Insights into an Unusual Nonribosomal Peptide Synthetase Biosynthesis

IDENTIFICATION AND CHARACTERIZATION OF THE GE81112 BIOSYNTHETIC GENE CLUSTER

Received for publication, May 20, 2010, and in revised form, August 5, 2010 Published, JBC Papers in Press, August 14, 2010 DOI 10.1074/jbc.M110.146803

Tina M. Binz, Sonia I. Maffioli, Margherita Sosio, Stefano Donadio, and Rolf Müller

From the Department of Microbial Natural Products, Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Helmholtz Center for Infection Research (HZI), and Department of Pharmaceutical Biotechnology, Saarland University, Campus C2 3, Saarbrücken 66123, Germany and Naicons Scrl, Via G. Fantoli, 16/15, Milan 20138, Italy

The GE81112 tetrapeptides (1–3) represent a structurally unique class of antibiotics, acting as specific inhibitors of prokaryotic protein synthesis. Here we report the cloning and sequencing of the GE81112 biosynthetic gene cluster from Streptomyces sp. L-49973 and the development of a genetic manipulation system for Streptomyces sp. L-49973. The biosynthetic gene cluster for the tetrapeptide antibiotic GE81112 (getA-N) was identified within a 61.7-kb region comprising 29 open reading frames (open reading frames), 14 of which were assigned to the biosynthetic gene cluster. Sequence analysis revealed the GE81112 cluster to consist of six nonribosomal peptide synthetase (NRPS) genes encoding incomplete di-domain NRPS modules and a single free standing NRPS domain as well as genes encoding other biosynthetic and modifying proteins. The involvement of the cloned gene cluster in GE81112 biosynthesis was confirmed by inactivating the NRPS gene getE resulting in a GE81112 production abolished mutant. In addition, we characterized the NRPS A-domains from the pathway by expression in Escherichia coli and in vitro enzymatic assays. The previously unknown stereochemistry of most chiral centers in GE81112 was established from a combined chemical and biosynthetic approach. Taken together, these findings have allowed us to propose a rational model for GE81112 biosynthesis. The results further open the door to developing new derivatives of these promising antibiotic compounds by genetic engineering.

The emergence of multi-drug resistant microbial pathogens is driving the search for novel antibiotics with new mechanisms of action and natural products continue to provide original scaffolds affecting essential bacterial targets (1, 2). As it is currently understood, most antibiotics act in three basic ways; 1) inhibition of DNA replication and repair, 2) inhibition of cell wall biosynthesis, and 3) inhibition of protein biosynthesis. Protein translation and cell wall biosynthesis in bacteria are currently the targets for the majority of antimicrobial natural products. The former is the target of the macrolides (e.g. erythromycin), the aminoglycosides (e.g. kanamycin), the streptogramins, the lincomamides, tetracycline, and chloramphenicol as well as of other classes of compounds that are not in clinical use. These compounds have been shown to affect protein biosynthesis at various steps (3, 4).

Some aspects of protein translation are rarely targeted, e.g. translation initiation and polypeptide chain termination. A microbial product screening program aimed at discovering novel inhibitors of bacterial protein synthesis revealed the new tetrapeptide GE81112 compounds (1-3) (Fig. 1) to selectively inhibit the formation of the prokaryotic 30 S initiation complex with an IC<sub>50</sub> of 0.9 μM (5). To date, three GE81112 congeners, A (1), B (2), and B1 (3), have been described from a Streptomyces sp. (Fig. 1). Extensive NMR and MS studies revealed the tetrapeptides to comprise hydroxypipecolic and hydroxypentanoic acids, an (amino)histidine, and a hydroxychlorohistidine (5). A retro-biosynthetic analysis of the GE81112 core structure suggests a NRPS origin with additional tailoring steps occurring at some point during assembly. NRPS multienzymes are composed of successive catalytic units or “domains” that are themselves organized into biosynthetic modules which catalyze the assembly reactions in a coordinated, often “co-linear” manner. Normally, each module in the assembly line performs one cycle of chain extension (condensation of one residue into the growing peptide chain). A typical, minimal NRPS module consists of an adenylation (A) domain, a peptidyl carrier protein (PCP) domain (also referred to as a thiolation (T) domain), and a condensation (C) domain (6). However, a growing number of gene clusters encode systems that deviate in their domain organization from the standard C-A-PCP architecture and comprise partial modules or isolated domains acting in trans to complement the functionality of the multimodular NRPSs (7–9).

To better understand GE81112 biosynthesis and to generate structural analogues, we sought to develop a strategy for the cloning and identification of the biosynthetic gene cluster. To this end, the cluster was identified on two overlapping cosmids. In addition, we expressed five A-domains from the corresponding NRPS genes assigning of the configuration of most chiral centers have

* The work was funded in part by the Bundesministerium für Bildung und Forschung and the Deutsche Forschungsgemeinschaft (to R. M.).
2 The abbreviations used are: NRPS, nonribosomal peptide synthetase; A, adenylation; PCP, peptidyl carrier protein; T, thiolation; C, condensation; O-HPA, 5-hydroxy-2-aminopentanoic acid; Pip, piperolic acid.

The online version of this article (available at http://www.jbc.org) contains supplemental Table S1 and Figs. S1–S4.
allowed us to delineate this unusual NRPS pathway and propose a biosynthetic model for the GE81112 antibiotics.

**EXPERIMENTAL PROCEDURES**

**Bacterial Strains and Culture Conditions—Streptomyces sp.** L-49973 was grown in INAS (glycerol, 30 g liter⁻¹; soya extract, 15 g liter⁻¹; NaCl, 2 g liter⁻¹) or T6 (glycerol, 45 g liter⁻¹; soya extract, 25 g liter⁻¹; CaCO₃, 2 g liter⁻¹) media in baffled flasks for production of GE81112. Pre-cultures were grown in V6 medium (glucose, 20 g liter⁻¹; meat extract, 5 g liter⁻¹; yeast extract, 5 g liter⁻¹; peptone 5 g liter⁻¹; casein, 3 g liter⁻¹; NaCl, 1.5 g liter⁻¹; pH 7.5) and used to inoculate production cultures (1:100). The cultures were maintained at 30 °C and 180 rpm on a rotary incubator and harvested after 6 days.

**Escherichia coli** DH10B, **E. coli** ET12567/pUZ8002, and **E. coli** SURE were grown in liquid LB medium at 37 or 30 °C with the appropriate antibiotic selection. Antibiotic concentrations were as follows; apramycin (60 µg ml⁻¹), chloramphenicol (34 µg ml⁻¹), kanamycin (60 µg ml⁻¹), and ampicillin (100 µg ml⁻¹) were used for selection in **E. coli**. Apramycin (60 µg ml⁻¹) was used for selection of Streptomyces sp. L-49973 recombinants. Nalidixic acid (25 µg ml⁻¹) was used to select against **E. coli** donor after conjugation.

**Molecular Biology Methods**—The pET28b (+) (Novagen) and pCR2.1 TOPO (Invitrogen) cloning vectors were from commercial sources; pKC1132 and pOJ436 were described previously (10). Restriction enzymes were purchased from MBI Fermentas. All PCRs were carried out using Taq (MBI Fermentas) or Phusion (Invitrogen) polymerase. DMSO was added to the reaction mixture to a final concentration of 5%. Conditions for amplification with a Peqlab ThermoCycler were as follows: denaturation, 15–30 s at 95/98 °C; annealing, 8–20 s at 50–62 °C; extension, 15–60 s at 72 °C (30 cycles); final extension at 72 °C for 10 min. Oligonucleotides were obtained from Sigma. Plasmid DNA was isolated using the GeneJET™ Plasmid Miniprep kit (Fermentas). DNA fragments from agarose gel were isolated and purified using the NucleoSpin Extract II Miniprep kit (Fermentas). DNA was ligated into the pCR2.1 TOPO vector and sequenced. Several NRPS arrays (Hybond N+; Amersham Biosciences) were utilized for the screening of the cosmid clones, as described previously (13).

**Screening of a Streptomyces sp. L-49973 Genomic Cosmid Library—Cyclodeaminase and NRPS fragment sequences were initially used as probes under low stringency conditions. To create the cyclodeaminase probe, we used specific primers to amplify two known cyclodeaminase genes, tubZ (14) (primers TubZ_up and TubZ_down, 570-bp product) from the tubulin-sin cluster, and rapL (15), from the rapamycin cluster (primers RapL_up and RapL_down, 534-bp PCR product) (see the primers in supplemental Table S1). As an alternative approach, we designed an additional probe to identify NRPS genes. For this, we used degenerate NRPS primers; NRPS-A1-up and NRPS-H1-dn (supplemental Table S1). These primers amplify A-domains between the structural regions A3 and A6 (16), where the amino acid binding pocket is located, giving a 760-bp PCR fragment. The amplicons were labeled with digoxigenin and used to probe the cosmid library at low stringency (40 °C). Several cosmids that hybridized with both probes were subjected to PCR analysis for the amplification of the cyclodeaminase and NRPS A-domains. In the first screening we identified a cosmid encoding a putative cyclodeaminase but not the expected GE81112 biosynthetic enzymes. A segment of this cyclodeaminase was amplified using the primers Cyclo_probe_for and Cyclo_probe_rev (supplemental Table S1) and used under high stringency conditions (42 °C) for further screening of the library. Cosmids hybridizing with this probe were analyzed by PCR using primers to amplify the cyclodeaminase and NRPS A-domains again. The resulting PCR products were gel-purified and subcloned into pCR2.1 TOPO vector and sequenced. Several NRPS sequences were found from different cosmids, and the eight critical residues responsible for substrate recognition could be determined enabling an in silico prediction of the substrate specificity of each cloned A-domain fragment. Cosmids harboring A-domains that were predicted to activate piperidolic acid were digested with BamHI, and the restriction pattern was compared with identify similar cosmids. End sequencing of the cosmids was carried out using primers T4 and T7 (supplemental Table S1). One cosmid (BI11) was identified containing a large part of the GE81112 biosynthetic gene cluster. To find a cosmid that overlapped with the 3’ (T7) end of the cluster, a 1-kb fragment was amplified from the T7 end of cosmid BI11 (primers BI11-T7end-for and BI11-T7end-rev, supplemental Table S1) to serve as a probe. The cosmid library was then screened with the new probe, with hybridization at 42 °C. Among the identified cosmids, BA23 showed the smallest extent of overlap with BI11 based on PCR analysis and restriction digest. Cosmids BI11 and BA23 were shotgun-sequenced on both strands as described previously (17).

**Data Analysis**—The annotation analysis of the sequence data was performed through FramePlot analysis (FramePlot 4.Obeta) (23) and data base comparison with the basic alignment search tool (BLAST) on the server of the National Center for Biotechnology Information. For alignment analysis of the sequence data, ClustalW on the server of EMBL-EBI was used. Specificity of the A-domains was determined by using the NRPS predictor Bioinformatics Toolbox from University of Tübingen and polyketide synthase/NRPS analysis web site.
Conjugation and Generation of Mutant Strains of Streptomyces sp. L-49973—DNA manipulation was carried out with E. coli DH10B as the host strain. To generate knock-out mutants by means of insert-directed homologous recombination, a 572-bp internal fragment of gene getE was amplified with primers pipA_for and pipA_rev (supplemental Table S1). The fragments were cloned into pCR2.1TOPO, and the constructs were digested with EcoRI. The fragments were ligated into knock-out vector pKC1132 (10) digested with EcoRI. The final pKC1132-derived plasmids were introduced into Streptomyces sp. L-49973 by intergeneric conjugation with the methylation-deficient donor strain E. coli ET12567 containing the conjugative vector pUZ8002 (10). Mutants were analyzed by PCR using appropriate control primers. One primer was designed to bind to the integrated vector backbone (lacZ1, lacZ2), whereas the second primer was designed to target the genome sequence either up- or downstream of the integration site (A1_for, A1_rev) (supplemental Table S1). PCR of the mutants yielded distinct amplicons, whereas no products were detected from the wild type.

*Analysis of GE81112 Production in Streptomyces sp. L-49973*—Streptomyces strains (wild types and mutants) were cultured in 500-ml baffled shake flasks containing 100 ml of GE8112 production medium (INaS5 or T6) at 30 °C and 180 rpm. Recombinant strains were amended with apramycin (60 μg ml⁻¹). A square of agar from a sporulating SM agar plate was used for inoculation. After 6 days of cultivation, cells were harvested by centrifugation at 12,000 rpm for 5 min. The culture supernatant was extracted twice with ethyl acetate, evaporated, and redissolved in 500 μl of methanol. LC-coupled FT-Orbitrap-MS analysis was carried out with an Accella UPLC system (Thermo Electron Corp.) operating in positive ionization mode at a scan range of m/z 100–2000. A Hypersil Gold column (2.1 × 50 mm; Thermo Fisher Scientific) was used for separation with a solvent system consisting of H₂O (A) and acetonitrile (B), each containing 0.1% formic acid. A gradient of 5–95% B was applied over 10 min. Measurements were carried out in single ion mode. GE81112 compounds were identified by comparison to the retention times and the MS data of authentic single ion mode. GE8112 compounds were identified by comparison to the retention times and the MS data of authentic single ion mode. GE8112 compounds were identified by comparison to the retention times and the MS data of authentic single ion mode. GE8112 compounds were identified by comparison to the retention times and the MS data of authentic single ion mode.

*Construction of A-domain Overexpression Constructs*—The genes encoding for the five A-domains (GetEA1, GetGA2, GetGA3, GetMA5) of GE8112 were PCR-amplified from cosmids, BI11 and BA23. The forward and reverse primers were designed to target the genome sequence corresponding to NdeI/BamHI restriction sites (supplemental Table S1). A-domain fragment getA4 could not be amplified from the cosmids, probably due to the high GC content. Therefore, the A-domain sequence was synthesized, and the GC content was optimized for use in *E. coli* (ATG:biosynthetics GmbH). The fragment was obtained in pBluescript SK+ (pBSK) vector flanked by restriction sites NdeI and EcoRI (supplemental Fig. S1). All PCR products were cloned into the digested pET28b (+) vector using the corresponding NdeI/BamHI restriction sites. Fragment getA4 was obtained after restriction of pBSK/GetJA4 with NdeI and EcoRI and cloned into pET28b (+). Final plasmids were sequenced and transformed into *E. coli* Rosetta BL21 (DE3) pLysS/RARE for protein expression.

**Expression and Purification of the A-domains**—Purified A-domain pET28b (+) plasmids were transformed into *E. coli* Rosetta BL21 (DE3) pLysS/RARE competent cells for protein production and purification. Fresh transformants harboring the constructs were grown in LB-medium (1-liter batches started with 0.1% inocula from a 10-ml culture grown for 5 h at 37 °C) supplemented with kanamycin (50 μg ml⁻¹) and chloramphenicol (34 μg ml⁻¹). All cells were grown at 37 °C to an OD₆₀₀ of ~0.8. The cells were then induced with 1 mM isopropyl-β-D-thiogalactopyranoside to an end concentration of 0.2 mM and then grown at 16 °C overnight. The cells were harvested by centrifugation (6000 rpm, 10 min, 4 °C) and resuspended in buffer A (20 mM Tris-HCl, pH 7.8, 200 mM NaCl, and 10% (v/v) glycerol). The cells were then lysed (2 passes at 700 p.s.i., French press, SLM Amino, and the cell debris was removed by centrifugation (21,000 rpm, 10 min, 4 °C). Prepacked HisTrap™ HP columns were used for preparative purification of histidine-tagged recombinant proteins by immobilized metal ion affinity chromatography on the Akta prime™ plus system (GE Healthcare). 15-ml protein lysates were filtered through a sterile filter and loaded onto the 1-ml HisTrap column. Purification was performed as recommended in the GE Healthcare manual (HisTrap HP, Instructions 71-5027-68 AF). The desired protein was eluted from the column in a stepwise imidazole gradient with buffer B (20 mM Tris-HCl, pH 7.8, 200 mM NaCl, and 10% (v/v) glycerol and 60, 100, 200, 300, and 500 mM imidazole) for each protein per liter of culture. Fractions containing the pure target protein, as determined by SDS-PAGE, were combined and concentrated to ~200 μl by using Amicon Ultra PL-10 centricrons. Then 800 μl of storage buffer (50 mM Tris-HCl, pH 7.5, 50 mM NaCl, and 10% (v/v) glycerol) was added to the concentrated protein before flash-freezing in liquid nitrogen and storage at −80 °C. Protein concentrations were determined using the Bradford assay (Bio-Rad). 1–3 mg/ml purified protein were obtained for each protein per liter of culture.

**Determination of Substrate Specificity by ATP-[³²P]PP Exchange Assay**—To determine substrate specificity, ATP-[³²P]PP reactions (100 μl) containing Tris-HCl (pH 7.5, 75 mM), MgCl₂ (10 mM), dATP (5 mM), amino acid (5 mM), and protein (2 μg) were performed at 30 °C. [³²P]-Labeled tetrasodium pyrophosphate was obtained from PerkinElmer Life Sciences (NEN #NEX019). The reactions were started by the addition of [³²P]PP (0.1 μCi final amount) for up to 30 min before quenching with charcoal suspensions (500 μl, 1.6% (w/v) activated charcoal, 0.1 mM Na₃P₄O₁₀, and 0.35 mM perchloric acid in H₂O). The charcoal was pelleted by centrifugation before being washed twice with the wash solution (500 μl, 0.1 mM Na₃P₄O₁₀, and 0.35 mM perchloric acid in H₂O), resuspended in H₂O (500 μl), and counted by liquid scintillation (Beckman LS6500). The experiments were carried out in triplicate for each substrate concentration with a negative control (no amino acid).

**Chemical Analyses**—GE81112 was purified and hydrolyzed as described by Brandi *et al.* (5). Dehalogenation of GE81112 was carried out under H₂ at atmospheric pressure and room temperature in 10% acetic acid, with 10% palladium/carbon as a catalyst.
catalyst. Catalytic hydrogenation of hydroxypicolinic acid was performed under H₂ atmosphere (50 p.s.i.) at room temperature in aqueous NH₄OH with 10% palladium/carbon as a catalyst. Dihydroxylation of allylglycine was performed following published procedures (18). A sample of (2R,3R)-3-hydroxy-pipelicolic acid was kindly provided from Prof. Jieping Zhu (CNRS, Gif-sur-Yvette, France). Chiral GC-MS analyses were performed on the methyl esters, and trifluoroacetyl derivatives were analyzed using a Finnigan TSQ700 triple stage quadrupole mass spectrometer interfaced with a Varian 3400 gas chromatograph (5). NMR and 13C,1H NMR experiments were recorded at 400 MHz on a Bruker Advance spectrometer in D₂O or in D₂O acidified with trifluoroacetic acid (TFA).

RESULTS

Isolation of the GE81112 Biosynthetic Gene Cluster—To capture the GE81112 biosynthetic gene cluster (get, for GE81112 tetrapeptide) a cosmid library containing 2304 clones was generated from the genomic DNA of the GE81112 producer strain (Streptomyces sp. L-49973). Hybridization probes were designed by applying a retrobiosynthetic strategy that allowed us to predict some probable genetic elements of the get cluster from analysis of the metabolite structures (Fig. 1). On this basis we designed a set of probes using degenerate primers based on A-domains of streptomycte origin using the CODEHOP software (20); priming was targeted against the core A3 and A6 motifs. Furthermore, from the structure of 1 we predicted that the starter unit likely involves pipelicolic acid, which is known to be formed from lysine via the action of a lysine cyclodeaminase (21). As cyclodeaminase genes are relatively rare in bacterial genomes, this gene was used to design a second probe to identify the get cluster (22). We designed specific probes based on a cyclodeaminase sequence that was identified in previous experiments in the same strain.³ Colony hybridization of the Streptomyces sp. L-49973 cosmid library with the cyclodeaminase probe led to the identification of 7 cosmid hits. The cosmids were then determined to contain the targeted cyclodeaminase sequence by PCR. As we expected NRPSs to be encoded by the identified cosmids, we used the degenerate NRPS primers to amplify and sequence A-domain segments from these cosmids. Sequence analysis and prediction of the A-domain substrate specificity revealed one cosmid (BI11) containing an A-domain with predicted substrate specificity for proline/pipelicolic acid, as expected for the GE81112 starter unit (Table 1). Subsequent, complete sequencing of the cosmids revealed several genes expected for GE81112 biosynthesis. As the entire gene cluster was not present on the cosmids, we identified overlapping cosmids. In total, seven cosmids were identified, and after verification by PCR and restriction analysis, a single one (BA23) was selected for further analysis.

Sequence Analysis and Organization of the get Biosynthetic Gene Cluster—The two overlapping cosmids, BI11 and BA23, were sequenced. The obtained sequence was analyzed for the presence of putative open reading frames (orfs) with FramePlot 4.0beta (23), and preliminary functional assignments of individual orfs were made by comparison of the deduced gene products with proteins of known function in the BLAST data base (Table 2). Annotation of the two cosmids (BI11 and BA23) revealed 29 orfs, of which 14, designated getA-N, are postulated to be involved in the GE81112 biosynthetic pathway (Fig. 2A). The first gene predicted to be involved is getA, as the proteins encoded by the orfs upstream of getA show no homology to proteins involved in biosynthetic pathways. getA (852 bp) encodes a type II thioesterase. Type II thioesterases are present in many NRPS and polyketide synthease systems, where they perform crucial proofreading functions by hydrolyzing aberrant substrates from the respective carrier protein domains (24). getA is found within an operon that also harbors the genes getB-E, getB and getC encode proteins having homology to known ABC transporter systems (25). The next gene, getD, encodes the cyclodeaminase that was targeted in our library-probing strategy. It shows 52% identity to TubZ, the cyclodeaminase from Angiococcus disciformis (14). The first gene encoding a NRPS protein (a freestanding A-domain) is getE, which likely starts with a GTG and is preceded by a putative ribosome binding site (GGAG) 7 bp upstream of the start codon. The next gene, getF, is oriented in the opposite direction and encodes a protein with homology to a putative 1-(2S)-proline 3-hydroxylase. The following gene, getG (7212 bp), encodes another NRPS and is the likely starting point of a new operon which includes getH and getI. getH (1614 bp) encodes a di-domain NRPS (PCP-C) enzyme and again starts with a GTG, whereas getI exhibits homology to an oxygenase. The last segment of the cluster contains 5 genes (getJ-getN) and starts with another change in the transcription direction. It begins with getJ (2460 bp), which encodes a NRPS di-domain (A-PCP) enzyme. The next two genes, getK and getL, encode a GTPase protein and a halogenase, respectively. The NRPS-di-domain (A-PCP) encoding gene getM (1848 bp) is assumed to start with a TTG, with a ribosome binding site (AGGG) located 6 bp upstream. The last gene getN encodes a protein with homology to a type I thioesterase. The involvement of orfs 1–5 and 6–15 in GE811112 biosynthesis is unlikely but cannot yet be excluded. Further experiments will be carried out in the future to determine the exact boundaries of the gene cluster.

Analysis of NRPS Domains—For the seven orfs with homology to NRPS genes (Fig. 3A), the constituent domains were assigned using the polyketide synthase/NRPS predictor (PKS Analysis Web site) and confirmed by manual inspection with

³ T. M. Binz, S. I. Maffioli, M. Sosio, S. Donadio, and R. Müller, unpublished data.
GE81112 Biosynthetic Gene Cluster

Table 1
Prediction of substrate specificity of GE81112 A-domains based on the specificity-conferring codes of A-domains

| A-domain | Position of the amino acid within the A-domain | Predicted amino acid | Identity |
|----------|-----------------------------------------------|----------------------|----------|
| GetE\(_A_1\) | Asp | Val | Gln | Tyr | Ile | Ala | Gln | Val | Pro/Pip | 70 |
| GetE\(_A_2\) | Asp | Ala | Tyr | Asn | Leu | Gln | Leu | Ile | Orn/Gln/Asp | 70 |
| GetE\(_A_3\) | Asp | Ala | Val | Gly | Val | Gln | Gln | Val | Tyr/Trp | 70 |
| GetM | Asp | Ser | Ala | Ser | Thr | Ala | Gln | Val | His | 70 |
| GetA\(_A_4\) | Asp | Ser | Ala | Leu | Thr | Ala | Gln | Val | His | 70 |

Table 2
Predicted function of non-PKS/NRPS proteins present up- and downstream of the GE81112 biosynthetic gene clusters

| Protein | No. of amino acids | Proposed function of the homologous protein | Origin | Identity/similarity | Accession no. |
|---------|-------------------|-----------------------------------------------|--------|---------------------|---------------|
| Orf1    | 335               | Partitioning-binding protein                  | Nocardia farcinica ATCC 53653 | 64/75 | CBL193712.1 |
| Orf2    | 502               | Hypothetical protein                          | Streptosporangium roseum DSM 43021 | 52/67 | CBL193713.1 |
| Orf3    | 476               | Hypothetical protein                          | S. sviceus DSM 43021 | 58/74 | CBL193714.1 |
| Orf4    | 1203              | FtsK/SpoIIIE family protein                   | Bradyrhizobium sp. | 78/86 | CBL193710.1 |
| Orf5    | 1238              | Serine/threonine protein kinase                | Streptomyces avermitilis MA-4680 | 53/67 | CBL193711.1 |

Blast. To incorporate four amino acids into GE81112, the synthetases were expected to contain a loading module followed by three condensation modules with the standard C-APCP arrangement. Furthermore, the simplest model predicts that the four amino acid precursors would be incorporated in a co-linear manner: pipecolic acid-ornithine/glutamine/glutamic acid-histidine-histidine. However, the get NRPS genes exhibit a non-co-linear arrangement that did not fit our expected model (Fig. 3A). Although the domain complement necessary for the biosynthesis of a tetrapeptide could be identified, one extra A-PCP di-domain was present. Moreover, the NRPS modules show a highly split arrangement, as they occurred as freestanding domains (GetE) or di-domain units (GetH, GetJ, GetM) (Fig. 3A). Thus, from the domain assignment alone, the overall order of subunits could not be discerned. To determine the substrate specificity of the A-domains, the specificity-conferring code and eight conserved motifs were identified for each A-domain, and bioinformatic analysis was employed to predict their substrate specificities (Table 1 and supplemental Fig. S2) (26, 27). The first A-domain (GetE\(_A_1\)), encoded by getE as a discrete protein, was a candidate for starter unit selection as its predicted specificity was for the incorporation of proline/pipecolic acid (Table 1). The second A-domain (GetE\(_A_2\)) showed homology to ornithine/glutamine/asparagine-incorporating A-domains. The third A-domain (GetE\(_A_3\)) was predicted to be specific for the incorporation of tyrosine/tryptophan and the two latter domains GetE\(_A_4\) and GetE\(_A_5\) for histidine (Table 1). These results correlated well with the prediction that an ornithine and two histidines are incorporated into the GE81112 metabolites. However, it remained unclear why two A-PCP di-domains (GetM and GetE) are encoded by the cluster, as only one is required to give a full complement of four active modules. To check if any of the domains were inactive, the C- and PCP-domains were analyzed as well. The get cluster encodes three C-domains that aligned well with the C-domains from the rapamycin-, the gramicidin-, and the calcium-dependent antibiotic biosynthetic clusters (28–30). The seven conserved regions were identified in all of...
the C-domains and the same analysis was carried out for the PCP-domains, revealing the signature sequence and active serine residue, in each case (supplemental Fig. S3).

These results demonstrate that, in principle, all of the NRPS domains are active. Furthermore we annotated two discrete thioesteras (encoded by getA and getN), both containing the conserved motif GXSXG, present in functional enzymes (16). BLAST analysis (Table 2) revealed that the protein GetA is more related to type II thioesters, whereas GetN is more related to type I thioesters. This finding was surprising, as GetN is a discrete protein (like a type II thioesterase) not integrated into an NRPS-like type I thioesterase typically found in bacterial systems.

**FIGURE 2.** Organization of the GE81112 biosynthetic gene cluster in Streptomyces sp. L-49973. A, a schematic representation of the GE81112 biosynthetic locus and flanking ORFs in Streptomyces sp. L-49973 is represented on two overlapping cosmids. Proposed functions for individual ORFs are summarized in Table 2. B, shown is LTQ high resolution Orbitrap MS analysis of extracts of Streptomyces sp. L-49973 wild type and Streptomyces sp. L-49973::KO1 mutant, showing a base peak chromatogram (BPC) of Streptomyces sp. L-49973 wild type extract (m/z = 100 – 2000) and extracted ion chromatograms (EIC) of GE81112 compound B (2) with a molecular ion of the mass m/z = 659.2993 [M + H]⁺ from Streptomyces sp. L-49973 wild type and Streptomyces sp. L-49973::KO1 mutant extracts. GE81112 production is abolished in the mutant.

**FIGURE 3.** A linear model of GE81112 biosynthesis. A, genetic and modular organization of the GE81112 biosynthetic gene cluster is shown. Black arrows indicate NRPS genes, and white arrows non-NRPS genes. TEI, type I thioesterase. B, proposed biosynthesis of the fourth amino acid precursor is shown. Free (2S)-histidine is activated by the A-domain and loaded to the PCP of GetM. The PCP-bound histidine is chlorinated at position 6 by the halogenase GetL and hydroxylated at the β-position by GetI, although one of these reactions may not occur on the GetM-bound amino acid. It is not clear if the halogenation or hydroxylation occurs first, and it could be either way. The type II thioesterase GetA hydrolyzes the modified histidine to give the free amino acid. C, shown is the proposed biosynthesis of GE81112 congener A. Lysine is converted to piperolic acid by the cyclodeaminase GetD followed by hydroxylation catalyzed by GetF. (2S,3S)-Hydroxypiperolic acid is then activated by the A-domain GetE and loaded to the PCP of GetH. O-HPA, histidine, and 3-hydroxy-6-chlorohistidine are activated and loaded by modules 2, 3, and 4 in the next steps, and the final tetrapeptide is released by the type I thioesterase (TEI) GetN.
Confirmation of the Role of the get Cluster by Gene Inactivation—To verify the identity of the proposed GE81112 biosynthetic gene cluster, we inactivated getE, which encodes a free-standing A-domain. For this, a knock-out construct was designed by amplifying an internal fragment, which was then cloned into the knock-out vector pKC1132 (10). Initial attempts to transform Streptomyces sp. L-49973 with the knock-out plasmid were unsuccessful, necessitating the development of an adapted transformation method. Only by using a larger ratio of E. coli cells to recipient cells were we able to obtain several exconjugants containing the knock-out vector, indicating that the number of donor cells is crucial for the conjugation efficiency with this strain (31). The resultant mutants were verified by PCR against the apramycin resistance gene as well as an internal region from the genomic DNA. The mutants were then cultivated in production medium, and the extracts were analyzed for the presence of the GE81112 compounds by high resolution MS. This analysis clearly showed that GE81112 production was abolished in the mutants, confirming the role of the cloned gene cluster (Fig. 2B).

Biochemical Analysis of Adenylation Domains—To obtain experimental evidence for A-domain substrate specificity, we expressed the five get A-domains as N-terminal His6-tagged proteins. DNA fragments coding for the adenylation domains of getE (1 domain, GetEA, size 60.39 kDa), getG (2 domains, GetGA2 and GetGA3, sizes 61.50 and 63.13 kDa), getJ (1 domain, GetJA, size 60.0 kDa), and getM (1 domain, GetMA, size 63.40 kDa) were amplified from cosmids BI11 or BA23 and cloned into pET28b vectors. The constructs were confirmed by sequencing and transformed into E. coli Rosetta BL21 (DE3) pLysS/RARE. Cultivation was carried out at 16 °C. Expression was induced with 0.2 mM isopropyl-β-D-thiogalactopyranoside at A600 = 0.8 – 1. All proteins could be obtained in the soluble fraction and were used for the ATP-PPi exchange assay after purification.

The substrate specificity of the five purified adenylation domains was evaluated using the established ATP-PPi exchange assay (32, 33). Briefly, each protein was incubated with a panel of different amino acids, including the anticipated substrate of each A-domain. As a control, each protein was incubated in the absence of added amino acid. The results as shown in Fig. 4, A–E, indicate that GetEA activated l-(2S)-Pip (100%), d-(2R)-Pip (82.9%), and l-(2S)-Pro (80.3%). The background controls were between 0.85 and 3.79%, confirming that the measured activity reflected the true substrate preference of the A-domain. GetGA2 activated l-(2S)-ornithine (100%) as well as l-(2S)-Gln (62.7%) preferentially, whereas l-(2S)-Glu and l-(2S)-Asp were activated to a minor extent. GetGA3, GetJA, and GetMA all activated l-(2S)-His (100%) as well as l-(2S)-Lys to a minor extent (between 41 and 53%). So a clear preference for l-(2S)-His could be verified for all the three proteins. Taken together, these results establish that the five proteins exhibit the enzymatic activity of adenylation domains and show preference for substrates consistent with the GE81112 structure.

Stereoconfiguration of Chiral Centers—Before this work, the stereochemistry of the chiral centers in GE81112 was not verified. Here, the stereochemistry of the chiral centers was determined by comparing the optical rotation of the natural product with that of the corresponding racemate. The results showed that all five A-domains exhibit the expected stereochemistry, providing strong evidence for the correct assignment of the chiral centers.

FIGURE 4. Relative substrate specificities of internal adenylation domains from the GE81112 biosynthetic gene cluster. Internal adenylation domains GetEA, GetGA2, GetGA3, GetJA, and GetMA were investigated in terms of activity in the ATP-PPi exchange reaction with different amino acids and a control without amino acid. The highest activities were set at 100%. The background was below 10%. The specificities of the different domains coincide with the primary structures of the GE81112s.
known. The chemical analyses described here have confirmed the indication from the ATP-PPi exchange assays that all amino acids have the (2S)-configuration and provided preliminary evidence about the other stereocenters.

NMR analyses of GE81112 dissolved in DMSO-d6 (5) do not provide useful information about the relative stereochemistry due to generally broad signals. Improved resolution and shape of the signals were achieved using D2O (as such and acidified with TFA), allowing new assignments (Table 3). Under these conditions, the signal belonging to the 1H and 13C NMR chemical shift assignments of GE8112 congener A in D2O at 298 K. The assignments were made by analysis of COSY, TOCSY (two-dimensional total correlation spectroscopy), HMQC (heteronuclear multiple quantum coherence), and HMQC (heteronuclear multiple bond coherence) spectra. Numbering is according to previous work (5).

The assignments were made by analysis of COSY, TOCSY (two-dimensional total correlation spectroscopy), HMQC (heteronuclear multiple quantum coherence), and HMQC (heteronuclear multiple bond coherence) spectra. Numbering is according to previous work (5).

TABLE 3
1H and 13C NMR chemical shift assignments of GE8112 congener A in D2O at 298 K.
The assignments were made by analysis of COSY, TOCSY (two-dimensional total correlation spectroscopy), HMQC (heteronuclear multiple quantum coherence), and HMQC (heteronuclear multiple bond coherence) spectra. Numbering is according to previous work (5).

| Residue | Group | δ 1H | δ 13C |
|---------|-------|------|-------|
| AA1     | 2     | 4.09 | 60.3  |
|         | 3     | 4.51 | 64.6  |
|         | 4     | 1.82 | 28.2  |
|         | 5     | 1.75 | 15.7  |
|         | 6     | 3.02 | 43.6  |
| AA2     | 2     | 4.56 | 50.6  |
|         | 3     | 1.84 | 33.7  |
|         | 4     | 3.85 | 65.7  |
|         | 5     | 1.92 | 69.1  |
| AA3     | 2     | 4.67 | 52.2  |
|         | 3     | 2.98 | 27    |
| AA4     | 2     | 4.50 | 59.3  |
|         | 3     | 5.20 | 67.8  |
|         | 5'    | 6.97 | 118   |

Numbers in this column represent positions.

DISCUSSION
Many antibiotics target the prokaryotic translational apparatus, but few selectively inhibit initiation. Protein translation in prokaryotes is initiated by the binding of fMet-tRNA to the ribosomal P-site. Recently, the GE81112 tetrapeptides were shown to specifically inhibit this fMet-tRNA binding by blocking the P-site and, thus, represent a unique class of inhibitors with a new mode of action (35). Identification and biochemical characterization of the GE81112 biosynthetic gene cluster now provide insights into the biosynthesis of this unique family of secondary metabolites and sets the stage for the generation of new derivatives by genetic engineering. The identity of the cloned gene cluster was confirmed by inactivation of getE, which completely abolished GE81112 production. Although many of the enzymes encoded by the get cluster are consistent with GE81112 biosynthesis, the NRPSs are present with a highly split, non-co linear module arrangement (Fig. 3A). Unusual features include two freestanding A-PCP di-domains (encoded by getJ and getM) and a stand-alone A-domain (encoded by gene getE). Stand-alone A-domains have been identified in other nonlinear NRPS pathways, e.g. myxochelin (36) and yersiniabactin (37), whereas free-standing A-PCP di-domains are found in the zorbamycin and syringomycin gene clusters (38, 39). According to bioinformatic analysis, all NRPS domains are predicted to be functional (supplemental Fig. S2 and S3). Furthermore, because the get cluster encodes five distinct A and PCP domains instead of the four predicted for a tetrapeptide, the order of subunits (and corresponding modules) could not be predicted from sequence alone. The established specificity for the five A-domains and analysis of the functions predicted from the other enzymes encoded by the cluster allow us to draw a first model for GE81112 formation (Fig. 3C).

Accordingly, the biosynthesis of GE81112 starts with the formation of (2S)-piperoc acid from (2S)-lysine via the action of the putative cyclodeaminase GetD (Fig. 3C). Piperoc acid is directly activated by the A-domain GetEA1 and then loaded onto the adjacent PCP domain. Consistently, an ATP-PPi exchange assay confirmed that GetEA1 preferentially recognizes (2S)-piperoc acid (Fig. 4A). However, the piperoc acid moiety in GE81112 is hydroxylated at the β-position. There is precedent for β-hydroxylation to occur at three different stages; on the free amino acid (40), whereas the amino acid is tethered to the PCP (41), or on the mature peptide after thioesterase hydrolysis from the NRPS (42, 43). As GetEA1 could not be directly assayed with hydroxypiperoc acid due to the commercial
GE81112 Biosynthetic Gene Cluster

unavailability of this compound, we suggest that this domain might also recognize hydroxypypecolic acid. In any case, the presence of the 3-hydroxyl group on the pipicolate moiety is essential for activity, as GE81112 derivatives lacking this group are 3 orders of magnitude less active in the translation assay than the parent compounds.4

It was not obvious from which amino acid the second building block is derived. For GE81112 congeners A (1) and B (2), the second amino acid residue is 5-hydroxy-2-aminopentanoic acid (O-HPA), which is hydroxylated at position 4 and O-carbamoylated at position 5, whereas congener B1 (3) contains a hydroxylated and carbamoylated ornithine residue (Fig. 1). We hypothesize that the A-domain GetGA2 is responsible for the incorporation of the second amino acid in GE81112. A recent example in the biosynthesis of the nucleoside antibiotic polyoxin shows that O-carbamoyl-polyhydroxypentaenoic acid is generated from free (2S)-glutamate by stepwise reduction, O-carbamoylation, and hydroxylations (44). The complete building block is then attached to the nucleoside. Similarly, GE81112 congeners A and B could derive from activation of the O-HPA substrate (or of a precursor) by the A-domain GetGA2 and loading on the corresponding PCP (Fig. 3C). However, we could not identify candidate genes in the get cluster for O-HPA biosynthesis, so they may be encoded elsewhere in the genome (45, 46). According to our hypothesis, the A-domain GetGA2 must show a broad substrate acceptance, activating O-HPA- and ornithine-based amino acids to generate the different GE81112 congeners. As O-HPA or any derivatives were not commercially available, we tested the likely precursors of O-HPA, (2S)-glutamic acid and (2S)-glutamine in addition to (25)-ornithine. Activation of all the three amino acids (ornithine, Glu, and Gln) was observed (Fig. 4B) with a preference for (25)-ornithine. The reduced degree of activation of glutamic acid and glutamine may indicate that modified versions of glutamic acid and glutamine are preferred, consistent with the hypothesis that O-HPA is formed before loading to the PCP. It should be noted that the major congeners of the GE81112 complex produced by Streptomyces sp. L-49973 are A and B,5 indicating that in vivo O-HPA or a precursor is the preferred substrate for the NRPS.

The third amino acid incorporated into the GE81112 peptide is either histidine (1) or aminohistidine (2 and 3). In silico analysis of GetGA3 predicts specificity for tyrosine or tryptophan rather than histidine. This prediction may indicate a general preference of the A-domain for aromatic amino acid residues, which might account for activation of both histidine and aminohistidine. However, it is not clear how the aminohistidine moiety is generated: “amination” might occur at the PCP-bound histidine after release of the peptide from the NRPS, or alternatively, aminohistidine could derive from the cyclization of arginine by nucleophilic attack of the γ-carbon. According to the ATP-PPi exchange assay, histidine is the preferred substrate, and no activation was observed with arginine, tryptophan, or tyrosine (Fig. 4C). These data suggest that aminohistidine is not generated from PCP-bound arginine, tryptophan, or tyrosine, but we could not establish whether GetGA3 also recognizes aminohistidine, as this amino acid is not commercially available. Thus, the nature and timing of histidine amination remains unclear.

A 3-hydroxy-6-chlorohistidine is the last amino acid to be incorporated in GE81112. Halogenation is a common modification found in bioactive natural products (47–49), and the timing of halogenation reactions has been shown to vary. For example, in rebeccamycin, pre-assembly line chlorination of tryptophan occurs (50), whereas a PCP-bound threonine is chlorinated for syringomycin (51). The putative halogenase GetL is the likely candidate for histidine chlorination in GE81112. GetL shows homology to non-heme iron-dependent oxygenases/hydroxylases and is, therefore, assumed to hydroxylate histidine at the β-position. Figs. 3, B and C, illustrate the proposed mechanism for the chlorination/hydroxylation of histidine and its incorporation into GE81112. As there is evidence from sequence analysis and in vitro experiments (supplemental Fig. S2 and Fig. 4E) that the A-PCP domains of the seemingly superfluous GetM protein are active, this di-domain may play a role in the biosynthesis of the fourth amino acid precursor. Indeed, there are a number of examples in which specialized A-PCP di-domains are essential for generating NRPS precursors (52–56). Here, we propose that the A-domain GetMA3 activates (2S)-histidine, consistent with the results of the ATP-PPi exchange assay, and tethers it to the PCP of GetM. Hydroxylation and/or halogenation reactions then occur on the PCP-bound amino acid as catalyzed by GetL and GetL, respectively (Fig. 3B). The thioesterase GetA would then release the modified amino acid, as described for BarC from the barbamide biosynthetic pathway (53). In the subsequent step, the free, modified histidine is activated by the A-domain GetJA4 and loaded onto the PCP of module 4 (Fig. 3C). This is supported by the ATP-PPi exchange assay showing that GetJA4 has specificity for (25)-histidine (Fig. 4D). Again, the relative timing of chlorination and hydroxylation is unclear, as it is conceivable that one of these reactions may occur on the GetJ-tethered amino acid or peptide. In Fig. 3C, GetA is predicted to function as a type II thioesterase to release the modified amino acid from GetM. Alternatively, GetA could act as an aminoacyltransferase, shuttling the PCP-bound modified histidine from GetM to GetJ. There are several examples of such PCP-to-PCP shuttling reactions (38, 55), and the aminoacyltransferases identified to date have been assigned into two groups (38); one group, comprising SyrC, CmaE, and ZmbVIIId, contains a GXXG motif at the active site, and these enzymes are predicted to act as acyltransferases, with the active-site cysteine shown to be directly involved in aminoacyl transfer (52); the second group includes the acyltransferases BarC (53) and CouN7 (54), with an active site serine in the GXXG motif, and are predicted to function as normal thioesterases and not as aminoacyltransferases (supplemental Fig. S4). GetN is expected to catalyze the release of the peptide from the assembly line (Fig. 3C).

The combination of genetic, biochemical, and chemical analyses demonstrate that GE81112 is composed of L-amino acids only. This result is consistent with the observation that it is a

4 S. I. Maffioli, unpublished results.
5 M. Sosio and S. I. Maffioli, unpublished results.
Acknowledgments—We thank Dr. Shwan Rachid for helpful advice on the construction of the cosmID library of Streptomyces sp. L-49973. We thank Dr. Kira Weissman and Dr. Luke Simmons for critical reading of the manuscript and helpful discussions. We thank M. Scharfe and H. Böcker for sequencing of the cosmids at the Helmholtz Centre for Infection Research, Braunschweig. We are indebted to former colleagues at Vicuron Pharmaceuticals for fermentation and chemical support.

REFERENCES

1. Hopwood, D. A., Levy, S., Wenzel, R. P., Georgopapadokou, N., Balfz, R. H., Bhavnani, S., and Cox, E. (2007) Nat. Rev. Drug Discov. 6, 8–12
2. Baker, D. D., Chu, M., Oza, U., and Rajagaria, V. (2007) Nat. Prod. Rep. 24, 1225–1244
3. Walsh, C. (2003) Antibiotics: Actions, Origins, Resistance, American Society for Microbiology, Washington, D.C.
4. Kohanski, M. A., Dwyer, D. J., and Collins, J. J. (2010) Nat. Rev. Microbiol. 8, 423–435
5. Brandl, L., Lazzarini, A., Cavalletti, L., Abbondio, M., Corti, E., Cicilato, I., Gastaldo, L., Marazzi, A., Ferroggi, M., Fabbretti, A., Maio, A., Colombo, L., Donadio, S., Marinelli, F., Losi, D., Gualerzi, C. O., and Selva, E. (2006) Biochemistry 45, 3692–3702
6. Finkel, R., and Marahiel, M. A. (2004) Annu. Rev. Microbiol. 58, 453–488
7. Hutchinson, C. R. (2003) Proc. Natl. Acad. Sci. U.S.A. 100, 3010–3012
8. Wenzel, S. C., and Müller, R. (2005) Curr. Opin. Chem. Biol. 9, 447–458
9. Wenzel, S. C., and Müller, R. (2007) Nat. Prod. Rep. 24, 1211–1224
10. Kieser, T., Bibb, M., Buttner, M. J., Chater, K. F., and Hopwood, D. A. (2001) Practical Streptomyces Genetics pp. 562 and 566, The John Innes Foundation, Norwich, England
11. Sambrook, J., and Russell, D. W. (2001) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
12. Dürr, C., Schnell, H. J., Luzhetsky, A., Murillo, R., Weber, M., Welzel, K., Vente, A., and Bechtold, A. (2006) Chem. Biol. 13, 365–377
13. Rachid, S., Krug, D., Kunze, B., Kochens, I., Scharfe, M., Zabirske, T. M., Blocker, H., and Müller, R. (2006) Chem. Biol. 13, 667–681
14. Sandmann, A., Sasse, F., and Müller, R. (2004) Chem. Biol. 11, 1071–1079
15. Khaw, L. E., Böhme, G. A., Metcalfe, S., Staunton, J., and Leadlay, P. F. (1998) J. Infect. Dis. 178, 809–814
16. Marahiel, M. A., Stachelhaus, T., and Mootz, H. D. (1997) Chem. Rev. 97, 2651–2674
17. Silakowski, B., Schairer, H. U., Ehret, H., Kunze, B., Weinig, S., Nordsiek, G., Brandt, P., Blocker, H., Höfle, G., Beyer, S., and Müller, R. (1999) J. Biol. Chem. 274, 37391–37399
18. Girard, A., Greck, C., and Genêt, J. P. (1998) Tetrahedron Lett. 39, 4259–4260
19. Dong, L., and Müller, M. J. (2002) J. Org. Chem. 67, 4759–4770
20. Rose, T. M., Schultz, E. R., Henikoff, J. G., Pietrokovski, S., McCullam, C. M., and Henikoff, S. (1998) Nucleic Acids Res. 26, 1628–1635
21. Gatto, G. J., Jr., Boyne, M. T., 2nd, Kelleher, N. L., and Walsh, C. T. (2006) J. Am. Chem. Soc. 128, 3838–3847

A. Maio and S. Donadio, unpublished results.

substrate of the oligopeptide permease in some bacterial species.6 Many unusual features were found in the biosynthesis of this all-l-riboside binding tetrapeptide. The NRPS modular architecture includes A-domains that incorporate unusual amino acids, such as hydroxypropionic acid, hydroxypropyntoic acid, and hydroxychlorohistidine. Although many questions about its biosynthesis remain, the availability of the GE81112 cluster as well as tools for genetic manipulation now provide a platform for attempts to decipher these issues and to generate new GE81112 derivatives by genetic engineering.