ChemPubSoc Europe

DOI: 10.1002/cbic.201402434

CHEMBIOCHEM Communications

Automated Identification of Depsipeptide Natural Products by an Informatic Search Algorithm

Michael A. Skinnider, Chad W. Johnston, Rostyslav Zvanych, and Nathan A. Magarvey*[a]

Nonribosomal depsipeptides are a class of potent microbial natural products, which include several clinically approved pharmaceutical agents. Genome sequencing has revealed a large number of uninvestigated natural-product biosynthetic gene clusters. However, while novel informatic search methods to access these gene clusters have been developed to identify peptide natural products, depsipeptide detection has proven challenging. Herein, we present an improved version of our informatic search algorithm for natural products (iSNAP), which facilitates the detection of known and genetically predicted depsipeptides in complex microbial culture extracts. We validated this technology by identifying several depsipeptides from novel producers, and located a large number of novel depsipeptide gene clusters for future study. This approach highlights the value of chemoinformatic search methods for the discovery of genetically encoded metabolites by targeting specific areas of chemical space.

Microbial natural products are an important source of therapeutics,[1] due in part to their diverse chemical scaffolds.[2] Nonribosomal depsipeptides are one of the larger families of these bioactive secondary metabolites.[3] These small molecules are biosynthesized by massive assembly line-like enzymes known as nonribosomal peptide synthetases (NRPSs),[4] and are defined by the presence of a variable number of both ester and amide bonds. Many depsipeptide NRPSs contain integrated adenylation-ketoreductase (A-KR) domains that lead to the activation, reduction, and incorporation of α-ketoacids, thereby facilitating ester bond formation within the peptide backbone.[5,6] Depsipeptides are notable for their potent bioactivities as anticancer, antimicrobial, and antiviral agents, including several compounds currently in use in the clinic.[6] In spite of intense industrial and academic programs for isolation of these peptides over the last half century, preliminary genomic sequence analyses have revealed that as few as 10% of genetically encoded natural products have been identified,[7] which indicates that there are likely many undiscovered depsipeptides. As such, novel methods to access the biosynthetic potential of microbial genomes could unlock a large resource of unknown chemistries and, possibly, therapeutic small molecules.[8]

[a] M. A. Skinnider,[+] C. W. Johnston,[+] R. Zvanych, Prof. N. A. Magarvey
Department of Chemistry and Chemical Biology
Department of Biochemistry and Biomedical Sciences
M. G. DeGroote Institute for Infectious Disease Research
McMaster University, Hamilton, Ontario, L8N 3Z5 (Canada)
E-mail: magarv@mcmaster.ca

[+] These authors contributed equally to this work.

We previously developed an informatic search algorithm for natural products (iSNAP),[9] and demonstrated that it facilitated the identification of known nonribosomal peptides in liquid chromatography with tandem mass spectrometry (LC–MS/MS) data of microbial extracts. The iSNAP method is inspired by database-dependent proteomic mass spectrometry, capitalizing on reliable amide bond fragmentation in peptides by generating in silico libraries of peptide natural product fragments that can be correlated to real MS/MS fragments observed in the LC–MS/MS data. By building on this experimentally validated method for identifying known peptide natural products, we have recently introduced several improvements to facilitate the identification of genetically predicted natural products (C.W.J. et al., unpublished results). Despite considerable success in locating known and novel peptidic natural products, we observed that the reliance of the algorithm on amide bond cleavage impaired the discovery of depsipeptide natural products. To confirm this hypothesis and correct the problem, we developed a novel bioinformatic method to identify integrated A-KR domains and their substrates within microbial genome sequences, with the goal of identifying depsipeptide biosynthetic gene clusters, prioritizing strains predicted to produce novel depsipeptide metabolites, and detecting their small-molecule products in LC–MS/MS chromatograms.

Adenylation domain substrates have classically been determined by the identification of ten key amino acid residues, which were proposed to constitute a specificity-conferring code on the basis of the crystal structure of the gramicidin S synthetase phenylalanine adenylation domain.[10] While this approach has been useful in deciphering the substrate specificity of nonribosomal peptides, integrated A-KR didomains contain an atypical active site arrangement, which precludes the use of established codes. More recently, the use of ten letter codes has been superseded by the development of novel methods, including machine-learning techniques and the development of substrate-specific hidden Markov models,[11] which are more amenable to the identification of depsipeptide α-ketoacid substrates. In order to predict A-KR α-ketoacid specificity we collected a series of experimentally validated A-KR didomain sequences, which were used to construct a library of substrate-specific profile hidden Markov models. The resulting hidden Markov models and reference sequences, specific to pyruvate, α-ketoisovalerate, α-ketoisocaproate, and 3-methyl-2-oxopentanoate, are available at http://magarveylab.com/depsipeptide/. Using these bioinformatic tools, we carried out a virtual screening campaign for integrated A-KR domain-containing NRPSs within an in-house library of genome sequences from environmental actinomycete isolates. We were able to detect a putative depsipeptide biosynthetic gene cluster within the

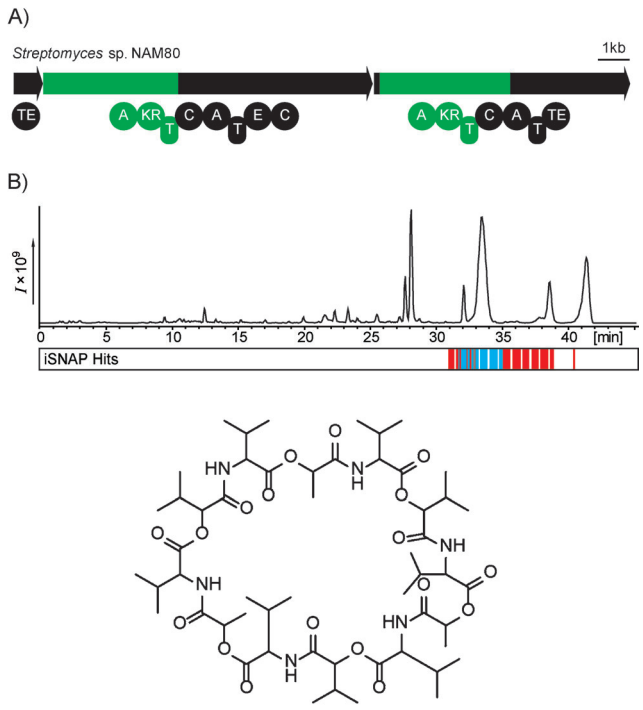

Figure 1. Identification of an integrated A-KR didomain NRPS and informatic dereplication of valinomycin. A) Hidden Markov models of A-KR didomains identified an A-KR NRPS in *Streptomyces* sp. NAM80, predicting a structure consistent with valinomycin (bottom). B) Implementation of in silico ester cleavage enables informatic detection of valinomycin and montanastatin (blue), along with predicted structural analogues (red) from a crude extract of NAM80.

genome of the environmental actinomycete NAM80 (Figure 1 A), which was later determined to be a strain of *Streptomyces fulvissimus*. Analysis with hidden Markov models identified two A-KR domains with putative substrates of pyruvate and α-isoketovalerate. This, in addition to predicted adenylation domain substrates, indicated significant homology to the biosynthetic machinery for the well-known ionophore valinomycin.[12] Analysis of liquid-culture extracts confirmed the presence of valinomycin, but identification proved impossible through the use of our informatic search algorithm alone. We hypothesized that the high ester bond content of valinomycin prevented the generation of appropriate hypothetical MS–MS fragments, which canonically only include amide cleavage. To facilitate the informatic detection of valinomycin, as well as the identification of genetically predicted depsipeptides, in silico identification and fragmentation of ester bonds was implemented within the iSNAP algorithm. Additional fragmentation settings were developed to allow the user the option to cleave bonds between the sp^3 oxygen and sp^2 carbon of an ester moiety, as well as the opposite bond to the sp^3 oxygen to account for alternative fragmentation patterns (designated "inverse ester" fragmentation). The addition of ester-cleavage functionality to the algorithm enabled the chemoinformatic detection of valinomycin—the trimer of the four-module assembly-line product—as well as montanastatin, the dimer (Figure 1 B). Applying techniques developed for the identification of genetically encoded peptide natural products (C.W.J. et al., unpublished results), we generated a library of hypothetical valinomycin and montanastatin structural variants that can be envisioned by classical assembly line promiscuity. Although symmetry of these molecules complicates the analysis, we were able to informatically detect variants of valinomycin and montanastatin with one (14 Da) or two (28 Da) additional CH_2 units (Figure 1 B). One false-positive scan (detecting destruxin A2) was observed from a total of 2904 scans, indicating that the implementation of ester cleavage did not considerably increase the false-positive discovery rate.

To better appreciate the utility of the ester cleavage functionality, we sought to evaluate the ability of iSNAP to dereplicate a diverse series of depsipeptides with varying numbers of ester bonds and ester-to-amide ratios. LC–MS/MS data of pure standards (daptomycin, syringomycin, aureobasidin, surfactin, and beauvericin) and depsipeptides found in bacterial extracts (fusaricidin, the antimycin and neoantimycin families, and JBIR-06) were analyzed with our iSNAP method, both with and without ester cleavage, in order to assess its impact on informatic depsipeptide detection. We concluded that, for depsipeptide molecules with a single ester bond, ester cleavage either did not impair or marginally improved dereplication (Figure 2). Although the number of matched fragments nearly always increased with additional bond-cleavage parameters, the scoring algorithm of iSNAP primarily considers the ratio of matched fragments to generated fragments. Thus, only sufficiently productive cleavage parameters will improve scoring. In contrast to molecules with single ester bonds, dereplication of depsipeptides with multiple ester bonds (including beauvericin and the antimycin-type molecules) was impossible in the absence of in silico ester fragmentation (Figure 2), demonstrating

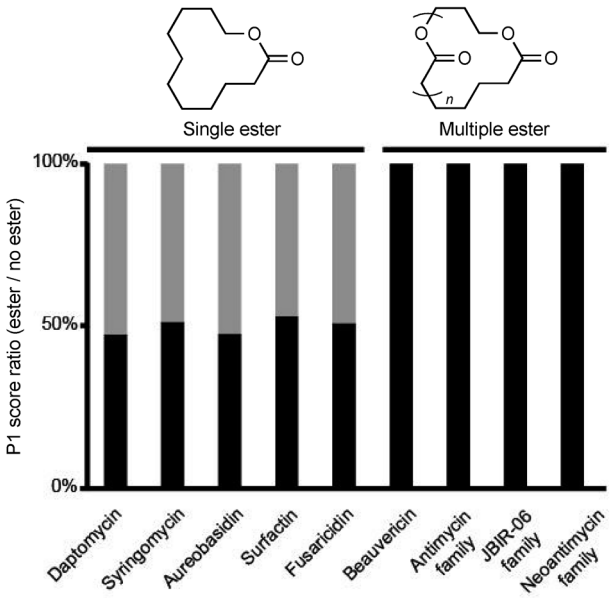

Figure 2. In silico ester cleavage is necessary for the informatic detection of multi-ester depsipeptides. Averaged P1 scores—representing the statistical confidence that matched MS/MS fragments correspond to a given structure—were compared between depsipeptides dereplicated with (black) or without (gray) in silico ester cleavage.

the utility of this new parameter in enabling informatic access to previously undetectable molecules.

We returned to our in-house-sequenced actinomycete library and attempted to locate non-A-KR depsipeptide biosynthetic gene clusters to provide a comprehensive demonstration of the utility of this novel fragmentation parameter. BLAST analysis revealed a biosynthetic gene cluster for a quinoxaline-type antitumor agent in an industrial *Streptomyces silvensis* strain, detected by the presence of key genes for quinoxaline biosynthesis.[13] Equipped with our comprehensive library of known depsipeptides, we analyzed crude culture extracts from *S. silvensis* to informatically identify quinoxaline-type metabolites. Informatic analysis in the absence of ester cleavage was incapable of revealing putative quinoxalines. However, the use of ester-cleavage functionality revealed the known quinoxaline echinomycin (quinomycin), with an HPLC retention time of 20.5 min (Figure 3). By generating theoretical fragments from

Figure 3. Identification of a quinoxaline NRPS and informatic dereplication of echinomycin. A) Genomic identification of a putative quinoxaline NRPS in *S. silvensis*. B) Implementation of in silico ester cleavage, in addition to thioether and amide cleavage, enables informatic detection of echinomycin (blue, bottom) from a crude extract of *S. silvensis*.

amide, thioether, and importantly ester bonds, we were able to identify this complicated, low-abundance natural product from a novel producer, demonstrating the utility of this flexible chemoinformatic platform for the discovery of desired metabolites.

The most commonly cited study of the gap between the biosynthetic potential of organisms and the number of natural products isolated and identified showed that, even in well-studied *Actinomycetes*, as few as 10–20% of genetically encoded natural products have been isolated.[7] Though compelling, this estimate only investigates a well-studied and productive genus, and does not focus on highly desired natural products such as depsipeptides or polyketides. To estimate the number of novel depsipeptide scaffolds present in the genomes of bacteria that have been sequenced, we applied our library of hidden Markov models to identify integrated A-KR domains within all available bacterial sequences.[14] Manual annotation of the known products revealed that while well-studied microbes such as Bacilli, *Streptomyces*, and Cyanobacteria had few novel depsipeptide scaffolds, a large number of putatively novel A-KR-containing depsipeptides could be found in Clostridia and poorly-studied Actinobacteria, as well as a diverse array of Gram-negative α-, β-, γ-, and δ-Proteobacteria. Novel depsipeptide gene clusters were found in organisms such as *Herpetosiphon* and *Mycobacteria*, which are poorly studied but have acknowledged biosynthetic potential,[15–16] as well as unstudied organisms such as *Glaciecola* and *Ruminococcus*, demonstrating that future microbial sequencing will continue to reveal new molecular scaffolds (Figure 4). We estimate that 36% of A-KR domain-containing depsipeptide assembly lines in sequenced genomes have been associated with known natural products, and thus that a multitude of novel depsipeptide chemical scaffolds await discovery.

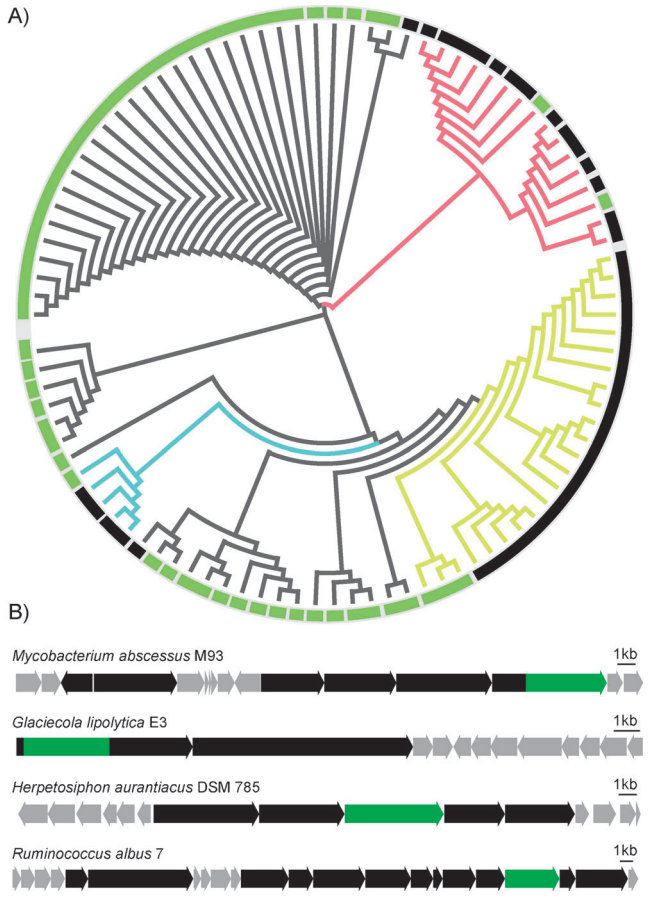

Figure 4. Hidden Markov models reveal an abundance of novel A-KR NRPSs in sequenced bacterial genomes. A) A-KR detecting HMMs reveal a diverse array of known (black) and novel (green) A-KR NRPS sequences in the genome sequences of *Streptomyces* (pink), Bacillus (light green), Cyanobacteria (blue), and other bacteria (dark gray). B) Examples of novel A-KR NRPS gene clusters from unstudied organisms, including biosynthetic genes (black) and highlighted A-KR didomains (green).

Depsipeptides are natural products with unique chemical scaffolds whose bioactivities have been optimized through evolution. Clinically validated therapeutics including daptomycin and romidepsin have demonstrated the utility of these molecules in treating disease, but genomic analysis suggests that many depsipeptide natural products remain undiscovered. In this work, we have introduced a bio- and chemoinformatic method for the genomic discovery and mass-spectral identification of depsipeptides. Development of a library of profile hidden Markov models enables the genomic identification of depsipeptide biosynthetic gene clusters and their specified substrates. An improved version of our informatic search algorithm has demonstrated the utility of in silico ester cleavage in the dereplication of a diverse range of depsipeptides within LC–MS/MS data from standards and crude extracts. More generally, our approach highlights the value of informatic strategies in the targeted exploration of natural-product chemical space.

Experimental Section

General experimental procedures: LC–MS data was collected using a Bruker AmazonX ion-trap mass spectrometer coupled with a Dionex UltiMate 3000 HPLC system, equipped with a Luna C_{18} column (50×4.6 mm or 150×4.6 mm; Phenomenex). The mobile phases were acetonitrile and water, each with 0.1% formic acid. For analytical flow rates, a UV/MS flow splitter of 10:1 was used. LC–MS spectral analysis was performed using Compass DataAnalysis 4.1 (Bruker). Valinomycin and echinomycin extract LC–MS/MS mzXML files are available at http://magarveylab.com/depsipeptide/

Bacterial strains and culture conditions: The environmental actinomycete NAM80 was isolated from a soil sample outside of McMaster University. *Streptomyces silvensis* was obtained from the American Type Culture Collection (ATCC 53525). Both strains were cultivated on GYM agar plates at 30 °C. For production of valinomycins, NAM80 was cultured in Bennett's medium (beef extract (1 g L^{-1}), yeast extract (1 g L^{-1}), NZ-amine (2 g L^{-1}), glucose (10 g L^{-1}), pH 7.3). For production of echinomycin, *S. silvensis* was cultured in VL55 medium (ATCC medium 2734).

Analysis of depsipeptide standards: Pure standards of daptomycin (Sigma), beauvericin (Sigma), syringomycin (Santa Cruz Biotech.), surfactin (Santa Cruz Biotech.), and aureobasidin (Clontech) were dissolved in methanol to a final concentration of 100 μg mL^{-1}. Fusaricidin A was identified after 96 h in a potato dextrose broth culture supernatant of *Paenibacillus polymyxa* (ATCC No. 21830), which had been extracted with HP20 resin (20 g L^{-1}; Dialon) for 2 h and eluted with excess methanol. Antimycin, JBIR-06, and neoantimycin were also identified from crude culture extracts which were previously described.[6] Samples were analyzed by iSNAP using standard settings, either with or without an optional single ester cleavage.

Genome sequencing: A single colony each of NAM80 and *S. silvensis* were used to inoculate two 50 mL cultures of GYM medium containing 0.5% glycine (GGYM). Cultures were grown for 96 h at 30° and 250 rpm shaking. An aliquot of each culture (500 μL) was centrifuged at 12*g* for 5 min, resuspended in SET buffer (500 μL; NaCl (75 mM), EDTA (25 mM, pH 8.0), Tris·HCl (20 mM, pH 7.5), lysozyme (2 mg mL^{-1})) and incubated for 2 h at 37 °C to induce cell lysis. Proteinase K and SDS were added after lysis to final concentrations of 0.5 mg mL^{-1} and 1%, respectively. The lysis mixtures were incubated at 55 °C for 2 h before the concentration of NaCl was adjusted to 1.25 M and the mixture was extracted twice with phenol/chloroform. Isopropanol was added (equivalent to 60% of the volume of the solution) to precipitate the genomic DNA, which was subsequently washed twice with 70% ethanol and resuspended in sterile water for sequencing. Genomic DNA was sent for library preparation and Illumina sequencing at the Farncombe Metagenomics Facility at McMaster University by using an Illumina HiSeq DNA sequencer. Contigs were assembled using the ABySS genome assembly program[17] and with Geneious bioinformatic software (Biomatters, Ltd).

Extraction and detection of valinomycins: NAM80 colonies from GYM agar plates were inoculated into GGYM cultures (50 mL) in sterile Erlenmeyer flasks (250 mL) and grown for 72 h at 250 rpm and 28 °C. This culture was used to inoculate a flask of Bennett's medium (50 mL), which was also grown at 28 °C and 250 rpm for 96 h. The supernatant of this culture was mixed with HP20 resin (20 g L^{-1}; Dialon) for 2 h and eluted with excess methanol. This supernatant extract was evaporated to dryness, reconstituted in 1 mL of methanol, and analyzed by LC–MS. Separation was achieved using a Luna C_{18} column (150×4.6 mm), with a mobile phase of aqueous acetonitrile (5% for 4 min, ramping to 100% by 30 min). The mobile phase flow rate was maintained at a constant 1.5 mL min^{-1}.

Extraction and detection of echinomycin: *S. silvensis* colonies from GYM agar plates were inoculated into GGYM cultures (50 mL) in sterile Erlenmeyer flasks (250 mL) and grown for 72 h at 250 rpm and 28 °C. This culture was used to inoculate a flask of VL55 medium (50 mL, ATCC medium 2734), which was also grown at 28 °C and 250 rpm for 96 h. The supernatant of this culture was mixed with HP20 resin (20 g L^{-1}; Dialon) for 2 h and eluted with excess methanol. This supernatant extract was evaporated to dryness, reconstituted in 1 mL of methanol, and analyzed by LC–MS. Separation was achieved using a Luna C_{18} column (150×4.6 mm), with a mobile phase of aqueous acetonitrile (5% for 4 min, ramping to 100% by 30 min). The mobile phase flow rate was maintained at a constant 1.2 mL min^{-1}.

Bioinformatic identification of nonribosomal depsipeptide synthetases: Sequences were collected from an in-house database of nonribosomal peptide synthetase gene clusters. Multiple sequence alignments in Stockholm format were created using Clustal Omega.[18] Hidden Markov models (HMMs) were then created from the Stockholm alignment files by using the hmmbuild program, which is part of the HMMER 3.1 package.[14] Analysis of environmental actinomycete genomes for depsipeptide biosynthetic gene clusters was performed using the hmmsearch program included in the same package.

In silico identification and fragmentation of natural product esters: Chemoinformatic analysis of depsipeptide molecules began with the identification of all ester bonds within the molecule. Informatic analysis of the chemical structures was performed using abstractions developed by the CDK.[19] Ester bonds were located informatically by searching the molecular structure for oxygen atoms with two bonds, one of which was a single bond to a carbon itself double-bonded to a second oxygen atom. The bond between the oxygen and the sp^2 carbon of the carbonyl group was then fragmented. The mass of the fragment containing the terminal oxygen was increased by one to simulate the adduction of a proton to form a hydroxy group. Additionally, in order to facilitate the dereplication of depsipeptides with unusual MS/MS cleavage patterns,

an "inverse ester" fragmentation option was also implemented. Identification of "inverse ester" bonds proceeded in the same manner as for ester bonds, but the non-ester bond to the sp^3 oxygen was fragmented, corresponding to ester cleavage with a simultaneous loss of water.

Detection of known and novel A-KR-containing depsipeptide biosynthetic machinery: Our A-KR didomain HMM was loaded into HMMer and used to identify A-KR didomains within a non-redundant protein database. A-KR didomain sequence entries were apparently exhausted in results with *E* values greater than 5.0×10^{-191}. Accordingly, over 400 sequence results were manually curated to remove non-A-KR results, duplicated sequences, and initiating A-KR didomains from natural product machinery for nostophycin,[20] aeruginoside,[21] auriporcine,[22] and microsclerodermin.[23] The remaining 97 individual entries—including several identical biosynthetic gene clusters—were cross-referenced against known depsipeptide biosynthetic gene clusters, identifying 23 novel A-KR containing gene clusters and 13 known A-KR gene clusters. 16S rRNA sequences for each A-KR bearing organism were used to generate a phylogenetic tree using the Geneious software, with a Tamura-Nei genetic distance model and Neighbor-joining as the tree building method.[24–25] The corresponding product was exported as a Newick tree, loaded into Dendroscope,[26] and exported as a PDF for annotation with Adobe Illustrator CS6.

Acknowledgements

This work was supported by generous gifts from McMaster University and the Canadian Institute of Health Research. We are grateful for the help of Dr. Xiang Li on this work.

Keywords: bioinformatics · chemoinformatics · depsipeptides · mass spectrometry · natural products

[1] D. J. Newman, G. M. Cragg, *J. Nat. Prod.* **2007**, *70*, 461.
[2] J. Clardy, C. Walsh, *Nature* **2004**, *432*, 829.
[3] F. Sarabia, S. Chammaa, A. Sanchez Ruiz, L. Martin Ortiz, F. J. Lopez Herrera, *Curr. Med. Chem.* **2004**, *11*, 1309.
[4] M. A. Fischbach, C. T. Walsh, *Chem. Rev.* **2006**, *106*, 3468.
[5] N. A. Magarvey, M. Ehling-Schultz, C. T. Walsh, *J. Am. Chem. Soc.* **2006**, *128*, 10698.
[6] S. A. Vanner, X. Li, R. Zvanych, J. Torchia, J. Sang, D. W. Andrews, N. A. Magarvey, *Mol. Biosyst.* **2013**, *9*, 2712.
[7] M. Nett, H. Ikeda, B. S. Moore, *Nat. Prod. Rep.* **2009**, *26*, 1362.
[8] C. T. Walsh, M. A. Fischbach, *J. Am. Chem. Soc.* **2010**, *132*, 2469.
[9] A. Ibrahim, L. Yang, C. Johnston, X. Liu, B. Ma, N. A. Magarvey, *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 19196.
[10] T. Stachelhaus, H. D. Mootz, M. A. Marahiel, *Chem. Biol.* **1999**, *6*, 493.
[11] B. I. Khayatt, L. Overmars, R. J. Siezen, C. Francke, *PLoS One* **2013**, *8*, e62136.
[12] A. M. Matter, S. B. Hoot, P. D. Anderson, S. S. Neves, Y. Q. Cheng, *PloS one* **2009**, *4*, e7194.
[13] K. Watanabe, K. Hotta, A. P. Praseuth, K. Koketsu, A. Migita, C. N. Boddy, C. C. Wang, H. Oguri, H. Oikawa, *Nat. Chem. Biol.* **2006**, *2*, 423.
[14] R. D. Finn, J. Clements, S. R. Eddy, *Nucleic Acids Res.* **2011**, *39*, W29.
[15] M. Nett, O. Erol, S. Kehraus, M. Köck, A. Krick, A. Eguereva, E. Neu, G. M. König, *Angew. Chem. Int. Ed.* **2006**, *45*, 3863; *Angew. Chem.* **2006**, *118*, 3947.
[16] J. R. Doroghazi, W. W. Metcalf, *BMC Genomics* **2013**, *14*, 611.
[17] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, I. Birol, *Genome Res.* **2009**, *19*, 1117.
[18] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, D. G. Higgins, *Mol. Syst. Biol.* **2011**, *7*, 539.
[19] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. L. Willighagen, *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 493.
[20] D. P. Fewer, J. Osterholm, L. Rouhiainen, J. Jokela, M. Wahlsten, K. Sivonen, *Appl. Environ. Microbiol.* **2011**, *77*, 8034.
[21] K. Ishida, G. Christiansen, W. Y. Yoshida, R. Kurmayer, M. Welker, N. Valls, J. Bonjoch, C. Hertweck, T. Börner, T. Hemscheidt, E. Dittmann, *Chem. Biol.* **2007**, *14*, 565.
[22] C. M. Theodore, B. W. Stamps, J. B. King, L. S. Price, D. R. Powell, B. S. Stevenson, R. H. Cichewicz, *PLoS One* **2014**, *9*, e90124.
[23] T. Hoffmann, S. Müller, S. Nadmid, R. Garcia, R. Müller, *J. Am. Chem. Soc.* **2013**, *135*, 16904.
[24] K. Tamura, M. Nei, *Mol. Biol. Evol.* **1993**, *10*, 512.
[25] N. Saitou, M. Nei, *Mol. Biol. Evol.* **1987**, *4*, 406.
[26] D. H. Huson, C. Scornavacca, *Syst. Biol.* **2012**, *61*, 1061.

Received: July 31, 2014
Published online on December 8, 2014