad availability of ‘omics technologies, we currently experience a paradigm shift in natural product research; for decades, the only way to get access to new compounds was to cultivate antibiotics-producing microorganisms, mainly fungi and bacteria, under different growth conditions, and then isolate and characterize the compounds with sophisticated analytical
chemistry. Nowadays, ‘omics approaches offer complementary access to natural products; by identifying natural product/secondary metabolite biosynthetic gene clusters (BGCs), it is possible to assess the genetic potential of producer strains and to more effectively identify previously unknown metabolites. While this approach has led to some renaissance of natural product research in academia and industry, this information will also be the basis to rationally engineer molecules or develop "designer molecules" using synthetic biology approaches in the future.

When the first whole genome sequences of the model streptomycete Streptomyces coelicolor A3(2)\(^1\) and the avermectin producer Streptomyces avermitilis\(^{13,14}\) were determined, both strains were found to possess more secondary metabolite BGCs than an initial estimation made based on the number of their already known secondary metabolites. This is especially remarkable as both strains have served as model organisms and – in the case of S. avermitilis – industrial production strains for many years and thus have been studied by many researchers all over the world. With the rise of novel sequencing technologies and a growing number of microbial whole genome sequences, it became evident that a high number of BGCs is a common feature among various groups of bacteria, for example actinomycetes.\(^\)\(^{12}\)

Although the diversity of natural product chemical scaffolds is vast, the biosynthetic principles are highly conserved for many secondary metabolites. There is a set of enzyme families, which are often and very specifically associated with the biosynthesis of different classes of secondary metabolites. Thus, sequence information of these known gene families can be used to mine genomes for the presence of secondary metabolite biosynthetic pathways.

There are two principal strategies in the implementation of bioinformatic tools. Rule-based approaches can be used to identify gene clusters encoding known biosynthetic routes with high precision. In the first step of the mining process, these tools identify genes encoding conserved enzymes/protein domains that have associated roles in secondary metabolism, for example the "condensation (C)," "adenylation (A)" and "peptidyl carrier protein (PCP)" domains of non-ribosomal peptide synthetases (NRPs). In the second step, predefined rules are used to associate the presence of such hits with defined classes of natural products. In the above example, a NRPS BGC can be simply and unambiguously identified if genes are present that code for at least one C-, A- and PCP domain. More complex rules may take into account whether specific genes are encoded in close proximity, for example type II polyketide BGCs can be detected using a rule that evaluates whether a ketosynthase α, a ketosynthase β/chain length factor and acyl-carrier protein are encoded by 3 individual genes in direct proximity. Such rule-based search strategies are, for example, implemented as one option in the pipeline antibiotic and Secondary Metabolite Analysis SHell (antSMASH),\(^{17,18}\) which, currently in its version 3, can detect 44 different classes of BGCs. Especially, clusters containing modular polyketide synthase (PKS) or NRPS genes can be easily detected by scanning the genome for genes that encode their characteristic enzyme domains, as also implemented in NaPDos,\(^{16}\) NP_searcher,\(^{17}\) GNP/PRISM,\(^{18}\) and SMURF.\(^{19}\) All these approaches are very precise in detecting gene clusters of known families and classes of which rules can be defined. Based on the prerequisite to have defined rules, these algorithms cannot detect novel pathways that use a different biochemistry and enzymes. To avoid this limitation, also rule-independent methods, which are less biased, have been developed, for example implemented in ClusterFinder\(^{20}\) and EvoMining\(^{21}\) (see below for details on how they work). These tools use machine learning-based approaches or automated phylogenomics analyses to make their predictions. For fungi, algorithms that evaluate transcriptome data can also efficiently predict clusters of co-transcribed genes.\(^{22}\)

As computational approaches to natural product discovery are rather a new and dynamic field, we intend to give an overview on existing computational tools and databases that help scientists solve the abovementioned tasks and develop perspectives on how these approaches will change the discovery of new natural products (Fig. 1).

## 2. Computational tools for natural product research

Recently, several reviews have been published, describing different strategies employed by the genome mining tools commonly used to detect secondary metabolite BGCs [e.g., references\(^{23-28}\)]. In this review, we therefore give a summarizing, but comprehensive up-to-date overview on the tools and databases that are currently available for mining for BGCs, analyzing biosynthetic pathways, combining genomic and metabolomic data, and generating genome-scale metabolic models of the secondary metabolite producers (Tables 1 and 2). More importantly, this overview information is coherently provided through the newly established Secondary Metabolite Bioinformatics Portal (SMBP) along with links to references and websites of the tools and databases. We also discuss perspectives on further development of the field.

### 2.1. Manual genome mining

Before automated tools (see below) became available, genome mining approaches have been undertaken by “manually” identifying key biosynthetic enzymes in genome data. For this, either amino acid sequences of characterized proteins of interest were used as queries for BLAST or PSI-BLAST,\(^{29}\) or – if alignments of a family of query sequences were available – these were used to generate profile Hidden Markov Models (HMMs) which served as queries using the software HMMer.\(^{30}\) Gene clusters were then identified by analyzing the genes encoded up- and downstream of the hit sequence. While this approach has been superseded by automatic tools for most of the commonly observed gene cluster types, it is still highly relevant for identifying gene clusters which are not covered by the rulesets of the common tools and where prototypes have just been discovered and described. The manual genome mining can be further improved with tools like MultiGeneBlast,\(^{31}\) which allow a BLAST-based analyses of whole operons or gene clusters.

### 2.2. Tools for identification of BGCs

Identifying BGCs with BLAST and HMMer works very well with low false positive rates for many different classes of secondary metabolites, for example polyketides (PKs) synthesized by type I or type II PKS, ribosomally and post-translationally modified peptides (RiPPs), or NRPs. Therefore, a number of tools have been developed that use rule-based approaches, i.e., the specific search for distinct enzymes or enzymatic domains (Fig. 1).

**BAGEL**\(^{26-30}\) is a web-based comprehensive mining suite to identify and characterize RiPPs in microbial genomes. **BAGEL** provides an annotation-independent identification of the genes encoding precursor peptides, classification of the RiPP types as well as a database of known RiPPs. Especially, in the field of identification of the BGCs of type I PKS, NRPS and hybrid PKS/NRPS, a wide variety of tools exist. **ClustScan**\(^{35}\) is a Java-based desktop application that offers mining for PKS and NRPS gene clusters in a convenient graphical user interface. **ClustScan** was used to compile and analyze the data contained in the ClustScan database (see below). NP_searcher\(^{17}\) is a web-based software program with an emphasis on structure prediction of the putative peptide or polyketide metabolites. NaPDos\(^{16}\) uses BLAST and HMMer to identify ketosynthase domain (in PKS) and condensation domain (in NRPS) encoding genes in genomic and metagenomic datasets and provides a detailed phylogenetic analysis of these domains which are then classified into functional categories. GNP/Genome search\(^{35,69,78}\) and GNP/PRISM\(^{18}\) are web-based tools to mine for and analyze PKS and NRPS pathways,
including identification of similar known pathways, the latter with an emphasis on the prediction of putative products. They are closely interconnected with the metabolomics platform iSNAP, which uses information on predicted products to identify corresponding peaks in liquid chromatography/tandem mass spectrometry (LC-MS/MS) data (see paragraph 2.6). The Secondary Metabolite Unknown Region Finder SMURF can detect fungal PKS, NRPS and terpenoid gene clusters involving a dimethylallyltryptophan synthase type prenyltransferases. With pipelines such as CLUster SEquence ANalyzer (CLUSEAN), there are also tools available that can automate the analysis of larger datasets using scripts instead of interactive web pages.

While the tools mentioned above are specialized in detecting and analyzing specific classes of secondary metabolites, antiSMASH provides detection rules for 44 different classes and subclasses of secondary metabolites. In addition to the identification of gene clusters, antiSMASH also provides detailed annotation of the domain structures of modular PKS and NRPS, analysis of lanthipeptide pathways, substrate predictions, genome-scale metabolic modeling and comparative genomics tools to identify conserved subclusters biosynthesizing building-blocks, similar gene clusters in other sequenced genomes and the Minimum Information about a Biosynthetic Gene cluster (MIBiG) dataset. With this functionality, antiSMASH currently is the most comprehensive software for mining microbial genomes for BGCs. In the future, it is planned to extend antiSMASH as a generic platform integrating various tools such as CRISPy-web, a web-based tool to design guide RNAs (sgRNAs) for CRISPR applications (Blin et al. in this issue).

All rule-based BGC-mining approaches can precisely identify BGCs of known biosynthetic types, but fail to identify pathways, which use non-homologous enzymes or enzymes with biochemistry that is presently unknown. However, there are some alternative approaches that try to identify BGCs independent of pre-defined rulesets. The software ClusterFinder, which also is implemented as an alternative cluster detection algorithm in antiSMASH, uses a HMM-based approach to detect chromosomal regions in genomes that aggregate protein domains associated with secondary metabolite biosynthetic pathways. The EvoMining approach identifies gene clusters based on the observation that many BGCs encode isoenzymes closely related to primary metabolism, but displaying a different phylogeny. By scanning the genomes for the occurrence of such enzymes, it is possible to detect secondary metabolite BGCs without respect to their conserved enzymology.

2.3. Tools for analyzing specific enzymes

In addition to the general genome mining tools mentioned above, a whole set of tools was developed specifically to provide automated specificity prediction for NRPS A-domains and to detect the enzymatic domains in multi-modular PKS and NRPS, such as SEARCHPKS or NRPS-PKS/SBSPKS. One of the hallmarks of computational analysis of secondary metabolite biosynthetic pathways was the deciphering of the NRPS A-domain specificity conferring code by Stachelhaus et al. and Challis et al., who found out that conserved amino acids near the active site of NRPS A-domains can be used to map the substrate specificity of these enzymes, which is an important prerequisite for the computational
Table 1
Comprehensive collection of freely accessible software programs and databases dedicated to natural product research. Only software programs and databases properly functioning as of December 2015 are listed in this table. A more comprehensive list can be found at the SMBP (http://www.secondarymetabolites.org).

| Software program or database | URL | Reference | Last publication or documented update | Main content and/or function |
|-----------------------------|-----|-----------|---------------------------------------|-------------------------------|
| **Tools for mining of secondary metabolite gene clusters** | | | | |
| 2metDB | http://secmetdb.sourceforge.net/ | 13–15 | 2015 | Standalone (Mac) tool to mine PKS/NRPS gene clusters |
| antiSMASH | http://antismash.secondarymetabolites.org | 27 | 2013 | Web application and standalone tool (LINUX, MacOS and MS Windows) to mine and analyze BGCs; includes comparative genomics tools and a homology-based metabolic modeling pipeline |
| BAGEL | http://bagel2.molgenrug.nl/ | 28–30 | 2013 | Web application to mine and analyze RiPPs |
| CLUSEAN | https://bitbucket.org/tilmweber/clusean | 31 | 2013 | Standalone (LINUX and MacOS) tool to mine and analyze BGCs, mainly PKS/NRPS |
| ClusterFinder | https://github.com/petercim/ClusterFinder | 20 | 2014 | Standalone tool (LINUX and MacOS) to identify BGCs with an non-rule based approach |
| eSNaPD | http://esnapd2.rockefeller.edu/ | 32–34 | 2014 | Web application to mine metagenomic datasets for BGCs |
| EvoMining | http://148.247.230.39/newevomining/new/evo-mining_web/index.html | 21 | 2015 | Web application for phylogenomic approach of cluster identification |
| GNP/Genome Search | http://magarveylab.ca/gnp/#/genome | 35 | 2015 | Web application to mine and analyze BGCs, mainly PKS/NRPS, including glycosylations and structure prediction |
| GNP/PRISM | http://magarveylab.ca/prism | 18 | 2015 | Web application to use transcriptome data to identify BGC coordinates in fungal genomes |
| MIDDAS-M | http://133.242.13.217/MIDDAS-M/ | 36 | 2013 | Web application to identify BGC coordinates in fungal genomes without transcriptome data |
| MIPS-CG | http://www.fung-metb.net/ | 37–8 | 2015 | Web application offering phylogenomic analysis of PKS-KS and NRPS-C domains |
| NaPDos | http://napdos.ucsd.edu/ | 16 | 2012 | Web application to mine PKS/NRPS/terpenoid gene clusters in fungal genome |
| SMURF | http://jcvi.org/smurf/index.php | 19 | 2010 | Java-based standalone tool to mine for PKS/NRPS BGCs |
| **Software for the analysis of type I PKS and NRPS pathways** | | | | |
| ClustScan Professional | http://bioserv.pbf.hr/cms/index.php?page=clustscan | 39 | 2008 | Web application/standalone tool (LINUX) to mine for PKS/NRPS BGCs |
| NPsearcher | http://dna.sherman.lsi.umich.edu/ | 17 | 2009 | Web application to mine for PKS BGCs |
| NRPS-PKS/SBSPKS | http://www.nri.ac.in/~pksdb/sbspks/master.html | 40–4 | 2010 | Web application to mine for PKS BGCs |
| SEARCHPKS | http://linux1.nii.ac.in/~pksdb/DBASE/pagesearchpks.html | 42 | 2003 | Web application to mine for PKS BGCs |
| **Software for predicting substrate specificities** | | | | |
| LSI-based A-domain function predictor | | | | |
| NRPS/pKS substrate predictor | | | | |
| NRPSpredictor/NRPSpredictor2 | | | | |
| NRPSp | http://www.nrpspp.com/ | 47 | 2012 | Web application to predict A-domain specificities |
| PKS/NRPS Web Server/Predictive Blast Server | | | | |
| SEARCHChTr | http://linux1.nri.ac.in/~pankaj/gt/gt_DB/html_files/searchgt.html | 37 | 2009 | Web application to predict A-domain specificities |
| SEQ-NRPS | http://services.birc.au.dk/seql-nrps/ | 48 | 2005 | Web application to predict glycosyltransferase specificities |
| | | | | | Web application to predict A-domain specificities |

(continued on next page)
| Software program or database | URL | Reference | Last publication or documented update | Main content and/or function |
|-----------------------------|-----|-----------|--------------------------------------|------------------------------|
| Databases focusing on gene clusters | | | | |
| Bactibase | http://bactibase.pfba-lab-tun.org | 50,51 | 2011 | Web accessible database of bacteriocins |
| ClusterMine360 | http://www.clustermine360.ca/ | 52 | 2013 | Web accessible database of BGCs |
| DoBISCUT | http://www.bio.nite.go.jp/pks/ | 53 | 2013 | Web accessible database of PKS/NRPS BGCs |
| IMG-ABC | http://img.jgi.doe.gov/abc | 54 | 2015 | Web accessible database of PKS/NRPS BGCs |
| MIBIG | http://mibig.secondarymetabolites.org | 55 | 2015 | Web accessible database of BGCs, tightly integrated into JGI's IMG platform |
| Recombinant ClustScan Database | http://csdb.bioserv.pbf.hr/csdb/RCSDB.html | 56 | 2015 | Web accessible repository of BGCs |
| Antibioticome | http://magarveylab.ca/antibioticome | Unpublished | 2015 | Database of in silico recombinated BGCs |
| ChEMBL | https://www.ebi.ac.uk/chembl/ | 58 | 2015 | Web accessible database on bioactive compounds with drug-like properties |
| ChemSpider | http://www.chemspider.com/ | 59 | 2015 | Web accessible database on bioactive compounds with drug-like properties |
| KNApSAcK database | http://kanaya.aist-nara.ac.jp/KNApSAcK/ | 60 | 2015 | Web accessible database on compounds; standalone version of KNApSAcK metabolite database available |
| NORINE | http://bioinfo.lifl.fr/norine | 61,62 | 2015 | Web accessible database on NRPs |
| Novel Antibiotics Database | http://www.antibiotics.or.jp/journal/database/database-top.htm | Unpublished | 2008 | Web accessible database on compounds |
| PubChem | http://pubchem.ncbi.nlm.nih.gov/ | 63 | 2015 | Web accessible database on compounds and bioactivities; source data available for download |
| StreptomeDB | http://www.pharmaceutical-bioinformatics.de/streptomedb | 64 | 2015 | Web accessible database on compounds produced by streptomycetes; download of compounds and metadata in SD format |
| Metabolomics tools | | | | |
| Cycloquest | http://cyclo.ucsd.edu | 65 | 2011 | Web application to correlate tandem MS data of cyclopeptides with gene clusters |
| GNPS | http://gnps.ucsd.edu/ | unpublished | 2015 | Generic metabolomics portal to analyze MS/MS data (dereplication and molecular networking) |
| GNPS/iSNAP | http://magarveylab.ca/gnp/ | 66,67 | 2015 | Web application to automatically identify metabolites in MS/MS data based on genomic data |
| NRPquest | http://cyclo.ucsd.edu | 70 | 2014 | Web application to correlate NRP tandem data with gene clusters |
| Pep2Path | http://pep2path.sourceforge.net | 71 | 2014 | Standalone application to correlate peptide sequence tags with NRP and RIPP BGCs |
| RiPPquest | http://cyclo.ucsd.edu | 72 | 2014 | Web application to correlate RiPP tandem data with gene clusters |
prediction of the biosynthetic products. The PKS/NRPS Web Server, Predictive Blast Server, and 2metDB deliver predictions based on BLAST analyses against the signatures determined by Challis et al. Later tools introduced the use of profile HMMs, for example an algorithm by Minowa et al., NRPSmap, NRPS/PKS substrate predictor, machine learning-based on transductive Support Vector Machines (SVMs), as for example implemented in NRPSpredictor. Latent Semantic Indexing, which is used by the LSI-based A-domain predictor or the Sequence Learner algorithm, which is used in SEQL-NRPS. There have also been first successful reports on using structural bioinformatics involving both crystal structure or homology models and docking analyses with putative substrates, which contributed to predicting substrate specificities of A-domains. However, this approach is currently very compute-intensive, and no automated tools have been reported so far. For other enzymes involved in secondary metabolism biosynthesis, only few tools are available. PKSIIfinder uses transductive SVMs to classify type III PKSs. SEARCHGTr currently is the only tool that offers prediction of glycosyltransferase specificities.

2.4. Databases focusing on biosynthesis genes and their clusters

All the tools mentioned in the previous section can be used to identify or analyze secondary metabolite BGCs or specific enzymes of the pathways in the user-submitted gene cluster/genome data. To allow cross-species comparison, several databases have been developed focusing on different aspects of secondary metabolism. The ClustScan database, DoSISCUI, and ClusterMine360 provide collections of a limited set of mostly hand-curated PKS and NRPS gene clusters. The recombinant ClustScan database r-CSDB in addition contains more than 20,000 in silico recombined sequences that are expected to produce novel molecules. Recently, a standard on MIBig has been developed. In the course of this project, a MIBig repository was generated, containing more than 1000 characterized BGCs; more than 400 of them were manually annotated and curated by the original researchers carrying out the experimental characterizations. In addition to these databases, data collections were also established based on large-scale sequencing efforts. The Integrated Microbial Genomes: Atlas of Biosynthetic Gene Clusters (IMG-ABC) is a huge data collection based on manually curated BGCs, but also includes automatically mined BGCs of public genome data and genomes that were sequenced at the US Department of Energy Joint Genome Institute (JGI). Currently, IMG-ABC is the largest collection of BGCs data.

So far, the genome data used for genome mining of whole biosynthetic pathways almost exclusively originated from cultivable organisms. Considering the fact that only a little percentage of environmental bacteria can be grown in culture, the unculturable microorganisms remain a huge and currently under-exploited resource. The environmental Surveyor of Natural Product Diversity (eSNAPD) is a system to map amplicon datasets to known BGCs. As eSNAPD also uses location metadata, the data can be analyzed based on not only the database but also on location information about the sampling sites.

2.5. Databases focusing on compounds

In addition to general public molecule databases, such as PubChem, ChEMBL, and ChEBI which contain information on a humongous volume of chemical compounds including secondary metabolites, commercial natural product compound databases are available, including antiBASE (Wiley-VCH, Weinheim, Germany), and the Dictionary of Natural Products (Taylor and Francis Group LLC, USA). Recently, several freely accessible or openly licensed databases have also been developed. The KNapsack website offers information on various secondary metabolites with respect to their basic chemical properties and bioactivities. Although the KNapsack system is mostly focused on plant metabolites, it also contains information on microbial bioactive compounds. A component of the KNapsack system dealing with metabolites can also be downloaded and used as a standalone Java-based tool. StreptomeDB is a database focusing on secondary metabolites isolated from streptomycetes. Bactibase is focused on ribosomally synthesized antimicrobial peptides, while NORINE is a hand-curated database of NRPs and their activities.
2.6. Metabolomics tools for natural product identification

LC-MS and nuclear magnetic resonance (NMR)-based metabolomics approaches gain increasing importance in natural product studies [for reviews, see references 90,91]. While some of the tools or databases on natural product compounds and their BGCs already have histories of more than ten years, first computational approaches have been published only very recently that use chemoinformatic approaches to automatically classify and map metabolomics (i.e., MS and MS/MS data) to natural product families and corresponding biosynthetic pathways. This has been especially successful for identifying peptides (RiPPs and NRPs) in the mass spectra of complex samples. Software programs for these approaches include Pep2Path,92 RiPPquest,93 NRPPquest,72 and Cycloquest.68 The GNPS/I5NAP (From Genes to Natural Products) – web application provides a user-friendly interface to carry out analyses of MS/MS data of NRP producing strains.95,96,97 Signals corresponding to NRPs or NRP-analogs are detected by comparison to databases containing computationally generated fragments of known secondary metabolites (e.g., those extracted from NOR2NE51 or PubChem63). Recently, I5NAP has also been extended to identify PK compounds and analogs of known molecules.98

The Global Natural Products Social Molecular Networking system (GNPS) provides workflows for automated spectra deconvolution, molecular networking to identify compound families and dereplication against a database of known molecules (unpublished). In addition to the analysis function, GNPS has a social network component that allows users to share their mass spectrometry datasets (including continuous identification by re-analyzing the deposited datasets against updated spectra libraries) or datasets of reference compounds.

2.7. High-throughput metabolic modeling tools

The availability of genomic information allows generation of genome-scale metabolic models, which have now become one of standard tools in systems biology and metabolic engineering communities. This technology enables linking between genotype, including BGCs of secondary metabolites, and metabolic phenotype of secondary metabolite producing microorganisms. A genome-scale metabolic model is a type of mathematical model that is based on mass balances of all the metabolites known/predicted to be present in an organism of interest and is represented in a large-scale stoichiometric matrix that can be simulated with various numerical optimization tools.99 One of the unique features of genome-scale metabolic model is description of gene-protein-reaction (GPR) associations in a Boolean format; the GPR associations logically connect genomic information with the organism’s metabolism, and hence enable prediction of various metabolic phenotypes using gene-level information. In the field of secondary metabolites, genome-scale metabolic models have largely contributed to studies on (i) predicting intracellular flux distributions of actinomycetes under specific environmental/genetic conditions90,96 and (ii) gene manipulation targets for overproduction of target secondary metabolites.96,98

Although development of a genome-scale metabolic model is a laborious and time-consuming procedure, involving a total of 96 steps in a protocol,97 a large fraction of the procedure can now be automated. Such high-throughput metabolic modeling tools allow streamlined system-wide metabolic studies for newly sequenced genomes of actinomycetes and other secondary metabolite producers whose number keeps growing due to increased attentions on novel antibiotics discovery. Among currently available high-throughput metabolic modeling tools, to our knowledge, only ModelSEED has been deployed to reconstruct multiple actinomycete species in a high-throughput manner for large-scale metabolic studies.99 Currently available high-throughput modeling tools are summarized in Table 2. For a detailed comparison of high-throughput metabolic modeling tools, see Hamilton and Reed.100 and Dias et al.108 Finally, a challenge for modeling secondary metabolite producers is that all the available metabolic modeling tools do not consider secondary metabolite biosynthesizing reactions and their relevant precursors, and the fact that most secondary metabolites are biosynthesized in stationary phase and not in the exponential growth phase, which stands against the pseudo-steady state assumption of this modeling approach. These special circumstances will therefore require additional efforts in optimizing the metabolic models.

3. The secondary metabolite bioinformatics portal

The field of secondary metabolite bioinformatics is drastically changing with new tools being released and old services discontinued. We therefore started the web-portal SMBP as a one-stop access point containing a manually curated collection of all the relevant tools and databases for ‘omics-based secondary metabolism research, including short descriptions of the tools, literature references and links to the web sites and/or download pages (Fig. 2). Currently, the tools and databases are assigned to one (or more) categories of contents functionalities covering secondary metabolite compounds, genome mining, PKS/NRPS analysis, specificity predictors, metabolomics analysis, metabolic modeling and generic tools. A full text search engine provides easy access to the relevant information. The SMBP is openly available at http://www.secondarymetabolites.org, and the MarkDown source code for the portal is available at https://bitbucket.org/secmetbioinl/portal.

4. Future challenges

Despite significant advances on computational approaches to identify and characterize BGCs, there still exist several challenges that have to be addressed in the near future.

Even for the well-studied secondary metabolite classes such as PK or NRP pathways, prediction of the core scaffold structure of a compound is incomplete because the biochemical knowledge on these systems is not yet implemented in the software (relatively easy to fix in this case) or the relevant biochemical knowledge is not sufficiently available to be the basis for the implementation of novel computational algorithms (more difficult to overcome than the former case). In particular, for machine learning-based approaches, the availability of medium- to large-scale biochemical data required to train good models is very limiting in many cases.

Another unsolved problem is currently inaccurate prediction of gene cluster borders. The most widely used genome mining software antiSMASH simply assigns n kb upstream or downstream of the core biosynthetic genes to the cluster (for example, n = 20 kb for PKS and NRPS clusters, and n = 10 kb for lanthipeptides). SMURF, which addresses fungal PK, NRP and terpenoid metabolites, uses a different approach; a statistical analysis of 22 clusters of the model strain Aspergillus fumigatus led to the identification of a total of 27 protein domains, which commonly co-occur with the PK, NRP and terpenoid biosynthetic genes. The occurrence of these domains in genes flanking the core biosynthetic genes, together with the intergenic distance, is then considered to calculate the cluster borders.19 Another promising approach to predict BGC borders is to use comparative genomics data; genes within a putatively identified BGC, which are conserved among other producers of similar compounds, are likely to belong to the BGCs, whereas genes not belonging to the cluster are more divergent. An algorithm implementing this strategy for filamentous fungi (NRRPS-CS) has been described by Takeda et al.17 For fungal BGCs, it has further been demonstrated that – in addition to the mining and analysis methods described above – transcriptome data can provide valuable
For prokaryotes, to our best knowledge, no such observations have been reported so far.

Analyses involving the integration of different “kinds” of data (e.g., genome with transcriptome or metabolome data) generally suffer from a very poor integration of different functionalities available across the tools and the requirement of specific input and output formats; all these barriers make using relevant software programs difficult for researchers not familiar with bioinformatics. In fact, this is a chronic problem in bioinformatics and systems biology in general. Advances in integrating heterogeneous ‘omics data would offer new dereplication opportunities to identify already known metabolites at a very early stage of the metabolite discovery process. In relation to this, proteome data can deliver important information on secondary metabolite biosynthesis when they are correlated to metabolome data (e.g., obtained by LC-MS) and bioactivity profiles. Using a set of different growth conditions, which leads to the differential expression of BGCs and thus different bioactivity profiles, Gubbens et al. were able to correlate the expression levels of biosynthetic enzymes with the occurrence of secondary metabolites. Using this approach, it was possible to identify juglomycin C and the corresponding gene cluster in Streptomyces sp. MBT70. Furthermore, the power of combining large-scale genome and metabolome data was explored along with computational approaches to identify novel secondary metabolites. Doroghazi et al. identified 11,422 PKS-, NRPS-, NRPS-independent siderophores, lanthipeptides and thiazole-oxazole modified microcin gene

Fig. 2. A screenshot of the antiSMASH page in the Secondary Metabolite Bioinformatics Portal at http://www.secondarymetabolites.org.
clusters in 830 genome sequences of actinomycetes. The gene cluster sequences were then clustered based on a combination of different distance metrics, resulting in 4122 gene cluster families. For a subset of 178 analyzed strains, this network was then automatically correlated with high-resolution mass spectrometric data of known compounds leading to the automatic identification of 110 molecules and 27 molecule families. Thus, for some of these molecule families, previously unidentified gene clusters could be automatically related to the produced metabolite. Taken together, as demonstrated in the studies discussed above, it is highly desirable to interconnect the existing tools and data, and automate the analysis workflows for streamlined characterization of genomes and their resulting secondary metabolites. Current bottlenecks in such integrative approaches can be relieved by standardizing APIs and data structures for programmatic access of the different tools.

5. Implications for synthetic biology applications in natural product studies

While the availability of computational tools provides new possibilities for identifying and characterizing novel secondary metabolites, such tools are also essential for the development of synthetic biology strategies, which aim at the efficient production of rationally designed molecules. While there exist several generic synthetic biology tools to predict, prioritize, model, select and implement pathways, as reviewed in reference [11], only few reports exist on their use to engineer natural product biosynthetic pathways.

Especially, engineering PKS and NRPS megasynthases will need further emphasis; from a formal perspective, these modular enzymes are excellent candidates for synthetic biology approaches because they display a modular organization and a well-defined split-up of “enzymatic tasks” and tempt to easy plug-and-play approaches. Although there are many successful module and/or domain replacements reported during the last 15 years that led to rationally designed molecules or combinatorially engineered products, the failure rates are still high and the yields obtained with the engineered assembly lines usually decrease severely. The main reason for this is likely that for designing the modified enzymes, mostly sequence divergence at the linker regions between the enzymatic domains or even trial-and-error approaches might have caused the suboptimal performance of the engineered assembly lines (i.e., inactivity or drastically decreased yields) as they interfered with the 3D structure and the intra- and intermolecular protein–protein interactions within the highly complex megaenzymes. Because structural data of not only separate enzymatic domains but also complete modules for both NRPS [20,121] and type I PKS [22,121] recently became available, they now offer the molecular background to overcome current challenges in engineering the PKS or NRPS assembly lines. In the same line, biochemical studies have been carried out, which specifically address how different domains interact with one another within the PKS or NRPS assembly lines and may help better understanding of the molecular mechanisms within the assembly lines. In this context, it has been learned that this knowledge has yet to be integrated into synthetic biology design software. Certainly, these approaches will be supported by the availability of heterologous expression and genome engineering tools like CRISPR, which recently also became available for secondary metabolite producers. These technologies will drastically reduce the efforts to generate the required recombinant strains and thus allow the high-throughput generation of many variants.

6. Conclusions

Genome mining and other “omics-based approaches to identify and characterize secondary metabolites and their producers have become essential technologies complementing the classical approaches of natural product discovery. This trend is manifested by an increasing number of new and improved bio- and chemoinformatic tools and databases bridging computational biology and wet-lab work in the field. Because of the ever-growing number of computational tools and databases dedicated to secondary metabolites, we herein release the SMBP (http://www.secondarymetabolites.org) where researchers in the field can explore diverse tools and databases in one stop. The SMBP is expected to enable users to compare tools for their utilities and make further contributions to the field of secondary metabolites.

Acknowledgments

The work of the authors is supported by a grant of the Novo Nordisk Foundation, Denmark.

References

1. Taylor J, Hafner M, Verushalmi E, Smith R, Bellasio J, Vardavas R, et al. Estimating the economic costs of antimicrobial resistance. Models and results. California: RAND Corporation SM and Cambridge, UK: The Wellcome Trust; 2014.
2. KPMG LLP UK. Report on the global economic impact of anti-microbial resistance, commissioned by the Wellcome Trust; 2014.
3. The Review on Antimicrobial Resistance. Antimicrobial resistance: tackling a crisis for the health and wealth of nations. London: The Review on Antimicrobial Resistance; 2012. <http://amr-review.org/sites/default/files/SECURING%20NEW%20DRUGS%20FOR%20FUTURE%20GENERATIONS%20FINAL%20WEB_0.pdf>.
4. Bush K, Courvalin P, Dantas G, Davies J, Eisenstein B, Huovinen P, et al. Tackling antibiotic resistance. Nat Rev Microbiol 2011;9:894–6.
5. Crigg GM, Newman DJ. Natural products: a continuing source of new drug leads. Biochim Biophys Acta 1830;2013:3670–95.
6. Fischbach MA, Walsh CT. Antibiotics for emerging pathogens. Science 2009;325:1089–93.
7. Newman DJ, Crigg GM. Natural products as sources of new drugs over the 30 years from 1981 to 2010. J Nat Prod 2012;75:311–35.
8. Bode HJ, Bethe B, Hofs R, Zeek A. Big effects from small changes: possible ways to explore nature's chemical diversity. ChemBioChem 2002;3:619–27.
9. Bentley SD, Chater KF, Cedeno-Tarrega AM, Challis GL, Thomson NR, James KD, et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3/21. Nature 2002;417:141–7.
10. Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, et al. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. Nat Biotechnol 2003;21:526–31.
11. Omura S, Ikeda H, Ishikawa J, Hanamoto A, Takashashi C, Shinose M, et al. Genome sequence of the industrial microorganism Streptomyces avermitilis, deducing the ability of producing secondary metabolites. Proc Natl Acad Sci U S A 2001;98:12215–20.
12. Weber T, Charasunti P, Musiol-Kroll EM, Jiang X, Tong Y, Kim HU, et al. Metabolic engineering of antibiotic factories: new tools for antibiotic production in actinomycetes. Trends Biotechnol 2015;33:15–26.
13. Weber T, Bliin K, Dudrella S, Krug D, Kim HU, Bruccoleri R, et al. antiSMASH 3.0 – a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res 2015;43:W237–43.
14. Bliin K, Medema MH, Kazempour D, Fischbach M, Breitling R, Takano E, et al. antiSMASH 2.0 – a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res 2013;41:W402–14.
15. Medema MH, Bliin K, Cimermancic P, de Jager V, Zakrzewski P, Fischbach MA, et al. antiSMASH: rapid identification, annotation and analysis of secondary metabolism biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res 2011;39:46.
16. Ziemert N, Podels S, Penn K, Badger JH, Allen E, Jensen PR. The natural product domain seeker NaPDoS: a phylogenetic bioinformatic tool to classify secondary metabolite gene diversity. PLoS ONE 2012;7:e34064.
17. Li MH, Ung PM, Zajkowska J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products, BNK Bioinformatics 2009;10:185.
18. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster AL, et al. Genomes to natural products: PreDictions Informatics for Secondary Metabolites (PREDISIM). Nucleic Acids Res (2015);doi:10.1093/nar/gkv1012.
19. Khalti N, Seufuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: genomic mapping of fungal secondary metabolism clusters. Fungal Genet Biol 2010;47:736–41.
20. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 2014;158:412–21.
21. Cruz-Morales P, Martinez-Guerrero CE, Morales-Escalante MA, Yáñez-Guerra LA, Kopp JF, Feldmann J, et al. Recapitulation of the evolution of biosynthetic
gene clusters reveals hidden chemical diversity on bacterial genomes. bioRxiv 2015; doi:10.1101/020503.

22. Andersen MR, Nielsen PF, Nuttallgaard A, Petersen LM, Zachariasen M, Hansen TJ, et al. Accurate prediction of secondary metabolism gene clusters in filamentous fungi. Nat Protoc 2013; 8:1099–107.

23. Bachmann BO, Grandal J, Nielsen PF, Methods for prediction of microbial polyketide and non-ribosomal peptide biosynthetic pathways from DNA sequence data. Methods Enzymol 2009; 458:181–217.

24. van Heel AJ, de Jong A, Mollabal-Lopez M, Kob J, Kuipers OP. BAGE: automated identification of genes encoding bacteriocins and (non- ) bactericidal posttranslationally modified peptides. Nucleic Acids Res 2013; 41:W448-53.

25. de Jong A, van Heel AJ, Kob J, Kuipers OP. BAGE2: mining for bacteriocins in genomic data. Nucleic Acids Res 2010; 38:W654-71.

26. Boddy CN. Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides. J Ind Microbiol Biotechnol 2014; 41:443–50.

27. Bremner MA, Rast J, Chapter 8. Methods for prediction of microbial polyketide and non-ribosomal peptide biosynthetic pathways from DNA sequence data. Methods Enzymol 2009; 458:181–217.

28. van Heel AJ, de Jong A, Mollabal-Lopez M, Kob J, Kuipers OP. BAGE3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Res 2013; 41:W448-53.

29. de Jong A, van Heel AJ, Kob J, Kuipers OP. BAGE2L: mining for bacteriocins in genomic data. Nucleic Acids Res 2010; 38:W654-71.

30. de Jong A, van Hijum SA, Bijlsma JJ, Kob J, Kuipers OP. BAGE: a web-based bacteriocin genome mining tool. Nucleic Acids Res 2006; 34:W273–9.

31. Rast J, Bremner MA, Rast J, Chapter 7. Methods for prediction of microbial polyketide and non-ribosomal peptide biosynthetic pathways from DNA sequence data. Methods Enzymol 2009; 458:181–217.

32. van Heel AJ, de Jong A, Mollabal-Lopez M, Kob J, Kuipers OP. BAGE3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Res 2013; 41:W448-53.

33. de Jong A, van Heel AJ, Kob J, Kuipers OP. BAGE2L: mining for bacteriocins in genomic data. Nucleic Acids Res 2010; 38:W654-71.

34. Boddy CN. Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides. J Ind Microbiol Biotechnol 2014; 41:443–50.

35. Bremner MA, Rast J, Chapter 8. Methods for prediction of microbial polyketide and non-ribosomal peptide biosynthetic pathways from DNA sequence data. Methods Enzymol 2009; 458:181–217.

36. van Heel AJ, de Jong A, Mollabal-Lopez M, Kob J, Kuipers OP. BAGE3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Res 2013; 41:W448-53.

37. de Jong A, van Heel AJ, Kob J, Kuipers OP. BAGE2L: mining for bacteriocins in genomic data. Nucleic Acids Res 2010; 38:W654-71.

38. Boddy CN. Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides. J Ind Microbiol Biotechnol 2014; 41:443–50.

39. Bremner MA, Rast J, Chapter 8. Methods for prediction of microbial polyketide and non-ribosomal peptide biosynthetic pathways from DNA sequence data. Methods Enzymol 2009; 458:181–217.

40. van Heel AJ, de Jong A, Mollabal-Lopez M, Kob J, Kuipers OP. BAGE3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Res 2013; 41:W448-53.

41. de Jong A, van Heel AJ, Kob J, Kuipers OP. BAGE2L: mining for bacteriocins in genomic data. Nucleic Acids Res 2010; 38:W654-71.

42. Boddy CN. Bioinformatics tools for genome mining of polyketide and non-ribosomal peptides. J Ind Microbiol Biotechnol 2014; 41:443–50.

43. Bremner MA, Rast J, Chapter 8. Methods for prediction of microbial polyketide and non-ribosomal peptide biosynthetic pathways from DNA sequence data. Methods Enzymol 2009; 458:181–217.
