Computational and Structural Biotechnology Journal 13 (2015) 192–203

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Mini Review

Genome mining: Prediction of lipopeptides and polyketides from *Bacillus* and related Firmicutes

Gajender Aleti, Angela Sessitsch, Günter Brader *

AIT Austrian Institute of Technology GmbH, AIT, Health & Environment Department, Bioresources Unit, Konrad Lorenz Strasse 24, A-3430 Tulln, Austria

ARTICLE INFO

Article history:
Received 1 November 2014
Received in revised form 15 March 2015
Accepted 16 March 2015
Available online 24 March 2015

Keywords:
Polyketides
Lipopeptides
Non-ribosomal protein synthetase
Genome mining
Structure prediction
Paenibacillus

ABSTRACT

Bacillus and related genera in the Bacillales within the Firmicutes harbor a variety of secondary metabolite gene clusters encoding polyketide synthases and non-ribosomal peptide synthetases responsible for remarkable diverse number of polyketides (PKs) and lipopeptides (LPs). These compounds may be utilized for medical and agricultural applications. Here, we summarize the knowledge on structural diversity and underlying gene clusters of LPs and PKs in the Bacillales. Moreover, we evaluate by using published prediction tools the potential metabolic capacity of these bacteria to produce type I PKs or LPs. The huge sequence repository of bacterial genomes and metagenomes provides the basis for such genome-mining to reveal the potential for novel structurally diverse secondary metabolites. The otherwise cumbersome task to isolate often unstable PKs and deduce their structure can be streamlined. Using web based prediction tools, we identified here several novel clusters of PKs and LPs from genomes deposited in the database. Our analysis suggests that a substantial fraction of predicted LPs and type I PKs are uncharacterized, and their functions remain to be studied. Known and predicted LPs and PKs occurred in the majority of the plant associated genera, predominantly in *Bacillus* and *Paenibacillus*. Surprisingly, many genera from other environments contain no or few of such compounds indicating the role of these secondary metabolites in plant-associated niches.

Contents

* Corresponding author.
E-mail address: guenter.brader@ait.ac.at (G. Brader).

http://dx.doi.org/10.1016/j.csbj.2015.03.003

1. Introduction

Bacteria are known to produce structurally diverse secondary metabolites including aminoglycosides, polyketides (PKs) and several small proteinaceous and peptidal structures such as bacteriocins, oligopeptides and lipopeptides (LPs) [1–3]. A substantial number of these metabolites have been described for their bactericidal, immune suppression and tumor suppression properties and represent potentially valuable agents in medical and veterinary medical applications, but especially PKs and LPs play also essential roles for applications in agriculture. They are vital for bacterial activities in suppressing disease pressure in plants by antimicrobial activities and activating plant defense and are important for biofilm formation and root colonization of crop plants [4–8]. LPs and PKs encompass a variety of cyclic, linear and branched structures and are generated by complex enzymes known as non-ribosomal peptide synthetases (NRPS) and polyketide synthases (PKS), respectively [9,10]. NRPS and type I PKS share to a large extent similar modular architecture and are largely organized into modules containing multiple domains, allowing the repetitive incorporation of building blocks into larger resulting compounds [11]. However, for the biosynthesis of smaller compounds (e.g. some siderophores), non-modular NRPS have been reported [12]. Often NRPS and type I PKS enzymes work using a co-linearity code, so that the recruitment of amino acids (for NRPS) and carboxylic acids (for PKS) for the biosynthesis and final structure assembly is the same as the order of catalytic domains in the genome [13,14]. This feature and insight into the architecture of modules and domains of NRPS and PKS often facilitate prediction of compound structures based on genomic sequences [15,16]. Nevertheless, variations from this conventional organization have been described and include for instance module iteration and skipping in several biosynthetic processes [17].

In this review, we will focus on Bacillales, an order belonging to the phylum Firmicutes, as genera within this order represent a rich source for diverse secondary metabolite gene clusters. Based on a recent whole genome mining study, 31% of the Firmicutes are estimated to harbor NRPS and PKS secondary metabolite gene clusters. 70% of these encode NRPS and 30% hybrid NRPS/PKS or PKS [18]. The total percentage of Firmicutes producing secondary metabolites is certainly higher, also because genes responsible for many common secondary metabolite classes (e.g. many oligosaccharides) are not detected by widely used prediction tools such as antiSMASH[19,20]. The distribution of NRPS and PKS gene clusters within different orders of the Firmicutes is not uniform and *Bacillus* and *Paenibacillus* from the order Bacillales dominate this secondary metabolite gene clusters count. These two genera in particular are well noted for their capability to produce structurally diverse LPs and PKs [4,7], but the genome information from most other Bacillales members remains largely untapped.

Despite the fact that next generation sequencing technology has contributed to the ample availability of the whole genome sequence data and a number of analysis tools for metabolite prediction exist [19–23], yet little is accomplished to explore the sequence wealth to identify novel LPs and PKs in these genomes and to predict uncharacterized secondary metabolites. We briefly review current knowledge on well characterized LPs and PKs from the Bacillales and show which novel compounds can be anticipated based on published Bacillales genome data using genome mining study and secondary metabolite prediction tools. The questions addressed here are to review the structural and functional information and the underlying gene clusters of known type I PKs and LPs produced by Bacillales and to elucidate by genome mining potential products of uncharacterized gene clusters and the potential of producing yet unidentified secondary metabolites of these types in distinct taxonomic groups of the Bacillales.

1.1. Bacillus and Paenibacillus polyketides

Polyketides are generated from simpler building units by repeated decarboxylation and condensation cycles on PKS enzymes [24]. The PKS machinery comprises three core domains: the acyl transferase (AT), the acyl carrier protein (ACP) and the ketosynthase (KS). The AT domain is responsible for activation and transfer of a simpler building unit (malonyl coenzyme A) to the ACP domain. The KS domain catalyzes decarboxylation and condensation reaction between the two ACP linked malonates [25]. Other domains include ketoreductases (KR) which catalyze hydroxy group formation, dehydratases (DH) which form double bonds after water elimination, enoyl reductases (ER) which catalyzes reduction reaction of the double bonds and methyl transferases (MT) which introduce methyl groups and branching in the carbon backbone. A phosphopantetheinyl transferase (PPT) encoded by a *sfp* gene is essential for the activation of the ACP domains [26,27]. The arrangement and the order of the catalytic domains within PKS influence PKs biosynthesis leading to a remarkable diversity in the PKs production. The PKS enzymes can be broadly categorized into three types, depending on the architecture of catalytic domains [28]. Type I PKS enzymes contain modules organized in multiple catalytic domains within a single protein that carry out decarboxylation and condensation steps to generate PKs from the starter unit malonyl-CoA [11]. In the type II and type III PKS enzymes, catalytic domains are found in separate proteins [28]. A large group of bacterial PKs are produced by modular PKS I enzymes with iterative KS, ACP and modification domains. These type I PKS mostly lack AT domains within the clusters, malonyl-CoA is transfered by acyl transferases acting in trans [29]. A large number of PKS is often found in association with NRPS as hybrid enzymes type I PKS-NRPS [30].

Metabolites produced by *Bacillus amyloliquefaciens* and *Bacillus subtilis* represent a substantial part of the diversity of LPs and PKs from the genus *Bacillus* [31,32]. The majority of the plant growth promoting and biocontrol agents commercially available are produced by these two species [4]. They produce three types of polyene PKs comprising bacillaene, difficidin and macrolactin [26,32]. *B. amyloliquefaciens* FZB42 contains a genome size of 3918 kb, of which nearly 200 kb are devoted to the production of polyketides. These three giant PKs gene clusters were assigned unambiguously by a mutagenesis study, utilizing MALDI-TOF MS and LC-ESI MS techniques [26]. In the genus *Paenibacillus* two PKs have been described so far. The underlying genetic cluster remains to be unambiguously identified in the case of paenimacrolidin [33], while for the recently described paenilamicins from *Paenibacillus larvae* also the responsible gene clusters have been reported [34]. In the following we describe the five known types of PKS from *Bacillus* and *Paenibacillus* in more detail.

1.1.1. Bacillaene

Bacillaene was first reported in the culture medium of *B. subtilis* strains 3610, and 55422 [35,36]. It has a linear structure comprising a conjugated hexaene (Fig. 2A) [35,36]. The biosynthesis of bacillaene has been described in *B. amyloliquefaciens* FZB42 and is encoded by a hybrid type I PKS-NRPS gene cluster called *bae* [26] (Fig. 1A). This cluster shares architectural characteristics with *pksX* of *B. subtilis* strain 168, presumably also encoding bacillaene [26]. The *bae* gene cluster contains five long open reading frames (ORFs) including *baeJ*, *baeL*, *baeM*, *baeN* and *baeR* [26]. The first and the second adenylation domains of *baeJ* are responsible for the incorporation of α-hydroxy-isocaproic acid and glycine, respectively. The third adenylation domain of *baeN* is involved in the incorporation of alanine [37]. Modules 4, 8 and 14 are splitted between adjacent genes (Fig. 1A). Three short ORFs found upstream of *baeJ* are *baeC, baeD, baeE*, encode for the three discrete AT domains that load malonyl-CoA [37]. Bacillaene and dihydrobacillaene are structural variants represented in this group of PKs [27,36] (Fig. 2A). Cell viable assays revealed that bacillaene selectively inhibits protein biosynthesis in prokaryotes, but not in eukaryotes,