Synthetic and Systems Biotechnology 1 (2016) 69–79

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: keaipublishing.com/synbio

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

Tilmann Weber [a,*], Hyun Uk Kim [a,b]

[a] *The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle Alle 6, 2970 Hørsholm, Denmark*
[b] *BioInformatics Research Center, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea*

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 10 December 2015
Accepted 26 December 2015
Available online

Keywords:
Antibiotics
Biosynthesis
Bioinformatics
NRPS
PKS
Natural product

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 at http://www.secondarymetabolites.org) 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.[4] 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,[5] 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.[6] Therefore, it is surprising that still the majority of newly approved small-molecule drugs are natural products or their derivatives.[7]

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,[8] and then isolate and characterize the compounds with sophisticated analytical

Abbreviations: A, adenylation domain; BGC, biosynthetic gene cluster; C, condensation domain; GPR, gene-protein-reaction; HMM, hidden Markov model; LC, liquid chromatography; MS, mass spectrometry; NMR, nuclear magnetic resonance; NRP, non-ribosomally synthesized peptide; NRPS, non-ribosomal peptide synthetase; PCP, peptidyl carrier protein; PK, polyketide; PKS, polyketide synthase; RiPP, ribosomally and post-translationally modified peptide; SVM, support vector machine.

[*] Corresponding author. The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kogle Alle 6, 2970 Hørsholm, Denmark. Tel.: +45 24 89 61 32; fax: +45 45 25 80 01.

E-mail address: tiwe@biosustain.dtu.dk (T. Weber).

Peer review under responsibility of KeAi Communications Co., Ltd.

http://dx.doi.org/10.1016/j.synbio.2015.12.002

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)[9] and the avermectin producer *Streptomyces avermitilis*[10,11] 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 (NRPSs). 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 antibiotics and Secondary Metabolite Analysis SHell (antiSMASH),[13–15] 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[23–26]]. 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,[75] 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.[76] 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,[77] 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[28–30] 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[39] 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,