Genome mining of lipopeptide biosynthesis of *Paenibacillus polymyxa* E681 in combination with mass spectrometry – discovery of the lipoheptapeptide paenilipoheptin

Joachim Vater,*[a] Stefanie Herfort,[a] Joerg Doellinger,[a] Max Weydmann,[a] Rainer Borriss,[b,c] and Peter Lasch[a]

Abstract: *Paenibacillus polymyxa* strains are qualified for agrobiotechnological uses such as plant growth promotion and for biocontrol strategies against deleterious phytopathogenic competitors in the soil depending on their attractive arsenal of bioactive compounds. On the other hand they are potent producers of antibiotics for medical applications. To identify new products of such organisms genome mining strategies in combination with mass spectrometry are the methods of choice. Here we performed such studies with the *Paenibacillus* strain E681. Bioinformatic evaluation of its genome sequence revealed four gene clusters A-D encoding nonribosomal peptide synthetases (NRPS). Accordingly, four lipopeptide families were detected by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS). Clusters A and D codify the well known fusaricidins and polymyxins. A yet unknown lipoheptapeptide was discovered and structurally characterized by de novo sequencing using MALDI-LIFT-TOF/TOF MS. It was designated as paenilipoheptin. From structure predictions we infer that the production of this agent is encoded by gene cluster C. Gene cluster B encodes the synthesis of tridecaptins, a family of open-chain lipotridecapeptides. Strain E681 produced new subspecies of such compounds (tridecaptins E) showing variations both in their fatty acid as well as in their peptide part.

Introduction

With the advances achieved by up-to-date nucleotide sequencing techniques a new era in drug discovery has been initiated accelerating the speed of genome sequencing of microorganisms enormously. This progress opened efficient pathways for the prediction of gene clusters encoding new bioactive compounds and for understanding the routes of the underlying biosynthetic processes. In particular, in the realm of *Bacillales* in the meantime hundreds of genomes have been completely sequenced. This immense array of data enabled the development of genome mining strategies for the detection, structure analysis and exploitation of yet unknown bioactive compounds[1-11]. As particularly successful revealed procedures, which combine bioinformatic genome screening with modern mass spectrometry utilizing the unsurpassed sensitivity and accuracy of such techniques for the detection of novel compounds and for structure elucidation[5].

Paenibacilli are a rich source of attractive secondary metabolites, among them nonribosomally produced peptides (NRP), polyketides (PK), NRP-PK hybrids and bacteriocins which are of great interest both for agrobiotechnological and medical uses[6-8,12]. The genomes of numerous *Paenibacillus* strains were completely sequenced[13-16], all of them being equipped with gene clusters encoding the biosynthesis of such compounds and revealing the molecular organization of the cell factories producing them. Genome mining studies have been initiated with several of these strains to detect and evaluate their biosynthetic capacity [6,7]. Structure prediction, in particular of nonribosomally synthesized peptides, has been reported, but the identification and structure analysis of several of these products are still missing. Our work is focused on this important task by a mass spectrometry based genome mining of representative *Paenibacillus* strains. Our present research mainly concerns *P. polymyxa* strains which are efficient plant growth promoting rhizobacteria (PGPR) colonizing surfaces of plant roots in the rhizoplane. In addition, because of their impressive arsenal of bioactive products they are of great benefit as biocontrol agents protecting plants against deleterious phytopathogenic microorganims in the soil. In particular, *P. polymyxa* strains E681 and M1 are distinguished as reference organisms to outline perspectives for biotechnological and medical applications. Though these strains were investigated in detail on the genomic level, precise data is still missing concerning the identification and structural characterization of a major part of the predicted products which is the aim of our work presented here.

Previously, we investigated the fusaricidins[17,18] and polymyxins[19] of strain M1 extensively. In this paper our efforts are focused on exploring the biosynthetic potential of strain E681. In its genome four gene clusters for the production of nonribosomal peptides were demonstrated encoding four lipopeptide families. According to the derived structure

[a] PD Dr. J.Vater, S. Herfort, Dr. J. Doellinger, M. Weydmann, Dr. P.Lasch
Robert Koch-Institut Berlin
ZBS6 Proteomics and Spectroscopy
Seestr. 10, 13353 Berlin (Germany)
E-Mail: VaterJ@rki.de

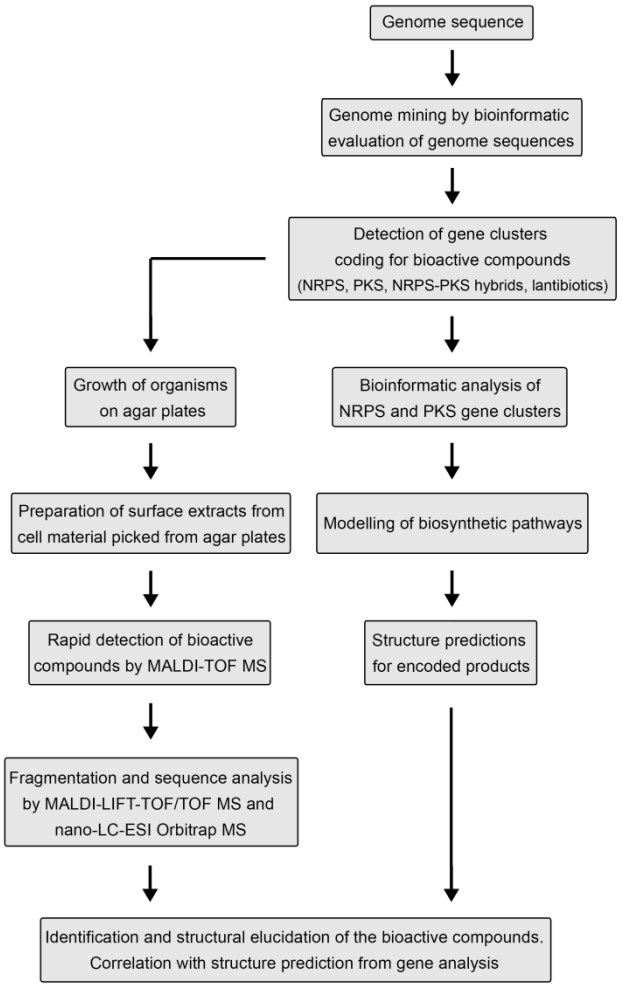
[b] Prof, Dr. R. Borriss
Humboldt Universität Berlin
Fachgebiet Phytomedizin
Lentzeallee 55-57, 14195 Berlin (Germany)

[c] Prof. Dr. R. Borriss
NordReet UG
Marienstr. 27a, 17489 Greifswald (Germany)

predictions two of them correspond to the well known fusaricidins and polymyxins. Most important, we identified a yet unknown lipoheptapeptide which we characterized in detail mass spectrometrically. This compound was designated by us as paenilipoheptin.Presumably, its biosynthesis is encoded by the third gene cluster C. Interestingly, the fourth gene cluster encodes new subspecies of tridecaptins, a family of linear tridecapeptides showing potent activities against Gram-negative organisms.

Results and Discussion

In this paper we explored the biosynthetic capacity of *P. polymyxa* E681 by genome mining in combination with sensitive advanced mass spectrometric techniques according to scheme 1.



Scheme 1. MS-combined genome mining of *Paenibacillus polymyxa*

Genetic and structure elucidation workflow

Starting from the genome sequence of E681[15] we performed genome mining studies to identify the gene clusters coding for the producer multienzyme systems of nonribosomal peptides using bioinformatic strategies. In particular, we used antiSMASH version 4 [20] which is one of the most efficient and commonly used programs to identify, annotate and evaluate biosynthetic gene clusters. Their architecture gives insight into the structural and functional organization of the corresponding multienzymes and allows modeling of the biosynthetic pathways as well as structure predictions of their products[21] which ultimately have to be detected in cultures of the producer organisms and structurally characterized by chemical and spectroscopic methods. For this task modern mass spectrometry, particularly MALDI-TOF and ESI MS, is the method of choice. For this purpose we cultivated *Paenibacillus* strains on agar plates and prepared surface extracts from picked cell material.
In this report we investigated the production of lipopeptides produced by *P. polymyxa* E681. Such compounds generally appear attached to the outer surface of microorganisms. Therefore, they can efficiently be made available for testing from surface extracts which are well suited for their rapid, sensitive detection by MALDI-TOF MS[17,18]. Mass spectrometric structure analysis started with the determination of their amino acid components by their immonium ions, followed by the detection of nearest neighbor relationships derived from fragment spectra. This information formed the basis for de novo sequencing of the peptide products for their identification and structural characterization.

Genome mining studies

In accordance with previous studies[6,7] bioinformatic evaluation of the genome sequence of E681[15] (NC 014483.2) using antiSMASH 4.0 (http://antismash.secondarymetabolites.org) revealed four gene clusters involved in lipopeptide biosynthesis. Their organization is summarized in Figure 1A-D. They were listed in the order of their appearance along the genome sequence. All these gene clusters show assembly lines of modular units dissected in functional domains (A, T; E) each in combination with a condensation domain C managing peptide bond formation between adjacent amino acid components which are collinearly arranged with the amino acid components in the peptide product. The tertiary structure of the reaction centers for amino acid binding and activation in the adenylation domain (A) allows prediction of the nature of the amino acid substrates[21] and defines the code for nonribosomal peptide biosynthesis. The module arrangement along the assembly lines in Figure 1 reflects the sequence of the peptide products. In this way the NRPS type cluster A (81,155 – 111,975) and D (4,536,009 - 4,562,187) were attributed to the biosynthesis of the well known fusaricidins[17,18,22-26] and polymyxins[19,27,28] which have been investigated in detail in previous work. Fusaricidin is produced by a single six-module enzyme, while in the biosynthesis of polymyxin 3 multifunctional enzymes cooperate.

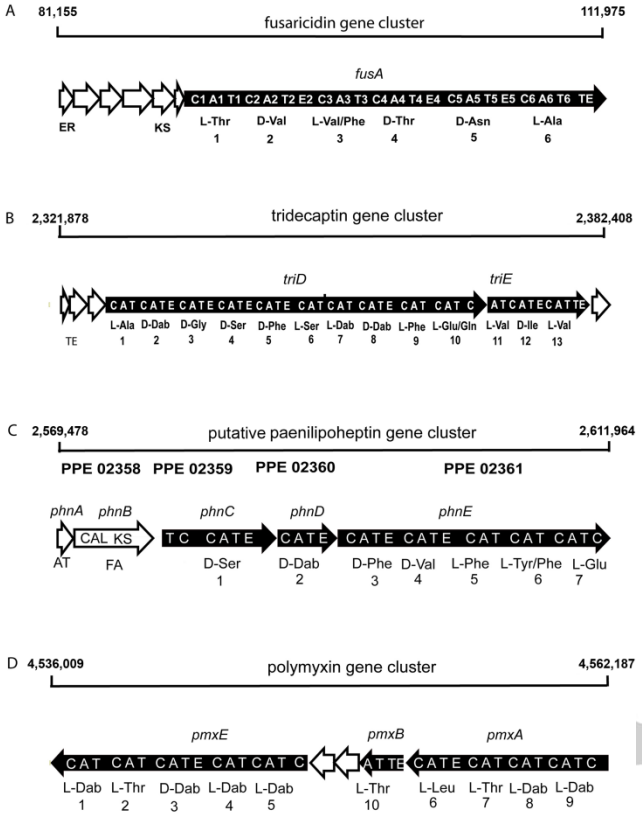


Figure 1. Organization of the gene clusters in the genome of *P. polymyxa* E681 coding for nonribosomal biosynthesis of the lipopeptides fusaricidins (A); tridecaptins (B), unknown heptapeptides (paenilipoheptins) (C) and polymyxins (D). The sequences of their peptide moieties were predicted by bioinformatic evaluation of the A-domains of their amino acid specific modules. Domain organization: A, adenylation domain; T, thiolation domain; C, condensation domain; E, epimerization domain; TE, thioesterase domain; CAL, CoA ligase domain; KS, ketosynthase domain; AT, aminotransferase domain.

Our work is particularly focused on the hybrid Nrps-Trans-AT-PKS gene cluster C (2,569,478 – 2,611,964) which encodes the biosynthesis of a yet unknown lipoheptapeptide product whose genes show 50% similarity to the paenilarvin gene cluster (BGC0000402_c1)[29,30]. The existence of such a compound has recently been predicted by genome mining studies performed by Aleti et al.[6] and Xie et al.[7], but its specific nature and structure remained unresolved. Cluster C comprises seven amino acid activating and condensing modules. The entire assembly line is subdivided into three genes coding for three peptide synthetases Phn C, D and E. Phn C and D are one module enzymes activating and epimerizing Ser and Dab, respectively. Phn E consists of 5 modules completing the biosynthetic process. Structure prediction on the basis of bioinformatic evaluation implies the following sequence:

D-Ser – D-Dab – D-Phe – D-Val – L-Phe – L-Tyr/Phe – L-Glu

The functional organization of the modules of cluster C allowed to assign the configuration of the amino acid components in the product. Modules 1-4 that are equipped with an epimerization domain (E) indicating a L- to D-conversion of the activated amino acid substrates. In its N-terminal region this gene cluster includes domains for a CoA-ligase, a ketosynthase unit involved in fatty acid or polyketide biosynthesis and an aminotransferase similar to the biosynthesis of iturin compounds (iturins, mycosubtilins and bacillomycins), well known lipoheptapeptides produced by numerous *B. subtilis* and *B. amyloliquefaciens* strains[31-33]. These features imply the presence of a N-terminal β-amino fatty acid. Here the fatty acid precursor activated by the CoA-ligase is elongated by two carbon atoms to its final length involving a ketosynthase. The β-keto-intermediate is ultimately converted to the β-amino fatty acid component which is then linked to the starting amino acid.

Remarkably, gene cluster C does not show a thioesterase domain at its 3'-end. Possibly termination of lipoheptapeptide biosynthesis may be accomplished by an external, free-standing thioesterase enzyme. This hypothesis remains to be clarified.

New information about the product pattern of *P. polymyxa* E681 was also achieved for another Nrps gene cluster B (2,321,878 – 2,382,408) which resembles the cell factory for the production of tridecaptins (Fig. 1B) which are open-chain lipotridecapeptides containing a β-hydroxy-fatty acid component[12,34-37].The tridecaptin gene cluster contains two genes *tri*D and *triE* coding for the two partial enzymes of tridecaptin synthetase providing an assembly line of 13 amino acid activating and condensing modules. *triD* comprises the genetic information for the assembly of the first ten amino acid substrates. TriE completes the biosynthesis of tridecaptin. Sequence prediction from bioinformatic evaluation of the *tri* gene cluster suggests the tridecapeptide shown in Figure. 1B.

L-Ala-D-Dab-D-Gly-D-Ser-D-Phe-L-Ser-L-Dab-D-Dab-L-Phe-L-Glu/Gln-L-Val-D-Ile-L-Val

The presence of an epimerization domain E in modules 2-5, 8 and 12 implies that the corresponding amino acid components in the tridecaptin sequence are present in D-configuration. The tridecaptin gene cluster does not contain the genetic information for the biosynthesis of its β-hydroxy fatty acid constituent similar to other well known lipopeptides, such as the surfactins and fengycins from *B. subtilis* and *B. amyloliquefaciens*, for examples[33].

Mass spectrometric identification and structure elucidation of the lipopeptide products of E681

In our work we combined genome mining of E681 with sensitive mass spectrometric detection and efficient structure analysis. In order to identify the NRPs produced by E 681 a rapid screening was performed by MALDI-TOF MS in surface extracts of *P. polymyxa* E681. For this purpose this organism was grown on agar plates for 72 h using the TSA medium. Cell material was picked and surface extracts were prepared by extraction with 50% acetonitrile/0.1% TFA.

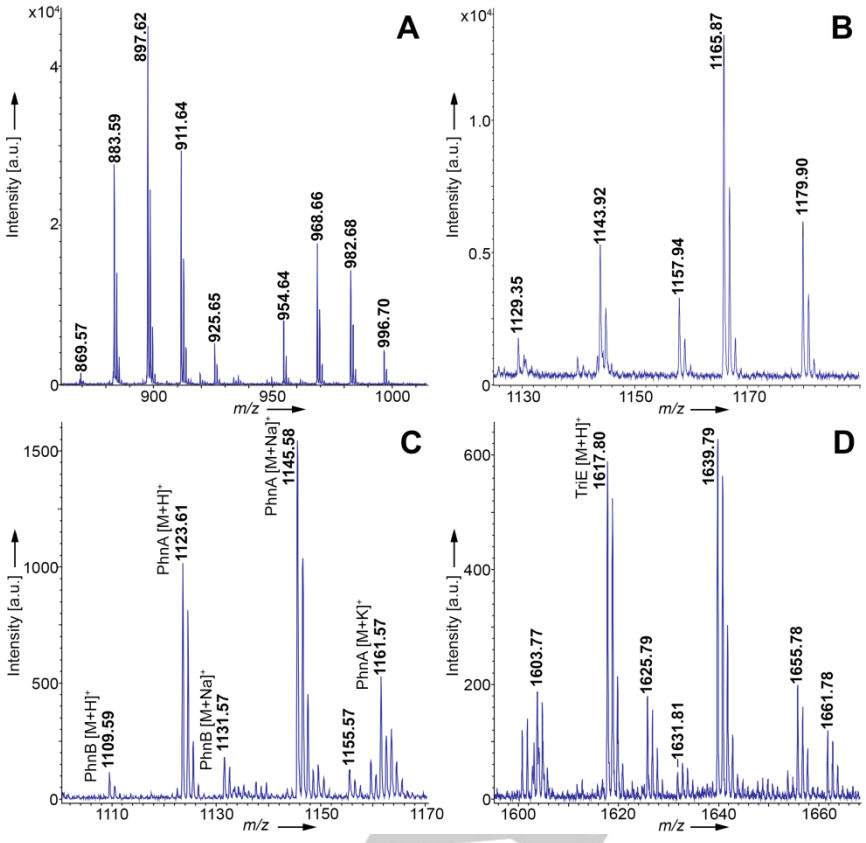


Figure 2. MALDI-TOF mass spectra of nonribosomal lipopeptides produced by *P. polymyxa* E681.
A) fusaricidins; B) polymyxins A; C) unknown lipoheptapeptides (paenilipoheptins) and D) tridecaptins.

In accordance with the genome mining studies four product families were detected in the extracts (Figure. 2A-D). They were presented with increasing mass numbers, as they appear in the mass spectrum. The ensembles found in the mass ranges *m/z* 860-1050 (Figure. 2A, corresponding to gene cluster A), *m/z* 1140-1200 (Figure. 2B, corresponding to gene cluster D) and *m/z* 1580-1680 (Figure. 2D, corresponding to gene cluster B) were attributed to the well known fusaricidins[17,18,22-26], polymyxins[19,27,28] and tridecaptins[12,34-37]. The main fusaricidin species of *P. polymyxa* E681 are fusaricidins A, B and E in contrast to strain M1 which exhibits fusaricidins A, B, C and D as the dominant products[17,18]. The polymyxins produced by strain E681 are of type A. They appear only in low amount. Their mass peaks were strongly superposed by those of yet unknown compounds showing mass peaks between m/z 1100-1180 (Figure. 2C). Therefore, the polymyxins could only be detected after HPLC separation and purification. For this purpose the complete product mixture of strain E681 was applied to a Zorbax 300 SP-C8 column and eluted by a gradient from 0-70% B as described in the Experimental Section.

Of particular interest in our research was the hitherto unknown peptide family in the mass range *m/z* 1100-1180 (Figure. 2C) which represents a mixture of two homologues with mass numbers $[M + H]^+$ = 1109.66 and 1123.59.

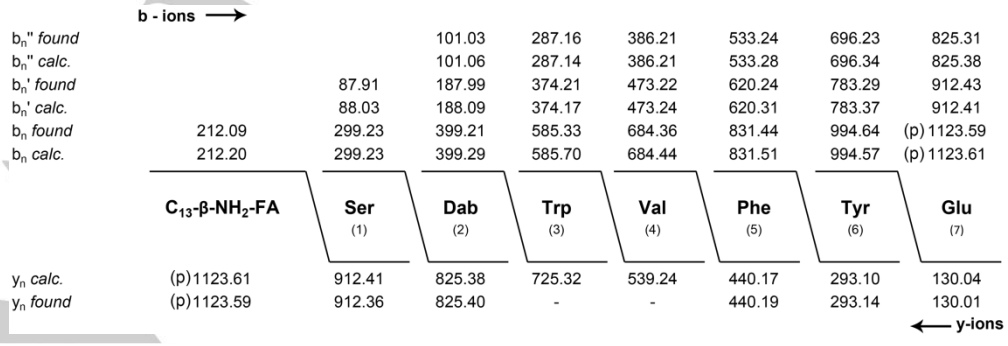


Figure 3. De novo sequencing of paenilipoheptin A (*m/z* 1123.61) predicted from gene cluster C obtained by genome mining of *P. polymyxa* E681 using MALDI-LIFT-TOF/TOF MS. b_8 and y_8 correspond to the parent ion p.

For interi

Table 1. Modeling of the paenilipoheptin A sequence by combination of nearest neighbor relationships (di-, tri- and tetrapeptide fragments) obtained by MALDI-LIFT-TOF/TOF MS.

								m/z	
a)	Dipeptide fragments							Calc.	Found.
C13-FA	Ser							299.23	299.21
	Ser	Dab						188.09	188.00
		Dab	Trp					287.14	287.14
			Trp	Val				286.15	286.19
				Val	Phe			247.14	247.12
					Phe	Tyr		311.13	311.15
						Tyr	Glu	293.10	293.12
b)	Tripeptide Fragments								
C13-FA	Ser	Dab						399.29	399.23
	Ser	Dab	Trp					374.17	374.21
		Dab	Trp	Val				386.21	386.22
			Trp	Val	Phe			433.22	-
				Val	Phe	Tyr		410.20	410.26
					Phe	Tyr	Glu	440.17	440.19
c)	Tetrapeptide Fragments								
C13-FA	Ser	Dab	Trp					585.37	585.31
	Ser	Dab	Trp	Val				473.24	473.20
		Dab	Trp	Val	Phe			533.28	533.24
			Trp	Val	Phe	Tyr		596.28	596.28
				Val	Phe	Tyr	Glu	539.24	-

C_{13}-FA means a C_{13}-β-amino fatty acid constituent.

These compounds were putatively related to gene cluster C in Figure. 1C. Our efforts were mainly focused on the identification and structure analysis of these unknown lipoheptapeptides.

On the basis of their biosynthetic features derived from genome mining studies we initiated *de novo* sequencing of compound A with a mass number of *m/z* 1123.59. Structural characterization was performed by fragment analysis applying MALDI-LIFT-TOF/TOF MS. Amino acid analysis detecting immonium ions ions indicated the presence of Val (*m/z* 72); Phe (*m/z* 120); Tyr (*m/z* 136) and Trp (*m/z* 156). Nearest neighbor relationships (di-, tri- and tetrapeptide fragments) were compiled and summarized in Table 1. Fragment analysis yielded the following amino acid sequence of the complete peptide part:

Ser(1)-Dab(2)-Trp(3)-Val(4)-Phe(5)-Tyr(6)-Glu(7)

It corresponds to the sequence of the b_n´-fragment ions in Figure. 3. This result is corroborated by the sequence of the b_n´´- ions starting from Dab(2).

The mass difference between the molecular mass of the complete product (*m/z* 1123.59) and the obtained heptapeptide portion (*m/z* 911.41) amounts to $[M + H]^+$ = 212.09 Da which is compatible with a C_{13}-β-amino fatty acid component assuming ring formation between its amino group with the C-terminal carboxyl group of L-Glu. The calculated mass number of this fatty acid residue is 212.20 corresponding to an elemental composition of $C_{13}H_{25}N_1O_1$. Starting with the b_1-fragment (*m/z* 212.09) of the C_{13}-fatty acid constituent the complete sequence of the ring-opened lipopeptide was determined as shown in Figure. 3 (b_n- and y_n-ions).

Table 1 shows a modeling of this sequence by combination of the obtained nearest neighbor relationships (di-, tri- and tetrapeptide fragments).

We designate these novel cyclic lipoheptapeptides produced by *P. polymyxa* E681 as paenilipoheptins. The subspecies with mass numbers *m/z* 1123.59 and *m/z* 1109.55 were assigned as paenilipoheptins A and B. The b_1-ion of paenilipoheptin B *m/z* 198.11 is 14 mass units lower than that of species A indicating a C_{12}-fatty acid component. This is corroborated by the finding that the y-type fragment ions of both subspecies exhibit the same amino acid sequence. According to the results of the mass spectrometric analysis the structure prediction in Figure. 1 has to be corrected for position 3. Here Trp was found instead of Phe as derived from the genetic level.

The lipopeptide products encoded by gene cluster B correspond to the tridecaptin family detected by MALDI-TOF MS in the range of *m/z* 1600-1680 (Figure 2D). Tridecaptins are open-chain tridecapeptides in combination with a ß-hydroxy-fatty acid attached to the N-terminus. Here we investigated the tridecaptin variant with the mass number *m/z* 1617.88 in detail. The sequence of this tridecaptin species was investigated by fragment analysis using MALDI-LIFT-TOF/TOF MS.

Table 2. Modelling of the sequence of the tridecaptin variant *m/z* 1617.88 by combination of nearest neighbor relationships (di-, tri- and tetrapeptide fragments) obtained by MALDI-LIFT-TOF/TOF MS.

												m/z	
a)	Dipeptide fragments											Calc.	Found.
C13-FA	Ala											284.22	284.23
	Ala	Dab										172.10	-
		Dab	Gly									158.08	158.20
			Gly	Ser								145.05	145.18
				Ser	Trp							274.11	274.29
					Trp	Ser						274.11	274.29
						Ser	Dab					188.09	188.22
							Dab	Dab				201.12	201.25
								Dab	Trp			287.14	287.33
									Trp	Glu		316.12	316.33
										Glu	Val	299.11	299.26
											Val	Ile/Leu	213.15	213.29
												Ile/Leu	$Ala(H_20)$	203.13	203.29
b)	Tripeptide fragments														
C13-FA	Ala	Dab												384.28	384.31
	Ala	Dab	Gly											229.12	229.15
		Dab	Gly	Ser										245.11	245.28
			Gly	Ser	Trp									331.13	331.35
				Ser	Trp	Ser								361.14	361.33
					Trp	Ser	Dab							374.17	374.43
						Ser	Dab	Dab						288.15	288.34
							Dab	Dab	Trp					387.20	-
								Dab	Trp	Glu				416.18	416.47
									Trp	Glu	Val			415.19	415.47
										Glu	Val	Ile/Leu		342.19	342.43
											Val	Ile/Leu	$Ala(H_20)$	302.19	302.38
c)	Tetrapeptide fragments														
C13-FA	Ala	Dab	Gly											441.30	441.36
	Ala	Dab	Gly	Ser										344.18	-
		Dab	Gly	Ser	Trp									431.19	431.51
			Gly	Ser	Trp	Ser								418.16	418.44
				Ser	Trp	Ser	Dab							461.20	461.49
					Trp	Ser	Dab	Dab						474.23	474.23
						Ser	Dab	Dab	Trp					474.23	474.23
							Dab	Dab	Trp	Glu				516.24	-
								Dab	Trp	Glu	Val			515.25	515.60
									Trp	Glu	Val	Ile/Leu		528.27	528.67
										Glu	Val	Ile/Leu	$Ala(H_20)$	431.23	431.50

C_{13}-FA means a C_{13}-β-amino fatty acid constituent.

Mass spectrometric amino acid analysis indicated the following components: Ala (*m/z* 44); Ser (*m/z* 60); Val (*m/z* 72); Ile/Leu (*m/z* 86) and Trp (*m/z* 159). On the basis of these results and the structure predictions derived from genome mining nearest neighbor relationships were compiled in Table 2. The mass difference between the molecular mass of the complete product (*m/z* 1617.91) and the tridecapeptide portion (*m/z* 1404.70) amounts to 213.17 Da which is compatible with a C_{13}-ß-hydroxy-fatty acid component. The calculated mass number of this fatty acid residue is 213.18 corresponding to an elemental composition of $C_{13}H_{25}O_2$.

Using this knowledge we performed de novo sequencing of the tridecaptin variant with the mass number *m/z* 1617.91. The obtained sequence is demonstrated in Figure. 4. Table 2 shows the modeling of this sequence by combination of the compiled nearest neighbor relationships (di-, tri- and tetrapeptide fragments). This tridecaptin contains two serine residues in positions 4 and 6 at which in the course of mass spectrometric fragmentation water is eliminated. Therefore, starting from these positions two series of fragment ions $b_n(-H_2O)$ and $b_n(-2H_2O)$ were observed which support the sequence obtained for the unmodified product (b_n-ions).

In our MALDI-LIFT-TOF/TOF MS studies it was not possible to discriminate between Leu, Ile and allo-Ile in position 12 of the tridecaptin sequence. Here the assignment was performed according to the structure prediction from our genome mining

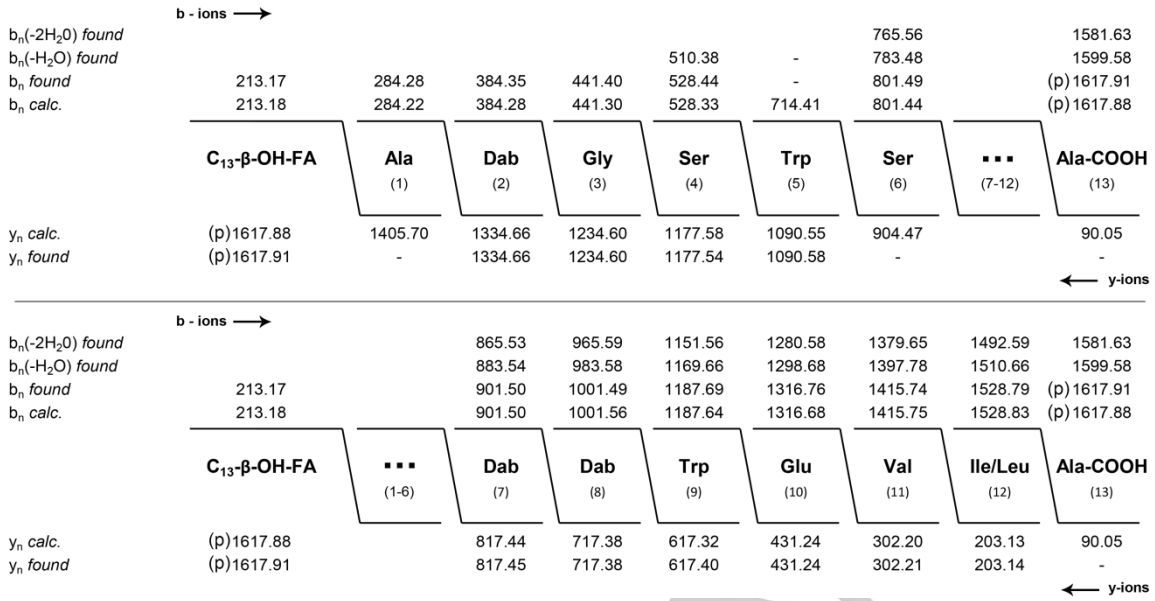
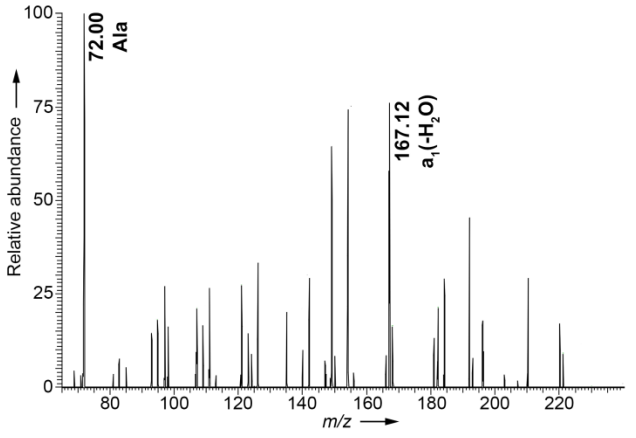


Figure 4. De novo sequencing of the tridecaptin variant with a mass number of *m/z* 1617.88 encoded by gene cluster B using MALDI-LIFT-TOF/TOF MS. b_{13} and y_{13} correspond to the parent ion p.

studies and comparison with the literature data[12,34-37]. The mass spectrometric results also did not allow to distinguish optical isomers. Therefore, the configuration of the amino acid components in paenilipoheptins and tridecaptins was derived from the domain organization of the modules along the assembly lines.

To clarify the N-terminal amino acid of the investigated tridecaptin species which could either be Ala or Val, we performed direct infusion nano-electrospray ionization experiments in the ion trap of a LTQ Orbitrap instrument. To discriminate between these two possibilities fragmentation of the triply charged parent ion with *m/z* 539.97 was performed in the linear trap quadrupole of the Orbitrap in a MS^4 experiment. Fragment b_2 was selected which contained the N-terminal amino acid attached to the ß-hydroxy fatty acid constituent. The decomposition of b_2 led along the pathways of fragment ions outlined in Figure. 5. The prominent fragment ions are *m/z* 167.12 and *m/z* 72.0 consistent with L-Ala forming the start amino acid of the tridecaptin product. The corresponding end point of *m/z* 139 expected in the case of valine as the endstanding amino acid was completely missing. Elimination of CO and H_2O is much more pronounced starting from *m/z* 284.28 (b_2) rather than from *m/z* 213.22 (b_1), because the intensity of the a_2-ion (*m/z* 256.27) is much higher than that of a_1 (*m/z* 185.19) starting from b_1. The molecular mass of $a_1(-H_2O)$ (*m/z* 167.12) is compatible with a hydroxy fatty acid residue of 13 carbon atoms attached to the N-terminus.



Pathways for the decomposition of fragment b_2

	b_2	-CO	a_2	$-H_2O$	$a_2(-H_2O)$	-Ala	$a_1(-H_2O)$
m/z found	284.28		256.25		238.21		167.12
m/z calc.	284.22		256.23		238.22		167.18

	b_2	-Ala	b_1	-CO	a_1	$-H_2O$	$a_1(-H_2O)$
m/z found	284.28		213.22		185.20		167.12
m/z calc.	284.22		213.18		185.19		167.18

Figure 5. Direct infusion nano-electrospray ionization ion trap MS^4 spectrum of the triply charged parent ion of tridecaptin (*m/z* 539.97) which was isolated in the ion trap of a LTQ-Orbitrap and selected for low-energy CID-type fragmentation. The obtained fragment ions were detected in the ion trap using enhanced resolution mode demonstrating the decomposition of fragment ion b_2 into the fatty acid constituent ($a_1(-H_2O)$; *m/z* 167,18) and the N-terminal amino acid Ala (*m/z* 72.04).

Discussion

The genome sequences of numerous *P. polymyxa* strains[13-16, 38-40] are now available for genome mining studies. Representative *P. polymyxa* strains are ATCC 842[14]; E681[15]; M1[16]; SQR-21[38]; SC2[39] and CR1[40] isolated from the rhizosphere of

cultivated plants. All of them are efficient producers of bioactive compounds. AntiSMASH-analysis[20] of the genome of strain E681 which has been investigated here, revealed an equipment of 33 gene clusters involved in synthesis of secondary metabolites, comprising four NRPS clusters (*fus*, *pmx, tri*, *phn* genes) as well as one cluster for the production of each a polyketide (Trans-AT-PKS-Otherks_NRPS) according to the definition given by antiSMASH analysis with similarity to bacillaene, a lantipeptide, a lassopeptide and a bacteriocin which still have to be identified. Characteristic features of all these organisms are genes involved in plant growth promotion coding for the production of plant hormone-like compounds, particularly indole-3-acetic acid, as well as volatiles, such as 3-hydroxy-2-butanone and 2,3-butanediol[39,40-43]. Furthermore, these organisms exhibit promising applications for biomass degradation and biofuel production[40].

Most of the nonribosomally produced peptides of the genus P*aenibacillus* are lipopeptides linking a peptide moiety with a fatty acid constituent. They form the dominant products of *P. polymyxa* strains and exhibit attractive antibiotic properties. They show a broad structural diversity of linear and cyclic species with peptide chain lengths between 6 and 13 amino acid components and characteristic variations of their amino acid moieties as well as their fatty acid side chains. The impressive collection of these compounds cover fusaricidins (lipohexapeptides containing a guanidinylated-β-hydroxy fatty acid)[17,18,22-26]; paenilipoheptins (lipoheptapeptides, β-amino fatty acid); octapeptins (lipooctapeptides, β-hydroxy fatty acid)[12,44,45] ; polypeptins (lipononapeptides, β-hydroxy fatty acid)[12,46,47]; polymyxins (lipodecapeptides, unsubstituted fatty acid components)[19,27,28], tridecaptins[12,34-37] (lipotridecapeptides, β-hydroxy fatty acid) and paenibacterins (cyclic lipotridecapeptides, pentadecanoic acid)[48,49]. Octapeptins and polypeptins are produced by *Paenibacillus elgii*[46], *ehimensis*[47] and *tianmuensis*[45]. Fusaricidins, paenilipoheptins, polymyxins and tridecaptins are products of *P. polymyxa* and *P. peoriae*. Generally, all these lipopeptides appear as families comprising numerous homologues.

The highlight of our work presented here is the discovery of a novel family of cyclic lipoheptapeptides produced by *P. polymyxa* E681 which we designated as paenilipoheptins. The gene cluster putatively involved in nonribosomal paenilipoheptin synthesis is not common in *P. polymyxa*. It was detected in addition to E681 only in the genome sequence of *P. polymyxa* strain YC0573 (CP017968.2) and in one strain classified as *P. peoriae* (CP011512.1), but in no other of the available complete genome sequences of this species including *P. polymyxa* M1.

Paenilipoheptins exhibit a heptapeptide moiety rich in aromatic amino acid components (Phe, Tyr, Trp) and contain a C_{12} or C_{13}-β-amino fatty acid which is linked via its amino group with the COOH-group of the C-terminal L-Glu. The presence of a β-amino fatty acid constituent in paenilipopeptins is implied from our genome mining studies because of the close relationship in the organization of the *phn* gene structure with that of iturin compounds produced by numerous *B. subtilis* and *B. amyloliquefaciens* strains[31,32] exhibiting domains for the final processing of the fatty acid component directly coupled with the assembly line for heptapeptide biosynthesis. This result is consistent with our mass spectrometric data of paenilipoheptins. Interestingly, *Paenibacillus larvae*, a honey bee pathogen inducing the american foulbrood, produces a family of lipoheptapeptides, the paenilarvins, which are directly related to iturins[29,30].

Cyclic lipoheptapeptides are frequently used motifs of *Bacillales*. Well known are the heptapeptide compounds surfactins and iturins produced by numerous *B. subtilis* and *B. amyloliquefaciens* strains[50,5] which exhibit manifold valuable activities, such as antimicrobial, antiviral and antimycoplasmatic effects which qualify them both for medical and biotechnological applications.

The main products of *P. polymyxa* strains usually are the fusaricidins, a large family of lipohexapeptides of unusual complexity containing a guanidino-C_{15}, C_{17} or C_{19} –β-hydroxy fatty acid. They appeared as a complex of more than 80 structural variants in culture filtrates of strain M1[17,18]. Fusaricidins are distinguished by potent activities against fungi of the genera *Fusarium* and *Aspergillus* as well as Gram-positive bacteria. They function as efficient biocontrol agents, effective against *Fusarium* strains[52,53] or *Phytophthora* blight infection in red pepper[54], for example.

Polymyxins are the lead compounds of *P. polymyxa* strains. They represent lipodecapeptides which are rich in α,γ-diamino-butyric acid (DAB). They are acylated either by (*S*)-6-methylheptanoic acid or (*S*)-6-methyloctanoic acid. So far more than 15 structural variants are known[28]. They display potent activities against Gram-negative bacteria and have been applied as commercial antibiotics until the seventieth of the last century. Then they were retracted from the market due to their harmful side activities, in particular, because of neuro- and nephro-toxic effects, but in recent times they again attain a reapproval for treatment of life-threatening infections against multi-resistant Gram-negative pathogens, such as *Klebsiella pneumonia*, *Pseudomonas aeruginosa* or *Acetinobacter baumannii*. Polymyxin B and colistin (polymyxin E) serve as last resort antibiotics for curing serious Gram-negative infections[28]. On the other hand, polymyxin P has been tested in plant protection as an efficient biocontrol agent against deleterious *Erwinia* strains which are inducers of the fire blight and soft rot diseases[19].

Tridecaptins[12,34-37] are linear cationic lipopeptides showing a sequence of 13 amino acids. Their N-terminal amino acid is linked to a β-hydroxy fatty acid side chain. Tridecaptins also form a family of subspecies that vary both in length and configuration of their fatty acid component as well as by amino acid substitutions in their peptide moiety. The hitherto reported tridecaptins exhibit side chains of 9-11 carbon atoms. In our work presented here we contribute to the diversity of such compounds presenting new species with $C_{11\text{-}13}$-β-hydroxy fatty acid components as well as specific amino acid replacements. These variants we designate as tridecaptins E relating to E681.

Similar to the polymyxins, tridecaptins are distinguished by potent activities against Gram-negative bacteria. They are less active against Gram-positive organisms. They have successfully been tested against multidrug resistant pathogens, such as *Klebsiella pneumoniae* and *Acinetobacter baumannii* [37].

Paenibacterin, another cyclic lipotridecapeptide, was isolated from *Paenibacillus* OSY-SE consisting of a N-terminal C_{15} fatty acyl chain and 13 amino acid residues[48]. It is rich in basic amino acids (Lys, Orn). Here the C-terminal Ile is connected to Thr in position 3 by an ester linkage. The paenibacterin gene cluster *pbt* encodes three NRPS multienzymes PbtA, PbtB and PbtC. PbtA and PbtB both consist of five modules, while PbtC is a three-modular enzyme[49]. Paenibacterin exhibits potent

activities against both Gram-positive as well as Gram-negative bacteria. It efficiently inactivates multiresistant pathogens, such as methicillin-resistant *Staphylococcus aureus* or *Salmonella enterica*, *serovar typhimurium* qualifying also this antibiotic for medical applications.

In summary, most of the lipopeptide products isolated from *P. polymyxa* strains display strong activities against multidrug-resistant organism covering both Gram-negatives as well as Gram-positives. Therefore, they are candidates which should be tested in clinical studies to overcome the insistent problem of antibiotic resistance. The investigation of the antimicrobial spectrum of the paenilipoheptins and biocontrol experiments with the lipopeptides produced by *P. polymyxa* E681 are in preparation.

Conclusions

The immense progress achieved in genome sequencing and the availability of a large number of microbial genomes initiated a new era in drug discovery by genome mining strategies. Here we performed such studies with the *P. polymyxa* strain E681 which is of high relevance both for medical and plant biotechnological uses. In particular, we studied the gene clusters coding for the biosynthesis of lipopeptides. E681 produces four families of such compounds, the fusaricidins, polymyxins, tridecaptins and a hitherto unknown family of lipoheptapeptides which were detected by MALDI-TOF MS and structurally characterized by MALDI-LIFT-TOF/TOF MS. They were designated as paenilipoheptins. They are cyclic compounds which contain a $C_{12\text{-}13}$-β-amino fatty acid integrated into the peptide ring. In addition, we contributed new $C_{12\text{-}13}$-tridecaptin variants (tridecaptins E). All these compounds are promising candidates as novel antibiotics.

Experimental Section

Chemicals

The matrix alpha-cyano-hydroxycinnamic acid was obtained from Bruker (Bremen, Germany). Acetonitrile (ACN, HPLC grade) was from Merck (Darmstadt, Germany). Trifluororacetic acid (TFA) was purchased from Sigma Aldrich (Steinheim, Germany).

Cultivation of *Paenibacillus polymyxa* E681

Paenibacillus polymyxa E681 was kept frozen at -20 °C using the MAST Cryobank™ system (MAST Diagnostica GmbH, Reinfeld, Germany) and grown on Tryptic Soy Agar (TSA, Soybean Casein Digest Agar) (Merck Millipore, Darmstadt, Germany) for 24; 48; 72 and 96 h at pH = 7.3. Optimal production of lipopeptides was achieved at 72 h.

Secondary metabolite gene cluster prediction and analysis tools

The architecture of the gene clusters were predicted using the antiSMASH program version 4[20]. To analyze adenylation domains of NRPS and AT domains of PKS, NRPS/PKS substrate predictor [21] was used. Firstly, Genebank accession numbers were given as input for antiSMASH. The predicted secondary metabolite gene clusters from antiSMASH comprised NRPS, PKS, hybrid PKS/NRPS, siderophoresand bacteriocins. The clusters responsible for biosynthesis of lipopeptides were analyzed. Finally, lipopeptide encoding gene clusters were subjected to BLAST to find the closest homologue available in the database.

HPLC fractionation of the products of *P. polymyxa* E681

Surface extracts of *P. polymyxa* E681 were prepared by extraction of cell material grown on TSA agar with 50% ACN containing 0.1% TFA (1 mL) and centrifuged at 10 000 rpm for 5 min and fractionated by reversed-phase HPLC (rp-HPLC) using an Agilent (1200 series) instrument (Agilent Technologies, Waldbronn, Germany). The organic phase (330 µL) was diluted with TFA (0.1%; 660 µL) and applied to a Zorbax 300 SP-C8 column (4.6 by 150 mm; 3.5 µm; rapid solution). The products of E681 were eluted by a two-step gradient of eluent B (from 0 to 70% in 70 min and from 70 to 95% in 5 min) at a flow rate of 0.5 mL/min followed by isocratic elution at 95% eluent B for 10 min at a flow rate of 1.0 mL/min. Eluent A was TFA in water (0.1%); eluent B was 99.9% ACN/0.1% TFA. Fractions of 1 mL were collected and evaporated to dryness in a SpeedVac evaporator (Uniequip, Martinsried, Germany). The dried material was dissolved in 50% aqueous ACN/0.1% TFA (30 µL) and tested by MALDI-TOF MS.

MALDI-TOF MS analysis of the E681 products

The products in surface extracts of *P. polymyxa* E681 and in HPLC fractions were detected by MALDI-TOF MS using a Bruker Autoflex Speed TOF/TOF mass spectrometer (Bruker Daltonik, Leipzig, Germany) with smartbeam™ laser technology using a 1 kHz frequency-tripled Nd-YAG laser (excitation wavelength 355 nm). Samples of the E681 extracts (2µL) and the HPLC-fractions were mixed with an equal volume of matrix solution (a saturated solution of α-cyano-4-hydroxycinnamic acid in aqueous ACN (50%) containing TFA (0.1%)), spotted on the target, air dried and measured. Mass spectra were obtained by positive ion detection and reflector mode. Monoisotopic masses were recorded. Parent ions $[M + H]^+$ were detected with a resolution of 10000.

Sequence analysis of the lipopeptide products of E681 was performed by MALDI-LIFT-TOF/TOF mass spectrometry in Laser Induced Decay (LID) mode, as outlined previously[55]. The product ions in the LIFT-TOF/TOF fragment spectra were obtained with a resolution of 1000. 5-10 single spectra were evaluated for fragmentation of a lipopeptide.

Direct infusion nano-electrospray ionization mass spectrometry

Tridecaptin was further characterized in MS^n experiments using direct infusion nano-electrospray ionization mass spectrometry. Dried HPLC fractions were resuspended in 50 % acetonitrile/0.1 % formic acid (20 µL) and directly sprayed into a LTQ Orbitrap Discovery™ mass spectrometer (Thermo Fisher Scientific, Rockford, IL, USA). Tridecaptin was ionized using a stainless steel emitter, I.D. 30 µm, (Proxeon, Odense, Denmark) at a spray voltage of 2.0 kV and a heated capillary temperature of 275°C at a flow rate of 500 nL/min. Full scan spectra were acquired in the *m/z* range of 300 – 2000 with a resolution of 30000 in the orbitrap. Triply-charged tridecaptin (*m/z* 539.98) was selected for low-energy CID-type fragmentation in the ion trap with normalized collision energies of 50 % (MS^2 and MS^3) or 40 % (MS^4) using an activation time of 30 ms. The *m/z* isolation width for MS^N fragmentation was set to 3 Th. The detection of the fragment ions was performed in the linear trap quadrupole (LTQ) using enhanced resolution mode. Automatic gain control was set to a target value of 10^6 for full scans and 5×10^3 for MS^N spectra with maximum allowed ion accumulation times of 500 ms and 200 ms, respectively. Ten single spectra were accumulated for MS^4 spectra. Raw spectra were processed using Xcalibur™ (Thermo Fisher Scientific, Rockford, IL, USA).

Keywords: Genome mining – mass spectrometry – peptides – *Paenibacillus polymyxa* E681 - paenilipoheptins - tridecaptins

[1] H. Gross, *Curr. Opin. Drug Discov. Devel.* **2009**, *12*, 207.

[2] G.L. Challis, *Microbiol.* **2008**, *154*, 1555.

[3] M. Zerikly, G.L. Challis, *ChemBioChem.* **2009**, *10*, 625.

[4] J.E. Velasquez, W. van der Donk, *Curr. Opin. Chem. Biol.* **2011**, *15*, 11.

[5] R.D. Kersten, Y-L. Yang, Y. Xu, P. Cimermancic, S.-J. Nam, W. Fenical, M. A. Fischbach, B.S. Moore, P.C. Dorrestein, *Nat. Chem. Biol.* **2012**, *7*, 794.

[6] G. Aleti, A. Sessitsch, G. Brader, *Compt. Struct. Biotechnol. J.* **2015**, *13*, 19.

[7] J. Xie, H. Shi, Z. Du, T. Wang, X. Liu, S. Chen, *Sci. Rep*. **2016**, *6*:21329;DOI:10.1038.

[8] X. Zhao, O.P. Kuipers, *BMC Genomics* **2016**, *17*, 882.

[9] J. I. Tietz, D. A. Mitchell, *Curr. Top. Med. Chem.* **2016,** *16*, 1645**.**

[10] W. Wohlleben, Y. Mast, E. Stegmann, N. Ziemert, *Microb. Biotechnol*. **2016**, *9*, 541.

[11] J. Paterson, G. Jahanshah, Y. Li, Q. Wang, S. Mehnaz, H.Gross, *FEMS Microbiol. Ecol.* **2017**, *93*, 1.

[12] S.A. Cochrane, J.C. Vederas, *Med. Res. Rev*. **2016**, *36*, 4.

[13] A.W. Eastman, D.E. Heinrichs, Z.C. Yuan, *BMC Genomics* **2014**, *15*, 851.

[14] H. Jeong, S.Y. Park, W.-H. Chung, S.H. Kim, N. Kim, S.-H. Park, J.F. Kim *J. Bacteriol*. **2011**, *193*, 5026.

[15] J.F. Kim, H. Jeong, S.-Y. Park, S.-B. Kim, Y.K. Park, S.-K. Choi, C.-M. Ryu, C.-G. Hur, S.-Y. Ghim, T.K. Oh, J.J. Kim, C.S. Park, S.-H. Park, *J. Bacteriol.* **2010**, *192*, 6103.

[16] B. Niu, C. Rueckert, J. Blom, Q. Wang, R. Borriss, *J. Bacteriol.* **2011**, *193*, 5862.

[17] J. Vater, B. Niu, K. Dietel, R. Borriss, *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1548.

[18] J. Vater, S. Herfort, J. Doellinger, M. Weydmann, K. Dietel, S. Faetke, P. Lasch, *J. Mass Spectrom.* **2017**, *52*, 7.

[19] B. Niu, J. Vater, C. Rueckert, J. Blom, M. Lehmann, J.-J. Ru, X.-H. Chen, Q. Wang, R. Borriss, *BMC Microbiol.* **2013**, *13*, 137.

[20] K. Blin, T. Wolf, M.G. Chevrette, X. Lu, C.J. Schwalen, S.A. Kautsar, H.G. Suarez Duran, E.L.C. de los Santos, H. U. Kim, M. Nave, J.S. Dickschat, D.A. Mitchell, E. Shelest, R. Breitling, E. Takano, S.Y. Lee, T. Weber, M.H. Medema, Nucleic Acids Res. **2017**, *45*, W36.

[21] B. I. Khayatt, L. Overmars, R. J. Siezen, C. Francke, *PLoS ONE* **2013**,*8*, e62136.

[22] Y. Kajimura, M *J.* Kameda, *Antibiot. (Tokyo)* **1996**, *49*, 129.

[23] Y. Kajimura, M. *J.* Kameda, *Antibiot. (Tokyo)* **1997**, *50*, 220.

[24] J. Kuroda, T. Fukai, M. Konishi, J. Ono, K. Kurusu, T. Nomura, *Heterocycles* **2000**, *53*, 1533.

[25] J. Kuroda, T. Fukai, T. Nomura, *J. Mass Spectrom.* **2001**, *36*, 30.

[26] S.-K. Choi, S.-Y. Park, R. Kim, C.-H. Lee, J.F. Kim, S.-H. Park, *Biochem. Biophys. Res. Commun*. **2008**, *365*, 89.

[27] S.-K. Choi, S.-Y. Park, R. Kim, S.-B. Kim, C.-H. Lee, J.F. Kim, S.-H. Park, J. Bacteriol. **2009**, *191*, 3350.

[28] T. Velkov, P.E. Thompson, R.L. Nation, J. Li, *J. Med. Chem.* **2010***, 53*, 898.

[29] S. Sood, H. Steinmetz, H. Beims, K.I. Mohr, M. Stadler, M. Djukic, W. von der Ohe, M. Steinert, R. Daniel, R. Müller, *ChemBioChem* **2014**,*15*, 1947.

[30] G. Hertlein, M. Seiffert, S. Gensel, E. Garcia-Gonzalez, J. Ebeling, R. Skobalj, A. Kuthning, R.D. Süssmuth, E. Genersch, PLoS One **2016**, 11, e0164656.doi:10.1371.

[31] E.H. Duitman, L.W. Hamoen; M. Rembold, G. Venema, H. Seitz, W. Saenger, F. Bernhard, R. Reinhardt, M. Schmidt, C. Ullrich, F. Leenders, J. Vater, *Proc. Natl. Acad. Sci. USA*, **1999**, *96*, 13294.

[32] K. Tsuge, T. Akiyama, M. Shoda, *J. Bacteriol*. **2001**, *183*, 6265.

[33] A. Koumoutsi, X.-H. Chen, A. Henne, H. Liesegang, G. Hitzeroth, P. Franke, J. Vater, R. Borriss, *J. Bacteriol*. **2004**, *186*, 1084.

[34] T. Kato, H. Hinoo, J.Shoji, *J. Antibiot. (Tokyo)* **1978**, *31*, 652.

[35] T. Kato, R. Sakazaki, H. Hinoo, J. Shoji, *J. Antibiot. (Tokyo)* **1979**, *32*, 305.

[36] C.T. Lohans, M.J. van Belkum, S.A. Cochrane, Z. Huang, C.S. Sit, L. M. McMullen, J. C. Vederas, *ChemBioChem.* **2014**, *15*, 243.

[37] S.A. Cochrane, C.T. Lohans, M.J. van Belkum, M.A. Bels, J. C. Vederas, *Org. Biomol. Chem.* **2015**, *13*, 6073.

[38] S. Li, D. Yang, M. Qiu, J. Shao, R. Guo, B. Shen, X. Yin, R. Zhang, N. Zhang, Q.Shen, *Genome Announcements* **2014**, *2*, pii: e 00281-14; doi 10.1128.

[39] M. Ma, C.Wang, Y. Ding, L. Li, D. Shen, X. Jiang, D. Guan, F. Cao, H. Chen, R. Feng, X. Wang, Y. Ge, L. Yao, X. Bing, X. Yang, J. Li, B. Du, *J. Bacteriol.* **2011**, *193*, 311.

[40] A.W. Eastman, B. Weselowski, N. Nathoo, Z.-C. Yuan, *Genome Announcements* **2014**, *2*, pii: e01218-13. doi: 10.1128.

[41] S. Lal, S. Tabacchioni, *Ind. J. Microbiol*. **2009**, *49*, 2.

[42] E. N. Grady, J. MacDonald, L. Liu, A. Richman, Z.-C. Yuan, *Microb. Cell Fact.* **2016**, *15*, 203.

[43] M. A. Farag, H. Zhnag, C.-M. Ryu, *J. Chem. Ecol*. **2013**, *39*, 100

[44] T. Velkov, K.D. Roberts, J. Li, *Nat. Prod. Rep*. **2017**, *34*, 295.

[45] C.-D. Qian, X.-C. Wu, Y. Teng, W.-P. Zhao, O. Li, S.-G. Fang, Z.-H. Huang, H.-C. Gao, *Antimicrob. Agents Chemother*. **2012**, *56*, 1458.

[46] C.-D. Qian, T.-Z. Liu, S.-L. Zhou, R. Ding, W.-P. Zhao, O. Li, X.-C. Wu, *BMC Microbiol*. **2012**, *12*, 197.

[47] Z. Huang, Y. Hu, L. Shou, M. Song, *BMC Microbiol*. **2013**, *13*, 87.

[48] Y. Guo, E. Huang, C. Yuan, L. Zhang, A.E.Yousef, *Appl. Environ. Microbiol*. **2012**, *78*, 3156.

[49] E. Huang, Y. Guo, A.E. Yousef, *Res. Microbiol*. **2014**, *165*, 243.

[50] J.-M. Bonmatin, O. Laprevote, F. Peypoux, *Comb. Chem. High Throughput Screen*. **2003**, *6*, 541.

[51] J. Vater, X. Gao, G. Hitzeroth, C. Wilde, P. Franke, *Comb. Chem. High Throughput Screen*. **2003**, *6*, 557.

[52] W. Raza, X.M. Yang, H.S. Wu, Y. Wang, Y.C. Xu, Q.R Shen,. *Eur. J. Plant Pathol*. **2009**, *125*, 471.

[53] I. Zalia-Kolsi, A. Ben Mahmoud, H. Ali, S. Sellami, Z. Nasfi, S. Tounsi, K. Jamoussi. *Microbiol. Res*. **2016**, *192*, 148.

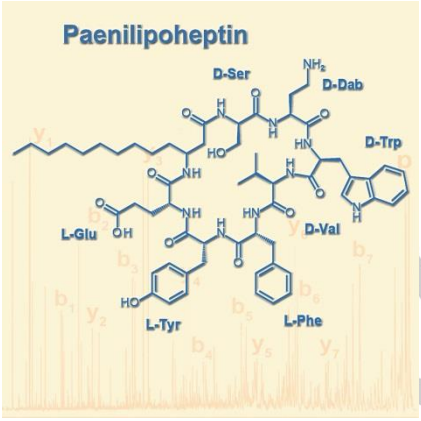
[54] S.H. Lee, Y.E. Cho, S.-H. Park, K. Balaraju, J.W. Park, S.W. Lee, K. Park, *Phytoparasitica* **2013**, *41*, 49.

[55] D. Suckau, A. Resemann, M. Schuerenberg, P. Hufnagel, J. Franzen, A. Holle, *Anal. Bioanal. Chem.* **2003**, *376*, 952.

Entry for the Table of Contents (Please choose one layout)

FULL PAPER

Analysis of the organization of the *phn* gene cluster led to the discovery of novel lipo-heptapeptides, the paenilipo-heptins, which were detected and structurally characterized mass spectrometrically



*Author(s), Corresponding Author(s)**

Page No. – Page No.

Title