Universidade de Aveiro Departamento de Biologia
2015

Tiago Filipe Melo Santos

Análise bioinformática do genoma de *Pedobacter* sp. NL19

Bioinformatics analysis of the *Pedobacter* sp. NL19 genome

DECLARAÇÃO

Declaro que este relatório é integralmente da minha autoria, estando devidamente referenciadas as fontes e obras consultadas, bem como identificadas de modo claro as citações dessas obras. Não contém, por isso, qualquer tipo de plágio quer de textos publicados, qualquer que seja o meio dessa publicação, incluindo meios eletrónicos, quer de trabalhos académicos.

Universidade de Aveiro Departamento de Biologia
2015

Tiago Filipe Melo Santos

Análise bioinformática do genoma de *Pedobacter* sp. NL19

Bioinformatics analysis of the *Pedobacter* sp. NL19 genome

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biologia Molecular e Celular, realizada sob a orientação científica da Doutora Sónia Alexandra Leite Velho Mendo Barroso, Professora Auxiliar com Agregação do Departamento de Biologia da Universidade de Aveiro e da Doutora Tânia Isabel Sousa Caetano, bolseira de Pós-Doutoramento no Departamento de Biologia da Universidade de Aveiro.

Dedico aos meus pais João e Maria, e irmão Miguel.

o júri

presidente

Prof. Doutora Maria de Lourdes Gomes Pereira
professora associada com agregação da Universidade de Aveiro

Prof. Doutora Isabel da Silva Henriques
investigadora auxiliar da Universidade de Aveiro

Prof. Doutora Tânia Isabel Sousa Caetano
bolseira de pós-doutoramento da Universidade de Aveiro

agradecimentos

Começo por agradecer à Professora Sónia pela orientação dada ao longo dos anos e por me ter aberto sempre as portas para continuar a trabalhar no Laboratório. Agradeço também pelo interesse e disponibilidade que sempre demonstrou para me ajudar.

Quero agradecer à Tânia, pela paciência, auxílio e orientações dadas ao longo do(s) ano(s). Sem a sua indispensável ajuda não teria conseguido realizar e apresentar o presente trabalho.

Um obrigado à Cláudia, por me ter aperfilhado e ajudado ao longo deste ano a desenvolver este trabalho, com a sua NL19.

Não posso deixar de agradecer também à Andreia pela sua dedicação, interesse e orientação ao longo destes anos, entre os quais foi minha orientadora.

Agradeço também a todos os membros do Laboratório que conheci e que me ajudaram e acima de tudo alegraram ao longo deste(s) ano(s), Cátia, Diana, Joana B., Joana L., Liliana, Marta, Margarida e Teresa.

Tenho a agradecer também a todos os amigos com quem me cruzei nestes últimos seis anos pelos bons momentos e amizade demonstrada.

Por último, agradecer aos meus pais por me ajudarem a chegar até aqui e pelo interesse pelo que fazia. Agradeço também ao meu irmão Miguel pela paciência e apoio ao longo destes anos. Um agradecimento especial também para o Snoopy, pelos seus 16 anos de amizade.

palavras-chave

Pedobacter sp. NL19; sequenciação; genoma; bioinformática; metabolitos secundários; lantipéptidos; filogenia; espectrometria de massa

resumo

O final do século XX marcou o advento da engenharia genética, que culminou com o desenvolvimento de diversas técnicas, como o PCR ou a sequenciação de Sanger. Isto permitiu o aparecimento de novas técnicas de sequenciação de genomas, conhecidas como *next-generation sequencing*. Uma das suas aplicações é a procura *in silico* de novos metabolitos secundários, sintetizados por microrganismos e com ação antimicrobiana. Os péptidos antimicrobianos podem ser classificados em péptidos ribossomais e péptidos não-ribossomais, de acordo com a sua biossíntese.

Os lantipéptidos são os péptidos ribossomais mais estudados, sendo caracterizados pela presença de lantioninas e metillantioninas na sua estrutura, que resultam de modificações pós-traducionais. Estes podem ser classificados em quatro classes consoante a sua maquinaria de biossíntese. Na classe I, resíduos de serina e treonina são desidratados no terminal C do péptido precursor por uma enzima LanB. Em seguida, estes resíduos sofrem ciclização por ação de uma enzima LanC, formando ligações de lantionina. A clivagem e transporte são posteriormente realizadas por duas enzimas LanP e LanT, respectivamente. Na classe II uma enzima bifuncional LanM é responsável pela desidratação e ciclização, e uma enzima LanT, pela clivagem e transporte.

Pedobacter sp. NL19 é uma bactéria de Gram-negativo, isolada a partir de lamas de uma mina de urânio abandonada, em Viseu (Portugal). Possui atividade antimicrobiana *in vitro* contra várias bactérias de Gram-positivo e de Gram-negativo. A sequenciação e análise do genoma desta bactéria permitiu identificar a presença de 21 *clusters* biossintéticos para metabolitos secundários, incluindo *clusters* que codificam para péptidos ribossomais e não-ribossomais. Foram identificados quatro *clusters* de lantipéptidos contendo péptidos precursores, enzimas de modificação (LanB e LanC) de classe I, e a enzima bifuncional LanT, de classe II. Este resultado revela a existência de *clusters* de genes híbridos, pouco descritos na literatura, possuindo características de duas classes distintas. A análise filogenética efectuada revelou que as enzimas destes *clusters* agrupam dentro da clade de bacteroidetes. Assim, verificou-se que outras espécies deste filo também possuem os *clusters* de gene híbridos de lantipéptidos, mostrando que esta não é uma característica rara neste grupo de organismos. Por fim, a análise de colónias da NL19 por MALDI-TOF MS permitiu detectar uma massa com 3180 Da, correspondente à massa prevista para um lantipéptido codificado por um dos *clusters* híbridos. Contudo, este resultado não é totalmente conclusivo e mais procedimentos experimentais terão que ser realizados para caracterizar totalmente o potencial destes péptidos. Assim, a análise realizada revelou que a bactéria NL19 possui potencial para produzir diversos metabolitos secundários, incluindo lantipéptidos que não se encontram ainda funcionalmente caracterizados.

keywords

Pedobacter sp. NL19; sequencing; genome; bioinformatics; secondary metabolites; lanthipeptides; phylogeny; mass spectrometry

abstract

The last decades of the 20th century defined the genetic engineering advent, climaxing in the development of techniques, such as PCR and Sanger sequencing. This, permitted the appearance of new techniques to sequencing whole genomes, identified as next-generation sequencing. One of the many applications of these techniques is the *in silico s*earch for new secondary metabolites, synthesized by microorganisms exhibiting antimicrobial properties. The peptide antibiotics compounds can be classified in two classes, according to their biosynthesis, in ribosomal or nonribosomal peptides.

Lanthipeptides are the most studied ribosomal peptides and are characterized by the presence of lanthionine and methylanthionine that result from post-translational modifications. Lanthipeptides are divided in four classes, depending on their biosynthetic machinery. In class I, a LanB enzyme dehydrate serine and threonine residues in the C-terminus precursor peptide. Then, these residues undergo a cyclization step performed by a LanC enzyme, forming the lanthionine rings. The cleavage and the transport of the peptide is achieved by the LanP and LanT enzymes, respectively. Although, in class II only one enzyme, LanM, is responsible for the dehydration and cyclization steps and also only one enzyme performs the cleavage and transport, LanT.

Pedobacter sp. NL19 is a Gram-negative bacterium, isolated from sludge of an abandon uranium mine, in Viseu (Portugal). Antibacterial activity *in vitro* was detected against several Gram-positive and Gram-negative bacteria. Sequencing and *in silico* analysis of NL19 genome revealed the presence of 21 biosynthetic clusters for secondary metabolites, including nonribosomal and ribosomal peptides biosynthetic clusters. Four lanthipeptides clusters were predicted, comprising the precursor peptides, the modifying enzymes (LanB and LanC), and also a bifunctional LanT. This result revealed the hybrid nature of the clusters, comprising characteristics from two distinct classes, which are poorly described in literature. The phylogenetic analysis of their enzymes showed that they clustered within the bacteroidetes clade. Furthermore, hybrid gene clusters were also found in other species of this phylum, revealing that it is a common characteristic in this group. Finally, the analysis of NL19 colonies by MALDI-TOF MS allowed the identification of a 3180 Da mass that corresponds to the predicted mass of a lanthipeptide encoded in one of the clusters. However, this result is not fully conclusive and further experiments are needed to understand the full potential of the compounds encoded in this type of clusters. In conclusion, it was determined that NL19 strain has the potential to produce diverse secondary metabolites, including lanthipeptides that were not functionally characterized so far.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACN	Acetonitrile
antiSMASH	Antibiotics & Secondary Metabolite Analysis SHell
BAGEL3	Bagel automated bacteriocin mining
Bp	Base pairs
CIPRES	CyberInfrastructure for Phylogenetic RESearch
Da	Dalton atomic mass
ddNTPs	Dideoxynucleotides
dH_2O	Distilled water
Dha	2,3-dehydroalanine
Dhb	2,3-dehydrobutyrine
dNTPs	Deoxyribonucleotide triphosphates
emPCR	Emulsion polymerase chain reaction
HT-NGS	High-throughput next-generation sequencing
Lan	Lanthionines
MALDI-TOF	Matrix-assisted laser desorption ionization time of flight
MeLan	Methylanthionines
MS	Mass spectrometry
NGS	Next-generation sequencing
NRPs	Nonribosomal peptides
ORF	Open reading frame
PCR	Polymerase chain reaction
PEGs	Protein encoding genes
PGAAP	Prokaryotic Genome Automatic Annotation Pipeline
RAST	Rapid Annotation using Subsystem Technology
RiPPs	Ribosomally synthesized and post-translationally modified peptides
rRNA	Ribosomal ribonucleic acid
ssDNA	Single-stranded DNA

List of works submitted as part of this dissertation

Santos T., Cruz A., Caetano T., Covas C., Mendo S. (2015) Draft genome sequence of *Pedobacter* sp. strain NL19, a producer of potent antibacterial compounds. Genome Announc. 3, 2, e00184-15.

Chapter I. Introduction

1.1 The foundation of genetics and the advent of DNA sequencing

The term biology first appeared in 1736 by Carl Linnaeus in his book *Bibliotheca botanica*. Linnaeus used this term to refer to botanists that studied the life cycle of plants (1). Since this definition, the meaning of biology has changed due to the improvement of laboratorial techniques and the knowledge of molecular and cellular pathways. A new science that emerged and changed the concept of biology was genetics. Gregor Mendel in the 19th century studied rules that explained the inheritance of biological traits, controlled by the genes (2). This contributed to the emergence and development of genetics, which main aim is to understand genes and their function (2). This new science developed with the studies of several scientists. Until the 1940s the genes were considered proteins but experiments by Avery, MacLeod and McCarthy in 1944, and Hershey and Chase in 1952, lead to the confirmation of DNA as the "transformation principle" and the hereditary material (2, 3).

The attempt to sequence DNA was first reported in 1973 by Gilbert and Maxam, which sequenced 24 base pairs of the *lac* operator using a method known as wandering-spot analysis. But the method was time consuming and laborious. In the mid-1970s, Frederick Sanger developed a faster and more efficient process, known as Sanger sequencing or chain termination method (4, 5). This method is based in annealing a short oligonucleotide complementary to the single-stranded DNA molecules allowing the addition of deoxyribonucleotide triphosphates (dNTPs) and fluorescent marked dideoxynucleotides (ddNTPs), in the new strand synthetized (4, 5). The DNA polymerase enzyme performs this synthesis and when ddNTPs are incorporated the elongation stops, due to lack of the 3'-hydroxyl group needed to form a bond between two nucleotides (4, 5). Stopping of the process generates molecules with different lengths, each ending with different ddNTPs (ddATP, ddCTP, ddGTP and ddTTP) that occupy a position equivalent in the template DNA. This allows the separation and discrimination according to the molecules ddNTPs through electrophoresis, in a polyacrylamide gel or a capillary tube gel system (4, 5). In 1983, K. Mullis revolutionized genetics by inventing the polymerase chain reaction (PCR). The basic aim of PCR is amplifying DNA and obtaining rapidly millions of

copies of this nucleic acid material, using specific reagents and conditions. This invention spurred the rise of DNA sequencing from the first generation sequencing to further generations of automated sequencing (5, 6). These methods are identified as high-throughput next-generation sequencing (HT-NGS) and can generate high-throughput data from whole genomes, faster and with reduced costs (5, 6).

1.2 Next-generation sequencing

The next-generation sequencing (NGS) methods comprise several massive parallel platforms that use different DNA sequencing approaches while sharing some similarities, e.g. the production and amplification of a library to generate an appropriate signal to the sequencing reaction (figure 1) (5-9). The first NGS sequencers developed used the sequencing-by-synthesis with pyrosequencing method and emulsion PCR (Roche) or bridge PCR (Illumina) to perform the DNA amplification (5-9). Currently, the new developed sequencers use real-time sequencing of a single DNA molecule (Oxford Nanopore Technologies) (5-9).

1.2.1 General overview of NGS platforms

The first massive parallel sequencing platform available in the market was the GS20 machine, from 454 Life Science that used pyrosequencing as the sequencing method (figure 1) (5, 10). This method was developed in the 1990s by P. Nyrén and was subsequently optimized in the beginning of the 21th century with the incorporation of the PCR step (5, 6, 8, 10), which allowed a faster and more economic sequencing method to be used in several genomic studies (5, 7, 9, 11). The 454 sequencing uses emulsion PCR (emPCR), to generate millions of DNA copies, previously fragmented and ligated to microbeads through specific synthetic DNAs, known as adapters (figure 1) (5, 9, 10).

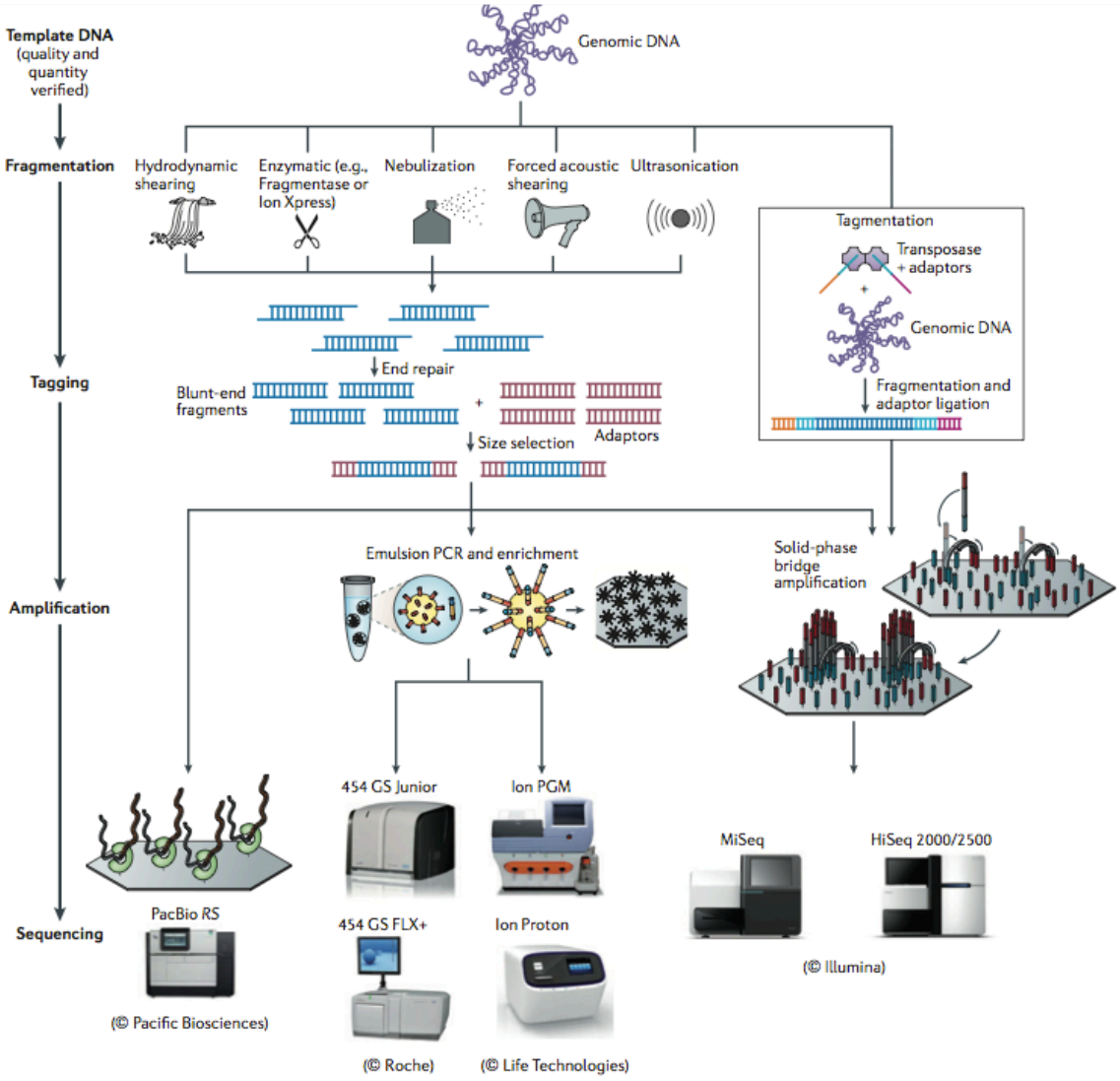


Figure 1| Next-generation sequencing platforms, in *Loman et al*. (2012) (6).

Other platforms using sequencing-by-synthesis method are Ion Torrent and Illumina HiSeq and MiSeq series. Illumina uses a different type of amplification method, the so-called bridge PCR. In this method, the fragments are in contact with a surface of a flow cell and the addition of nucleotides and enzymes to the surface initiates the solid-phase bridge amplification through the bending of the fragments (figure 1) (6, 8, 9).

Some differences between the HT-NGS technologies, such as the error rate and run time, for the high-end sequencers are shown in table 1. In a recent performance test Loman *et al*. (2012) compared the different available bench-top sequencers, Illumina

MiSeq, Roche 454 GS Junior and Ion Torrent PGM (6). MiSeq achieved a higher throughput per run and a lowest error rate but was the slowest instrument (6). Roche 454 GS Junior generated the longest fragments and the better assemblies (6). However, this platform is the most expensive and has the lowest throughput per run (6). Ion Torrent PGM was considered the fastest throughput instrument but produced smaller fragments (6). Furthermore, Ion PGM and 454 GS instruments introduced more sequencing errors in regions of homopolymers, generating assembly errors and ultimately frameshifts in coding regions (6).

Table 1| Comparison of the high-end high-throughput sequencing platforms, adapted from *Loman et al*. (2012) and *Mardis* (2011).

Company	Run time	Read Length (bp)	Error Rate (%)	Purchase Costs (US$)
Roche 454 GS FLX+	8 - 23 hours	500 - 800	1	100.000 - 500.000
Illumina HiSeq	1 - 11 days	2 × 100/150	>0.1	125.000 - 750.000
Ion Torrent	2 hours - 8 days	100 - 200	>0.06 – 0.01	80.000 - 350.000
PacBio *RS*	20 minutes	3.000 - 15.000	15	750.000
Oxford Nanopore	NA	NA	NA	NA

More recently, Oxford Nanopore Technologies developed a new sequencing technology, strand sequencing. This technique employs a protein nanopore consisting of a protein with a hollow tube in its core, which is inserted in a membrane created by a synthetic polymer with high electronic resistance (6, 12, 13). This feature enables the creation of a constant potential that results in the one-sided movement of the ssDNA, one base at the time, causing the potential disruption and allowing the sequencing of the DNA sequence (6, 12, 13). Currently, there is limited information about his sequencing performance since this technology is not commercially available yet (6, 12).

1.2.2 Ion Torrent PGM platform

In this study, the *Pedobacter* sp. NL19 genome was sequenced with Ion Torrent PGM (Life Technologies), a platform similar to 454. The main difference relies on the detection of hydrogen ions, released during the base incorporation (figure 2) (6, 8, 9).

The Ion PGM sequencing technology starts with separation and fragmentation of the double stranded genomic DNA, followed by the ligation of adapters, onto the ends of the ssDNA obtained (5, 9, 10). This ssDNAs are placed in contact with microbeads containing oligonucleotides complementary to the adapters, in a water-in-oil emulsion, allowing the ligation of each fragment to one microbead (5-7, 9). Subsequently, the emPCR produces millions of copies of the initial fragment. Then, the beads are loaded onto a special semiconductor chip containing millions of wells, to perform the sequencing-by-synthesis (6, 14). In this step, a new strand complementary to the ssDNA fragment is created, and in each sequencing cycle occurs the addition of one ddNTP (5-7, 9). This allows the determination of the ddNTP position at each cycle trough the release of hydrogen ions, which changes the pH of the solution allowing the detection by the ion sensor, thus converting chemical into digital information (5, 6, 8, 10).

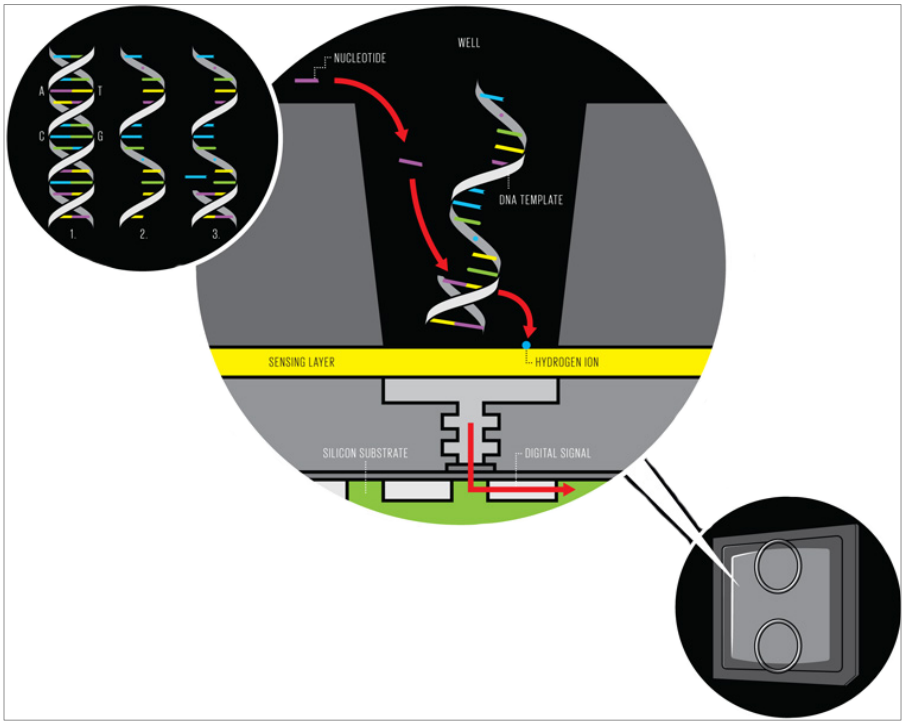


Figure 2| Ion semiconductor sequencing technology, in *Strickland* (2013) (15).

The increasing use of high-throughput sequencing technologies produced in the last decades great amounts of biological data, for instance microbial genome sequences (16). The data is deposited in public databases, e.g. National Center for Biotechnology Information (NCBI), which boosted the rising of bioinformatics field, a combination of biology and informatics. This area combines several computational programs with numerous functions, such as Clustal W/X (17) to create multiple sequence alignments or antiSMASH (18) for secondary metabolites mining, which retrieve relevant biological information from the data obtained by NGS. In fact, the generalization of whole-genome sequencing uncovered the underestimated potential of secondary metabolites clusters present in bacteria (19-21), which potentially exceeds the number of secondary metabolites already described (19).

1.3 Bioactive microbial metabolites

The use of natural products from plants or fungi is well documented since the Mesopotamian civilization (2600 B.C.) that used oils extracted from cypress and myrrh, which are still applied to treat inflammations, for example (22-24). This documentation is also observed in others civilizations, namely Egyptian with *The Ebers Papyrus*, Chinese with *The Chinese Materia Medica* and Persian with *Canon Medicinae* (22-24).

These natural products are compounds not essential for the organism growth, development or reproduction and therefore referred to as secondary metabolites. They most probably have a social and ecological role for the producers and are recognized as natural sources of potential drugs (22, 25, 26). These metabolites can be produce through or for adaptation to the environment pressure or also due to nutritional stress and/or presence of chemical, originating unique adaptations that can result in unique natural products (22, 26, 27).

The first active compound used pharmacologically was morphine, which is produced by *Papaver somniferum*, and discovered in the early 1800s by Friedrich Serturner (28). Other plant-derived compounds are salicin, from *Salix alba*, that originates

the anti-inflammatory agent, aspirin, and codeine, isolated from *Papaver somniferum*, and used extensively in human health (22, 25). However, perhaps the most important natural product discovered so far was penicillin, an antibiotic produced by the fungus *Penicillium notatum*. It was described in 1929 by Alexander Fleming and assured him the Nobel Prize award of Physiology and Medicine in 1945 (22, 29). This discovery represented the dawn of increased interest in new microbial antibiotics, leading to the identification of for instance, vancomycin, from *Amycolatopsis orientalis*, by Edmund Kornfeld in 1953 and erythromycin, from *Saccharopolyspora erythraea*, by Eli Lilly (22, 30). Despite this rush to discover new antibiotics through screenings of biological extracts, the pharmaceutical industries have changed their encouragement for this search towards high-throughput screening (HTS) during the last decades of the 20th century (20, 22, 31, 32). This technology has been combined with combinatorial chemistry to create compound libraries, facilitating the molecular target based drug approach, obtaining efficient "hits" (22, 31, 32). The HTS of natural products has several obstacles, such as the environmental conditions/variations and the need of a high number of specimens leading to loss of source and reproducibility, specially with marine organisms and higher plants (28, 33). Nevertheless, the traditional method is still used to discover microbial natural products, combining steps of collection and cultivation of strains with extraction and isolation of the compound of interest. However, the rediscovery rate using this method is higher (19). The combination of this method with an initial *in silico* analysis of the available microbial genomes and their biosynthetic pathways accelerate the efficiency in discovering new compounds, mitigating the rediscovery rate (21, 22).

Peptides with antimicrobial activity against bacteria and other microorganisms are an important class of natural products with wide occurrence in nature (34-36). In the last decades of the 20th century the numbers of antimicrobial peptides described are increasing (34). These peptides can be of two distinctive classes: the nonribosomal peptides (NRPs) and ribosomally synthesized and post-translationally modified peptides (RiPPs) (37).

1.3.1 Nonribosomal peptides

Nonribosomal peptides (NRPs) are synthesized by multifunctional enzymes designated nonribosomal peptide synthetases (NRPSs) (38-40). The NRPSs are organized in modules that are responsible for the recognition and assembly of the amino acids that constitute the peptide chain (38-40). By doing so, these enzymes are used as template and biosynthetic machinery, simultaneously. Each module is responsible for several chemical reactions steps and can be structurally subdivided in three main catalytic domains: the adenylation (A), the peptidyl carrier protein (PCP) and the condensation (C) domain (figure 5) (38-40). The A domain is responsible for recognizing and binding the correct amino acid, which is activated as an amino acyl-adenylate intermediate (38, 40). Thereafter, the PCP domain accepts the activated amino acid, keeping it bound to the enzymatic machinery, and transports it to the next center (38, 40). Finally, the C domain catalyzes the formation of peptide bond between the two activated amino acids. The final module of the NRPS contains the thioesterase (TE) domain that catalyzes the peptide release (38-40). The NRPs frequently have D-amino acids or methylated residues in their constitution. Also, they are frequently cyclic peptides. These properties are introduced by specific domains, which include the epimerization (E), the *N*-methylation (NMT) and the cyclization (Cy) domains, respectively (38, 40). Some nonribosomal peptides display a broad-spectrum and potent activity as antibiotics, immunosuppressants or antitumor, such as penicillin or tyrocidine (figure 3), cyclosporine A and bleomycin, respectively (41, 42).

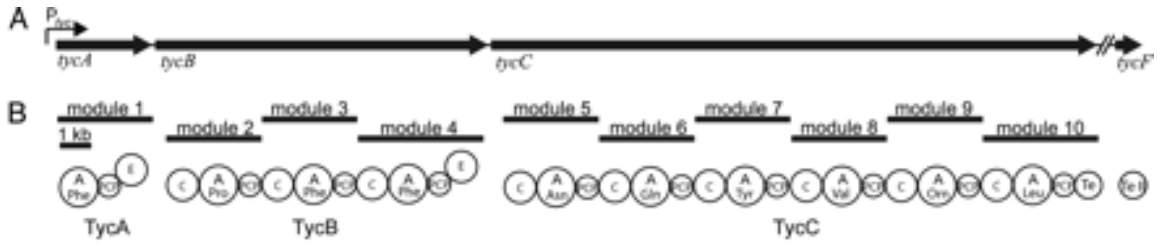


Figure 3| Example of a NRPs biosynthetic system (tyrocidine) encoded in three ORFs (tycABC) (A). Representation of the modules and domains that constitute the NRPs TycA, TycB and TycC (B), in *Hahn et al.* (2004) (42).

1.3.2 Ribosomally and post-translationally modified peptides: lanthipeptides

Ribosomally synthesized and post-translationally modified peptides or RiPPs, are peptides with diversified structures and widely produced in nature (43). Lanthipeptides are the most extensively studied class and are characterized by the presence of sulfur-to-carbon thioether cross-links designated as lanthionines (Lan) and methylanthionines (MeLan), that result from post-translational modifications (44). These modifications occur only in the mature form of the peptides and are formed by a specific biosynthetic machinery, comprising two phases (44). Initially, the Ser and Thr residues present in the core region of the precursor peptide are dehydrated to form 2,3-dehydroalanine (Dha) and (Z)-2,3-dehydrobutyrine (Dhb), respectively (44). Next, a reaction between Cys (which are exclusively found in the core region) and Dha and Dhb occurs to form the Lan and MeLan thioether rings, respectively (figure 4) (44). This is followed by the proteolytic cleavage of the N-terminal leader sequence to form the mature and active lanthipeptide (figure 4) (44).

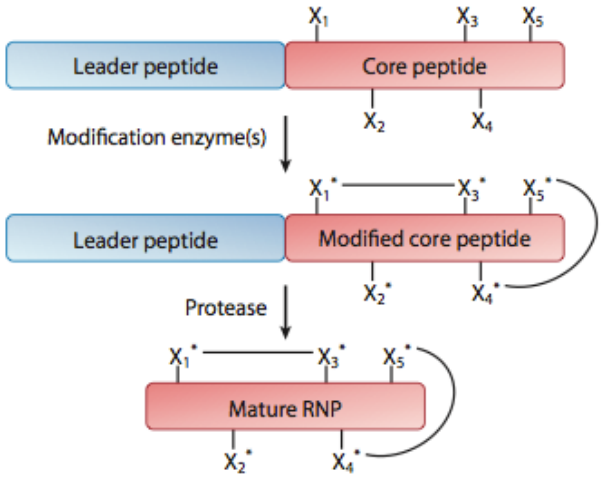


Figure 4| Representation of the general biosynthesis of lanthipeptides, where X_n represents a modified residue, in *Knerr et al.* (2012) (45).

Lanthipeptides are classified in four distinct classes according to their biosynthetic machinery, especially with regard to enzyme(s) responsible for the formation of Lan and

MeLan amino acids (44):

- In class I lanthipeptides, the biosynthesis of Lan and MeLan involves two different enzymes (figure 5): a dehydratase, LanB and a cyclase, LanC. LanB is responsible for the formation of the Dha and Dhb amino acids, through the enzymatic dehydration of Ser or Thr residues, respectively (46, 47). This enzyme also performs the addition of Cys thiols to these unsaturated amino acids. Then, LanC activates the Cys thiols for nucleophilic attack, leading to the formation of Lan or MeLan (44). Following these reactions, a protease LanP recognizes a conserved motif (TR, AR, PR or PQ) in the leader peptide and cleaves after it, producing an active peptide that is exported from the cell by the ABC transporter LanT (47). It has been described that the functionality of LanC enzyme is dependent on a zinc-binding motif (Cys-Cys-His). In the absence of this motif, the enzyme is incapable of achieving an accurate cyclization of the dehydroamino acids and Cys (46, 47).

- In class II lanthipeptides, a single lanthipeptide synthetase, the LanM, is responsible for the dehydration and cyclization reactions (figure 5) (44, 47). This enzyme contains a C-terminal cyclase domain with homology to LanC proteins and a N-terminal dehydratase domain that has no homology to LanB, but is responsible for the dehydration of Ser and Thr residues (44, 46, 47). Only one bifunctional enzyme, LanT, performs the cleavage and transport of the peptides. The N-terminal region of LanT is responsible for the recognition of a conserved motif (GG, GA or GS) and concomitant cleavage of the leader peptide and the C-terminal for the transport of the mature peptide (44, 45). Class II lanthipeptides include a special group of peptides such as lacticin 3147 and haloduracin, that are composed by two peptides: the α- and β-peptides (45). Each of these peptides is modified by different LanM enzymes, but their leader peptides and transport are catalysed by the same LanT protein. Their antibacterial activity results from the synergistic interaction of both peptides. Therefore, when the peptides act separately, it always result in the reduction or abolishment of their activity (45, 46).

- In class III lanthipeptides, the synthesis is achieved with the LanKC enzyme (45, 48). This enzyme contains a N-terminal lyase and central kinase domain, similar to serine/threonine protein kinases (44, 45). The C-terminus contains a putative cyclase

without the zinc-binding motif and with low similarity to LanC and LanM enzymes. (44-46). LanKC act as kinase-cyclase performing two-steps. First the phosphorylation of the precursor peptide LanA occurs, which is followed by the dehydration of Ser residues performed by the LanKC N-terminus (48). Then, its C-terminus performs the cyclization of cysteine residues in the C-terminus of LanA, despite the missing zinc-binding motif (45, 48). For the known class III lanthipeptides, e.g. labyrinthopeptin A2, the gene cluster lacked a protease enzyme and a recognition conserved motif for the cleavage of the leader peptide (45, 48).

- Class IV lanthipeptides are synthetized by a LanL enzyme that shares a similar lyase N-terminus and central kinase domains with LanKC (35, 45). Nonetheless, the LanL C-terminus domain contains the conserved zinc-binding motif, homologous to LanC and LanM (44-46). Furthermore, no protease was identified in the gene cluster, although the presence of a LanT-like enzyme was detected (45). Venezuelin, the first class IV lanthipeptide described, is produced by a similar pathway as class III lanthipeptides, with phosphorylation and dehydration of LanA performed by the N-terminus domains. Thereafter, the cyclization is achieved with the cyclase domain present in the C-terminus of LanL (44, 45).

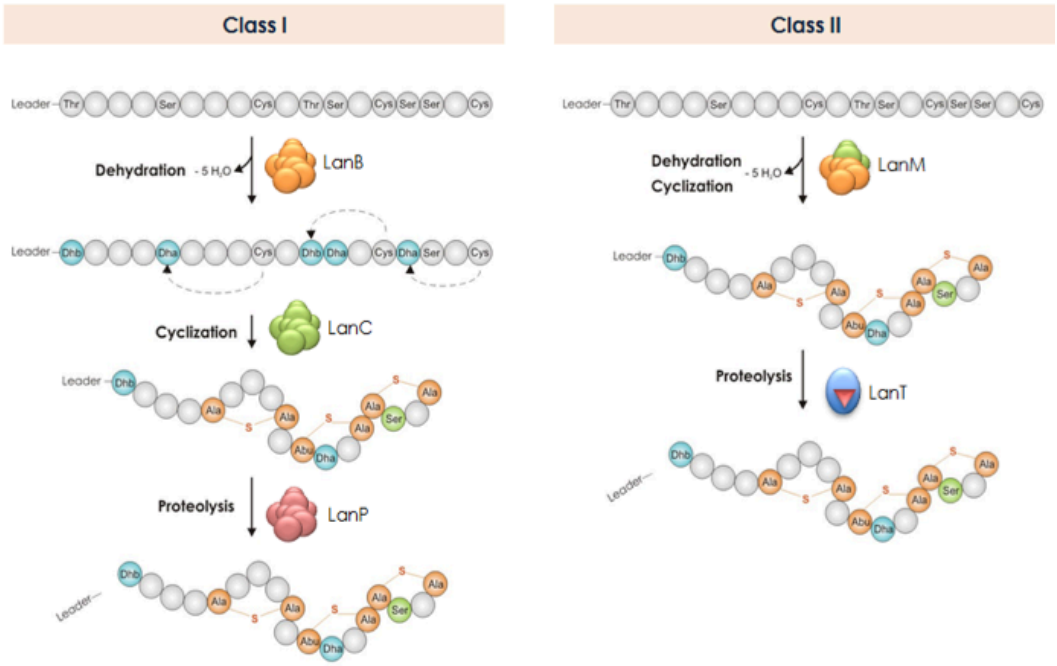


Figure 5| Schematic representation of the general biosynthesis of class I and class II lanthipeptides, in *Caetano* (2011) (49).

Lanthipeptides that exhibit antimicrobial activity are termed lantibiotics (35, 43). Generally, class I and class II lanthipeptides are lantibiotics. Class I nisin, mainly used as food preservative, is the most emblematic of these peptides and is the only commercially available lantibiotic (35, 43). Additionally, class I gallidermin and class II mersacidin exhibit potential activity as prophylactic agents, which prevent infections in medical devices (50). Other lantibiotics, namely class II duramycin and the semisynthetic actagardine exhibited great activity, in cystic fibrosis treatment and against infections caused by *Clostridium difficile*, respectively (35, 50). Interestingly, class II prochlorosins, discovered by genome mining of *Prochlorococcus* MIT9313, are the exception among class I and class II because they do not exhibit antimicrobial activity (50). As above referred, class III and IV lanthipeptides do not have antibacterial activity. However, they have other applications, namely antiviral activity *in vitro* exhibited by the class III labyrinthopetin and as biosurfactants, as is the case of class III SapT and SapB (50).

Generally, the essential genes involved in the biosynthesis, modification (LanB/LanC and LanM) and in the cleavage/export (LanP/LanT) of the precursor peptide (LanA) are clustered together, constituting the biosynthetic gene clusters (figure 6) (50). Additionally, regulatory (LanKR) and immunity systems (LanIFEG) genes can also be present in these clusters (50, 51). LanK, a membrane-bound histidine kinase and LanR, a transcriptional response regulator, form a two-component sensory system, regulating the lanthipeptide biosynthesis (50, 52). This system is activated after the induction of the autophosphorylation of a histidine residue in LanK, through extracellular signals or the lanthipeptide itself (52). The phosphate group released acts in the receiver domain of LanR that mediates the transcriptional activation of the lanthipeptide biosynthesis (50, 51).

Immunity systems, comprising the *lanIFEG* genes, are relevant in bacteria producing lantibiotics, by protecting themselves against their own produced compounds (49, 53). This system contain individual immunity proteins (LanI) and/or ATP binding cassette (ABC) transporters (LanFE(G)), different from LanT transporters, that can act coordinately or alone (49, 53). LanI proteins can be present in the bacterial supernatant as a free protein or associated with the membrane (49). The LanFE(G) system comprises

two different domains of ABC transporters (49). The association of two LanF ATPase subunits with two LanE or LanG integral membrane proteins export the lantibiotic from the cytoplasm (49). In nisin, the stimulation of the immunity system is achieved with the activation of the regulatory system nisKR (53).

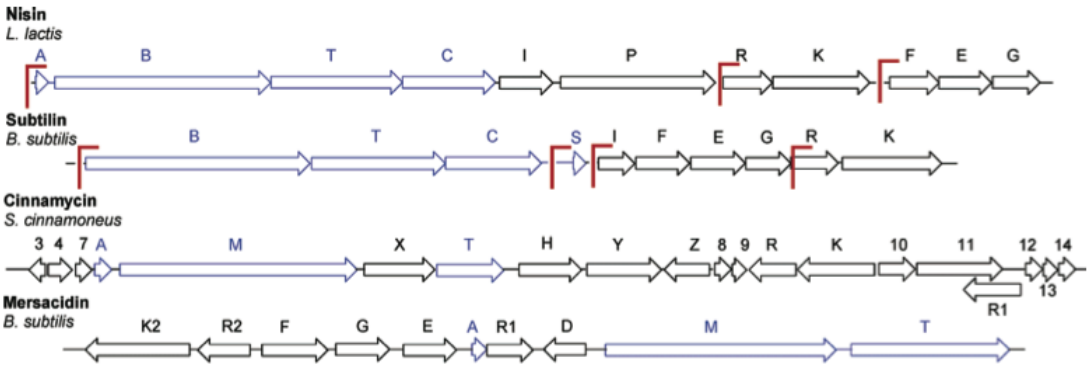


Figure 6| Schematic representation of four biosynthetic gene clusters of class I and class II lanthipeptides, in *Chatterjee* (2005) (54). In blue are genes essential for the biosynthesis of the mature peptide (A), *lanB* genes (B), and *lanC* genes (C) of class I members and *lanM* genes (M) for class II. Genes of class I proteases *lanP* (P) and transporter genes *lanT* (T) of both classes are also shown. Additionally, immunity (*lanIFEG*) and regulatory (*lanKR*) are also represented.

1.4 *Pedobacter* sp. NL19 strain

The bacterial strain NL19 was isolated in November of 2013, from a sludge water sample collected at a uranium mine located in Quinta do Bispo – Viseu (Portugal) that has been deactivated since 1991. Despite this, the water samples at the site still have high concentrations of both radionuclides and metals. The NL19 strain was selected for further studies due to its *in vitro* antibacterial activity against relevant Gram-positive and Gram-negative bacteria, including *Aeromonas hydrophila* ATCC 7966, *Listeria monocytogenes* 71 and *Klebsiella pneumoniae* ATCC 700603 (ESBL-producer). These bacteria are important infectious pathogens in the food and health sectors causing, for example, bacteremia to individuals (55). The 16S rRNA gene of NL19 was sequenced (GenBank acession KJ579161) and it was possible to affiliate this bacterium within the *Pedobacter* genus. This genus was proposed in 1998 and belongs to *Sphingobacteriaceae* family (56). Species from this

genus are characterized as Gram-negative rods, obligatory aerobic and heparinase producers (56). Since the proposal of this genus, 50 species have been already described in the *List of Prokaryotic names with Standing in Nomenclature* (accessed 24-06-2015) (57).

Pedobacter genus displays some interesting features with some members possessing antimicrobial activity. For instance, the production of bioactive compounds was observed in *Pedobacter* sp. SANK 72003, isolated from a rhizosphere soil sample in Japan (58). These compounds, designated pedopeptins, possess antibacterial activity against Gram-positive bacteria (58). Similarly, *Pedobacter cryoconitis* strain BG5, isolated in the Antarctic, was able to produce antibacterials, which were excreted to the supernatant and with activity against Gram-positive and Gram-negative bacteria (59). An interesting finding was observed for *Pedobacter* sp. V48, which shows antifungal activity if associated with *Pseudomonas fluorecens* (60). Additionally, by genome mining putative biosynthetic genes for lanthipeptides where found in *Pedobacter heparinus* DSM 2366, however, so far, no antibacterial activity was attributed to this strain (61).

1.5 Aim and objectives of this thesis

Pedobacter sp. NL19 displays an interesting antibacterial activity against clinically relevant Gram-positive and Gram-negative bacteria. Considering that bacterial resistance to antibiotics is a global problem and that new solutions are urgently required, further insights on the possible biotechnological potential of this strain were investigated. To that end we proceeded to sequencing the genome of NL19 strain. The genome was sequenced by Stabvida company.

Accordingly, the aim of this thesis was to identify and characterize putative biosynthetic gene clusters present in the genome of NL19 strain, mainly focusing on the biosynthesis of lanthipeptides. To achieve this, the following objectives were outlined:

1. Annotation of the draft genome;
2. Identification of gene clusters encoding the biosynthesis of secondary metabolites by *in silico* analysis;
3. Identification of the characteristic lanthipeptide encoding gene clusters;
4. Prediction of the molecular masses of the mature peptides that can be produced by lanthipeptides gene clusters identified in the genome of NL19;
5. MALDI-TOF MS detection of the lanthipeptides products corresponding to the predicted molecular masses.

The identification of some putative biosynthetic clusters will be valuable for further studies involving this strain, especially involving the assignment of antibacterial activity to some classes of compounds. Some of those may belong to the lanthipeptide family.

Chapter II. Materials and Methods

2.1 Genomic DNA extraction

The genomic DNA of *Pedobacter* sp. NL19 was extracted to perform the NGS procedure. The extraction was achieved using the DNeasy Blood & Tissue kit (Qiagen), according to the manufacturer's instructions and are below described.

GENOMIC DNA EXTRACTION
1. Prepare a culture of NL19 strain in 15 mL of Tryptic Soy Broth (TSB) and incubate for 24h at 26ºC, 180 rpm.
2. Harvest the cells in a microcentrifuge tube (eppendorf) by centrifuging for 10 min at 5,000 x *g*. Discard supernatant.
3. Resuspend pellet in 180 µl Buffer ATL.
4. Add 20 µl proteinase K and mix thoroughly by vortexing. Incubate at 56°C in a thermomixer, for 5 minutes.
5. Vortex for 15 seconds and add 200 µl Buffer AL to the microcentrifuge tube, and mix by vortexing.
6. Add 200 µl absolute ethanol, and mix by vortexing.
7. Pipet the mixture from the microcentrifuge tube into the DNeasy Mini spin column. Centrifuge at 6,000 x *g* for 1 minute.
8. Discard flow-through and collection tube and place the DNeasy Mini spin column in a new collection tube.
9. Add 500 µl Buffer AW1, and centrifuge for 1 minute at 6,000 x *g*.
10. Discard flow-through and collection tube and place the DNeasy Mini spin column in a new collection tube.
11. Add 500 µl Buffer AW2, and centrifuge for 3 minutes at 20,000 x *g*.
12. Place the DNeasy Mini spin column in a clean 1.5 ml microcentrifuge tube.
13. Pipet 200 µl Buffer AE directly onto the DNeasy membrane.
14. Incubate at room temperature for 1 minute.
15. Centrifuge for 1 minute at 6,000 x *g*.

2.2 Genomic DNA sequencing and assembly

The genome of *Pedobacter* sp. NL19 was sequenced in Stabvida company (Portugal) using the Ion Torrent™ Personal Genome Machine® (PGM) System, from Life Technologies, with an Ion 316 Chip v2 and a sequencing kit for read lengths of 400 bp per run.

The sequence reads obtained were trimmed, to enhance the data quality. The high quality reads were subsequently assembled using the CLC Genomics Workbench 7.0.3 program (CLC bio – Qiagen) using the detailed in table 2.

Word size was determined automatically by the same software, using the Bruijn graphs, depending on the total amount of the input data (62). Therefore, more data represents a longer word length, until a maximum word size of 64 bp (table 3).

The bubble size is defined as a bifurcation in the consensus assembly sequence into two reads, and the subsequent merging into one sequence again (62). It was determined automatically using CLC Genomics Workbench 7.0.3 program, where the bubble size value is set to 50 for reads shorter than 110 bp, and for longer reads, the size is set for the average read length (62). The assembly performed in this study used a bubble size representing the average read length of 291 bp (table 2).

Table 2| Assembly parameters.

Assembly parameters	
Word size	22
Bubble size	291
Minimum contig length [bp]	500
Mismatch cost	2
Insertion cost	3
Deletion cost	3
Length fraction	0.5
Similarity fraction	0.8

Table 3| Determination of word size by CLC bio assembler.

CLC bio - word size	
Word size 12	0 – 30,000 [bp]
Word size 13	30,001 – 90,002 [bp]
Word size 14	90,003 – 270,008[bp]
Word size 15	207,009 – 810,026 [bp]
Word size 16	810,027 – 2,430,080 [bp]
[...]	[...]
Word size 22	590,509,683 – 1,771,529,048 [bp]
Word size $^{(n+1)}$ ≤64	Amount of data $^{(\times 3)}$

2.3 Genome annotation

The annotation of *Pedobacter* sp. NL19 genomic contigs was performed with RAST 2.0 (Rapid Annotation using Subsystem Technology), an automated service allowing the annotation of complete or nearly complete bacterial and archaeal genomes (63). For the NL19 genome the classic RAST annotation scheme, RAST gene caller and the release 70 of FIGfams were used. Additionally, several other options such as, automatically fix errors and frameshifts and backfill gaps were selected to improve the annotation. The annotated contigs and protein encoding genes (PEGs) were downloaded to perform further genome analysis. Additional examination was performed with tRNAscan-SE-1.23 (64), to detect tRNA genes, and RNAmmer 1.2 server (65), to detect rRNA genes.

2.4 NCBI submission

The *Pedobacter* sp. NL19 genome was submitted to the National Center for Biotechnology Information (NCBI) database. In this submission, the genome was automatically annotated through the NCBI PGAAP (Prokaryotic Genome Automatic

Annotation Pipeline) and this annotation is publicly available at the NCBI database (accession JXRA01000000). All the accession numbers in the different databases for *Pedobacter* sp. NL19 are indicated in table 4. Furthermore, the genome availability and a short description of *Pedobacter* sp. NL19 potential was published in the Genome Announcements journal (66).

Table 4| Accession numbers for *Pedobacter* sp. NL19.

Accession numbers	
GenBank	JXRA00000000
BioProject	PRJNA273375
BioSample	SAMN03291000
SRA[1]	SRP052743[2]

[1] Sequence Read Archive: database of the raw reads after sequencing
[2] Release at: January of 2016

2.5 Gene Ontology Consortium (GO)

The PEGs obtained from RAST were analysed with the data-mining program Blast2GO (B2G). This program uses the Gene Ontology (GO) Consortium classification framework to perform the functional categorization of the input gene sequences in three independent ontologies: biological process, molecular function and cellular component classes (67, 68). The gene ontologies are associated to biological functions based on similarity, even for unknown sequences, allowing the organization in different functional classes (67, 68).

2.6 Identification of potential biosynthetic clusters for secondary metabolites

The identification of clusters encoding the biosynthesis of secondary metabolites was performed by analysing the *Pedobacter* sp. NL19 genomic contigs with two different

web-based platforms: antibiotics & Secondary Metabolite Analysis SHell (antiSMASH 3.0) (18) and Bagel automated bacteriocin mining (BAGEL3) (69). AntiSMASH identifies clusters by comparing each gene product to a curated database of known and key biosynthetic enzymes of secondary metabolite classes (18). Depending on the classification of the detected secondary metabolite gene cluster, additional algorithms predict the specificities and chemical structure of the gene cluster (18). BAGEL uses a different approach. It searches for ORF encoding protein domains that meet a set of pre-defined rules (69). Subsequently, an additional calling of small ORFs allows the identification of modification enzymes and the results are then searched against the BAGEL database (69) through BLAST (basic local alignment search tool) (70).

2.7 Phylogenetic analysis of the putative biosynthetic gene clusters

The amino acid sequences used for the phylogenetic analysis of the lanthipeptide synthetases LanB and LanC were obtained from the NCBI sequence database, using the lanthipeptide enzymes identified in the *Pedobacter* sp. NL19 genome as source. The sequences selected for the analysis had an expected value $< 10^{-8}$ and query coverage > 75%. Furthermore, the sequences of LanB and LanC enzymes belonging to other bacterial groups that were used by Zhang *et al.* (2012) (44) were also included in the analysis. For the phylogenetic analysis of LanT transporters were selected species with lanthipeptide synthetases clustered closely to the NL19 enzymes, enzymes from known class I and class II transporters. The accession number of the ORFs used and their respective source organisms are listed in tables 11-15 (appendix 1).

The sequences were aligned using ClustalW algorithm (71) within the CyberInfrastructure for Phylogenetic RESearch (CIPRES) portal (version 3.3) (72). The phylogenetic analysis of the selected lanthipeptide synthetases sequences was performed using MrBayes v3.2.3 (73) within the CIPRES portal. The analysis consisted of two runs of eight chains each (seven heated and one cold), with variable number of generations until a convergent standard deviation of split frequencies < 0.01 was reached. The parameters

were sampled every 1,000 generations and the final posterior probabilities considered were averaged over the final 75% trees generated due to the 25% burn in parameter. Also, a mixed amino acid model with a proportion of sites designated invariant (+I), and rate variation among sites modeled after a gamma distribution (+G) was used, with all variable parameters predicted by the program. The phylograms were prepared with FigTree v1.4.2 software (74).

The maximum likelihood models prediction was achieved with ProtTest 3 (75) to determine the best substitution model, and PhyML v3.0 (76) to estimate the phylogenies with branch support determined by SH-like approximate likelihood-ratio test (aLRT) statistics. The best models selected were the LG model with gamma distributed with invariant sites and equilibrium frequencies (G+I+F) for LanBs and LanTs, WAG model with gamma distributed and equilibrium frequencies sites (G+F) for LanCs.

2.8 Confirmation of the nucleotide sequences of *lanA* and *lanB* genes from *Pedobacter* sp. NL19

The nucleotide constitution of the structural genes (*pedA*) present in the lanthipeptide clusters that were identified in the NL19 genome was confirmed using Sanger sequencing reaction. This procedure was performed to assure a correct prediction of the molecular masses for their respective mature peptides. The LanB proteins of the clusters *ped8* and *ped17* were by two different ORFs. Part of these genes were also re-sequenced in order to understand if this prediction is correct or if it resulted from sequence errors introduced by NGS. The amplification of the genes was carried out with the proofreading *Pfu* DNA polymerase (Thermo Scientific) using the conditions and the amplification parameters described in table 5 and table 6. The primers used were constructed using Primer3 (77) and are listed in table 7. After amplification, the PCR products were sequenced at Stabvida company (Portugal).

Table 5| PCR reaction used for amplification of *pedA* and *pedB* genes.

Reagents	Final concentration	Volume (µL)
10X *Pfu* Buffer with $MgSO_4$	1X	5
10 mM dNTPs mix	0.2 mM	1
10 µM *Primer* FW	0.3 µM	1.5
10 µM *Primer* RV	0.3 µM	1.5
2.5 U/µL *Pfu* DNA Polymerase	1.25 U/µL	0.5
DNA total	1.7 ng	1
dH_2O	-	up to 50 µL

Table 6| Conditions used for the amplification of *pedA* and *pedB* genes.

	Temperature	**Time**	
Initial denaturation	95ºC	120 seconds	
Denaturation	95ºC	30 seconds	30 cycles
Annealing	50ºC - 60ºC	30 seconds	
Extension	72ºC	90 seconds	
Final extension	72ºC	300 seconds	

Table 7| Primers used and respective product size for the amplification of *pedA* and *pedB* genes.

Primers				
Gene	Primer name	Primer	$T_{annealing}$	Product size (bp)
pedA8	pedA8 Fw	5´-ccaggagttccagtggtcat-3'	52ºC	730
	pedA8 Rv	5´-tttcttcagctggctcctgt-3'		
pedA14	pedA14 Fw	5´-cagtccattggtttctggttg-3'	50ºC	812
	pedA14 Rv	5´-cctgcttctccgtcaaaaag-3'		
pedA15	pedA15 Fw	5´-tttccacgattccattacgg-3'	52ºC	866
	pedA15 Rv	5´-tcaggaccaattgagcgaat-3'		
pedA17	pedA17 Fw	5´-ttttggcattgacacaacct-3'	54ºC	752
	pedA17 Rv	5´-aataatacgtcccgctgcaa-3'		
pedB8	pedB8 Fw	5´-tcagggcttgcctaccttta-3'	60ºC	1223
	pedB8 Rv	5´-actgcttccaaaagctggtt-3'		
pedB17	pedB17 Fw	5´-aaagacgggttaagcgatga-3'	54ºC	1301
	pedB17 Rv	5´-aaaagcaaaccgatctgtgg-3'		

2.9 Mass spectrometry of *Pedobacter* sp. NL19 colonies and supernatant

The analysis of the PedA precursor peptides allows the prediction of a maximum molecular mass for all the mature peptides, assuming that the proteolytic removal of the leader sequence is performed after the GG-motif and considering the number of Cys residues (appendix 6, table 17). However, other lower molecular masses corresponding to different degrees of dehydration (Ser and Thr amino acids) can also be considered (appendix 6, table 17). Therefore, the production of these peptides can be further investigated using matrix-assisted laser desorption ionization time of flight mass spectrometry (MALDI-TOF MS). The method is based on the ionization of crystallized samples by short laser pulses, then the detector measure the time of flight of the ions by a mass-to-charge ratio (78, 79). Different conditions were used, including the analysis: i) of colonies, ii) of the extracts resulting from washes of colonies with two types of solvents and iii) of culture supernatant. The detailed preparation of the samples are described below in detail. The mass spectra was obtained using negative-ion reflector mode with gamma range from 1,000 to 4,000 in Applied Biosciences MDS SCIEX (Life Technologies). The mass spectra was visualized using Data Explorer® Software (Applied Biosciences).

I) PREPARATION OF NL19 COLONIES
1. Prepare a culture of NL19 strain in 15 mL of Tryptic Soy Broth (TSB) and incubate for 24h at 26ºC, 180 rpm.
2. Distribute 5 µL of this culture in the center of a Tryptic Soy Agar (TSA) plates and incubate for the appropriate time (24h, 48h or 1 week) at 26ºC.
3. Use a 1 µL loop to transfer colonies to two spots of the MALDI-TOF plate.
4. Cover with 0.4 µL of the sinapinic acid matrix one of the spots and the other with 0.4 µL of the α-cyano-4-hydroxycinnamic acid matrix.
5. Allow the plate to dry before analysis.

II) PREPARATION OF NL19 COLONIES WASHES FOR MS ANALYSIS

PROTOCOL 1

1. Prepare a culture of NL19 strain in 15 mL of Tryptic Soy Broth (TSB) and incubate for 24h at 26ºC, 180 rpm.
2. Distribute 5 µL of this culture in the center of a Tryptic Soy Agar (TSA) plate and incubate for the appropriate time (24h, 48h or 1 week) at 26ºC.
3. Use a 1 µL loop to transfer colonies to 100 µL of 50% ACN:dH_2O placed in a microcentrifuge tube. Repeat procedure twice for the same solution. Vortex the solution.
4. Centrifuge at 2,290 x *g* for 5 min and collect the supernatant to clean microcentrifuge tube.
5. Place 0.4 µL of this supernatant in two spots of the MALDI-TOF plate.
6. Cover with 0.4 µL of the sinapinic acid matrix one of the spots and the other with 0.4 µL of the α-cyano-4-hydroxycinnamic acid matrix.
7. Allow the plate to dry before analysis.

PROTOCOL 2

1. Prepare a culture of NL19 strain in 15 mL of Tryptic Soy Broth (TSB) and incubate for 24h at 26ºC, 180 rpm.
2. Distribute 5 µL of this culture in the center of a Tryptic Soy Agar (TSA) plate and incubate for the appropriate time (24h, 48h or 1 week) at 26ºC.
3. Use a 1 µL loop to transfer colonies to 500 µL in a microcentrifuge tube. Repeat procedure twice for the same solution. Vortex the solution.
4. Centrifuge for 5 minutes at 2,290 *x g* and remove the supernatant. Repeat the centrifugation for 5 minutes at 9,168 x *g* and remove the supernatant.
5. Dissolve the bacterial pellet in 200 µL of a 1:1 solution of acetonitrile (100%) formic acid (20%), followed by vortex.
7. Centrifuge for 4 minutes at 9,168 x *g* and collect the supernatant to a clean microcentrifuge tube.
8. Place 0.4 µL of this supernatant in two spots of the MALDI-TOF plate.
9. Cover with 0.4 µL of the sinapinic acid matrix one of the spots and the other with 0.4 µL of the α-cyano-4-hydroxycinnamic acid matrix.
10. Allow the plate to dry before analysis.

III) PREPARATION OF NL19 SUPERNATANT FOR MS ANALYSIS

1. Prepare a culture of NL19 strain in 15 mL of Tryptic Soy Broth (TSB) and incubate for 24h at 26ºC, 180 rpm.

2. Measure the optical density of the culture and inoculate three erlenmeyers flasks containing 25 mL of TSB (25%), obtaining a final cell density of OD=0.02.

3. Grow the cultures for 7 days at 26ºC, 180 rpm.

4. Centrifuge the liquid culture for 10 minutes at 8,875 x *g*.

5. Collect the supernatant and store it at -80ºC in 13mL falcon tubes (each containing 4 mL of supernatant).

5. Freeze-dry the supernatant through lyophilisation.

Chapter III. Results and Discussion

3.1 Genomic DNA sequencing statistics

The Ion Torrent sequencing generated 3,722,256 reads, with approximately 128× coverage, and a mean read length of 293 bp (table 8). The assembly of the high quality reads, obtained using the CLC Genomics Workbench, yielded a genome size of 5,988,703 bp and the assembly of 201 contigs. The contigs range from 499 to 180,550 bp in length, and have a *N*50[3] of 57,428 (table 9).

Table 8| Sequencing run statistics.

Sequencing statistics	
Total number of bases [Mbp]	1,090,77
Number of Q20 bases [Mbp]	996,24
Mean read length [bp]	293
Final library reads [number]	3,722,256

Table 9| Assembly statistics.

Assembly statistics	
*N*75 [bp]	34,375
*N*50[3] [bp]	57,428
*N*25 [bp]	90,903
Minimum contig length [bp]	499
Maximum contig length [bp]	180,550
Average contig length [bp]	29,795
Contig count	201
Total contigs length [bp]	5,988,703
Number of aligned reads	3,011,491
Total number of aligned bases	769,207,848
Average coverage	128

[3] *N*50: represents the statistic in which 50% of the entire assembly is contained in contigs with equal or longer length (80).

3.2 Automatic genome annotation

The gene annotation of the contigs was performed in RAST, which identified 380 subsystems (collections of functionally related protein families), 5,382 protein-encoding genes and 57 RNAs (figure 7). Additionally, tRNAscan-SE-1.23 (64) detected 52 tRNA genes and RNAmmer 1.2 (65) detected 5 rRNA genes (three 5S, one 16S, and one 23S).

Considering the biological function of the proteins, RAST identified 82 genes coding for stress response, including detoxification (10 genes), oxidative (38 genes) and osmotic stress (15 genes). Also, genes related with resistance to metals, arsenic (3 genes), cobalt-zinc-cadmium (40 genes) and zinc (1 gene), were identified. This suggests that NL19 can withstand metal contaminated environments.

Furthermore, genes encoding antibiotic resistance to beta-lactams (15 genes), fluoroquinolones (4 genes) and streptothricin (2 genes) were also identified. Genes coding for multidrug resistance efflux pumps (23 genes) were also identified.

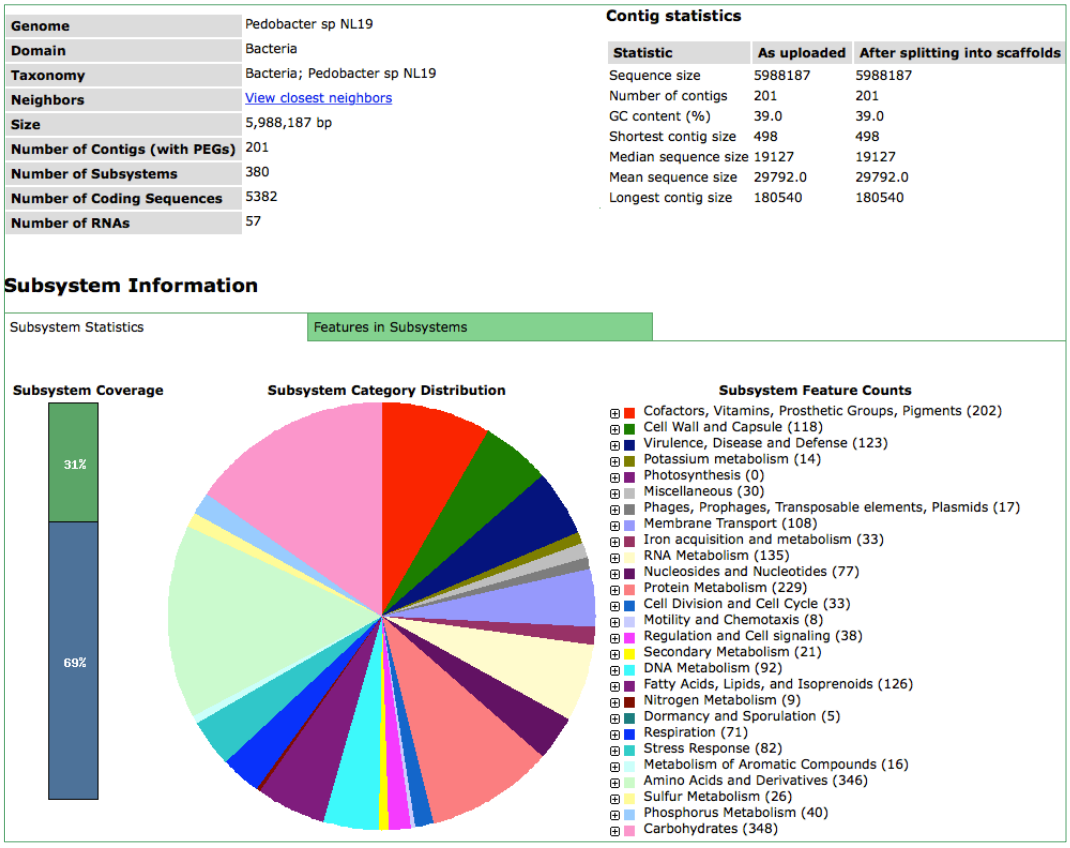


Figure 7| Result obtained with the RAST annotation of the *Pedobacter* sp. NL19 genome.

A second automatic annotation was performed with NCBI PGAAP during the genome submission at NCBI. The overall result was similar to the previous annotation because it identified 5,273 genes, 52 tRNAs and 6 rRNAs (three 5S, one 16S, and two 23S). However, NCBI detected one more 23S rRNA than RNAmmer due to probable disruption of the contigs, resulting in truncated rRNAs, and therefore erroneous annotation. Since this is a draft genome the correct numbers of rRNAs cannot be fully identified. There are only two complete genomes of *Pedobacter* genus: the type strain *Pedobacter heparinus* DSM 2366 (accession number CP001681) and *Pedobacter saltans* DSM 12145 (CP002545). In these, 7 rRNAs (one 5S, three 16S and three 23S) and 12 rRNAs (four 5S, four 16S and four 23S) were respectively identified. Overall, the GC content predicted by RAST, for the genomic contigs of *Pedobacter* sp. NL19 was 39.0%. This content is similar to other sequenced genomes of the same genus, with a range of GC content between 34.4% to 44.9% (60).

3.3 Gene Ontology (GO)

Blast2GO (B2G) was used for the Gene Ontology terms association through functional categorization of the predicted PEGs. The proteins were classified in the three main domains: i) biological process, ii) molecular function and iii) cellular component (figure 8).

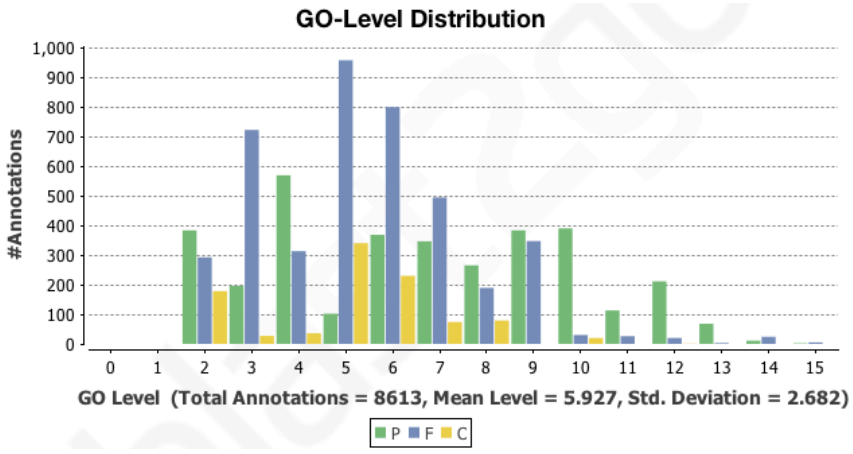


Figure 8| GO level distribution in the three main domains, (P) biological process, (F) molecular function and (C) cell component.

The distribution of the annotated PEGs in the different GO levels exhibited a concentration in levels 2, 4 and 6-10 for biological process (P), 3 and 5-7 for molecular function (F) and 5-6 for cellular component (C) (figure 8). Considering that deeper levels indicate precise accuracy in the GO annotation (80) the results showed a good term categorization. The sequences were categorized at GO level 2 in 13 biological process, 11 molecular function and 6 cellular component categories. Considering this, a nested graphical representation was constructed, until the GO level 3 distribution (figure 9 to 11).

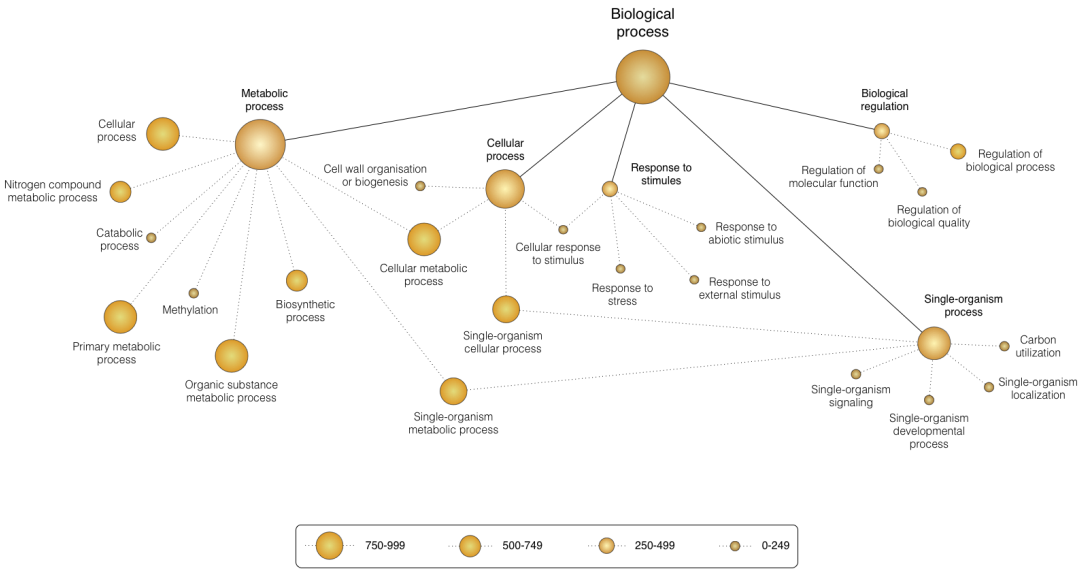


Figure 9| Nested representation of biological process until GO level 3 classified using Blast2GO. The five main classes are highlighted in bold (GO level 2). The size of the text and circles are proportional to the total of sequences associated.

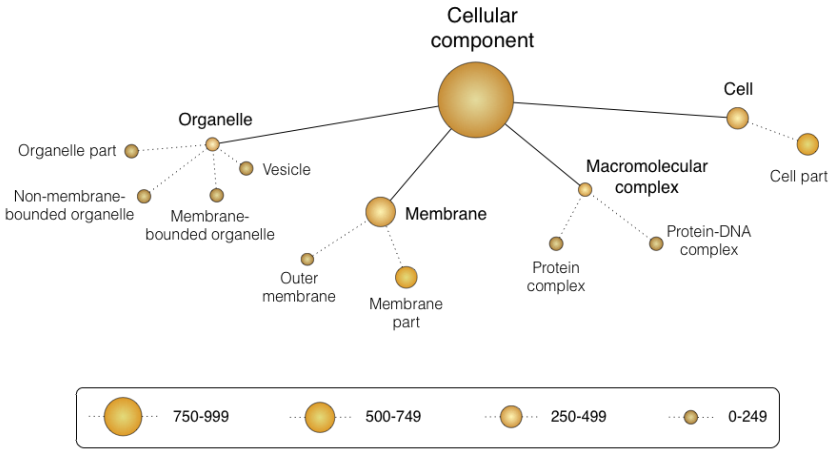


Figure 10| Nested representation of cellular component until GO level 3 classified using Blast2GO. The four main classes are highlighted in bold (GO level 2). The size of the text and circles are proportional to the total of sequences associated.

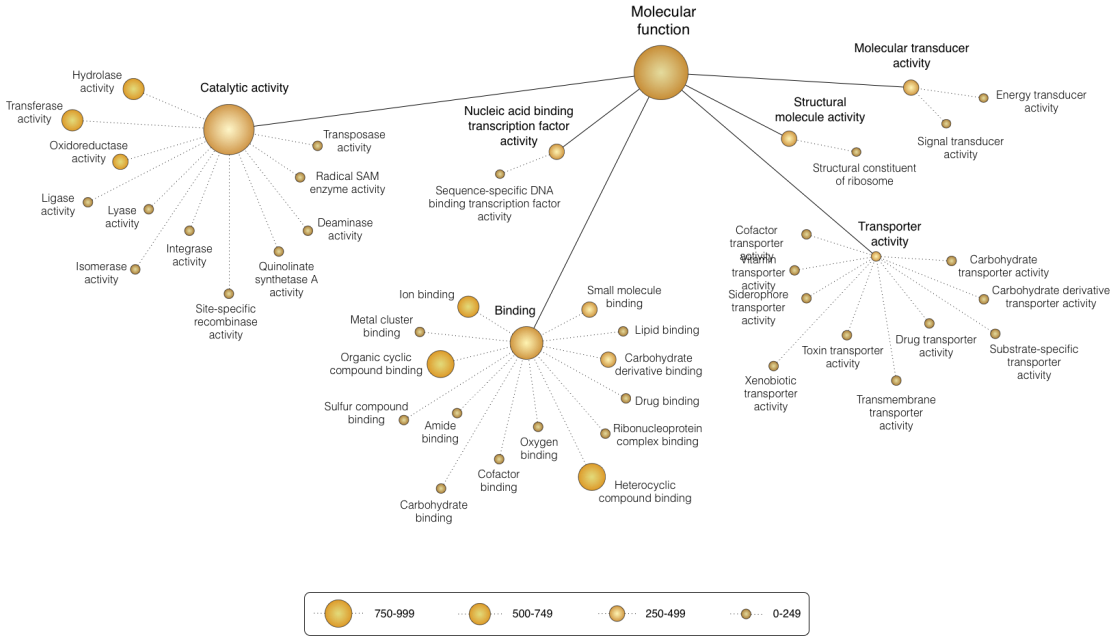


Figure 11| Nested representation of molecular function until GO level 3 classified using Blast2GO. The six main classes are highlighted in bold (GO level 2). The size of the text and circles are proportional to the total sequences associated.

The most abundant biological process GO terms were involved in metabolic process (1,918 sequences, 33.76%), cellular process (1,362 sequences, 23.97%), single-organism process (1,198 sequences, 21.09%), localization (352 sequences, 6.20%), biological regulation (348 sequences, 6.13%), response to stimuli (261 sequences, 4.59%) and signalling (145 sequences, 2.55%).

Most of the molecular function GO terms were involved in catalytic activity (1,853 sequences, 50.86%), binding (1,096 sequences, 30.09%), transporter activity (248 sequences, 6.81%), nucleic acid binding transcription factor activity (119 sequences, 3.27%) and molecular transducer activity (105 sequences, 2.88%).

For cellular component the most abundant GO terms were involved in membrane (514 sequences, 43.16%), cell (494 sequences, 41.48%), macromolecular complex (113 sequences, 9.49%) and organelle (66 sequences, 5.54%).

3.4 Identification of clusters encoding the biosynthesis of secondary metabolites

The annotation obtained with RAST predicted the presence of 21 genes that can be related with the biosynthesis of secondary metabolites by NL19 strain (table 10). These included clavulanic acid biosynthesis (1 gene), alkaloid biosynthesis from L-lysine (1 gene), auxin biosynthesis (5 genes) and lanthionine synthetases (14 genes).

Table 10| Contigs containing genes related with the biosynthesis of secondary metabolites in the draft genome of NL19 strain, as predicted by RAST annotation.

	Protein	Contig	PEG	Domains
Clavulanic acid biosynthesis	Clavaldehyde dehydrogenase	109	226	Clavulanic acid dehydrogenase
Alkaloid biosynthesis	Deoxyhypusine synthase	25	1882	Deoxyhypusine synthase
Auxin biosynthesis	Tryptophan synthase alpha chain	89	4946	Ttryptophan synthase (TRPS) alpha subunit
	Aromatic-L-amino-acid decarboxylase	38	2726	DOPA decarboxylase family
	Anthranilate phosphoribosyltransferase	89	4949	Anthranilate phosphoribosyltransferase
	Tryptophan synthase beta chain	89	4947	Tryptophan synthase-beta
	Phosphoribosylanthranilate isomerase	89	4948	Phosphoribosylanthranilate isomerase
Lanthionine synthetases	Protein-L-isoaspartate O-methyltransferase	98	5118	S-adenosylmethionine-dependent methyltransferases, class I
	Dihydropyidine synthase	56	3502	SpaB C-terminal
	Cyclase LanC	17	1463	LanC
		34	2547	LanC
		36	2695	LanC
	Dehydratase LanB	17	1461	Dehydratase C-terminal & SpaB C-terminal
		17	1462	Dehydratase N-terminal
		24	1808	Dehydratase & SpaB C-terminal

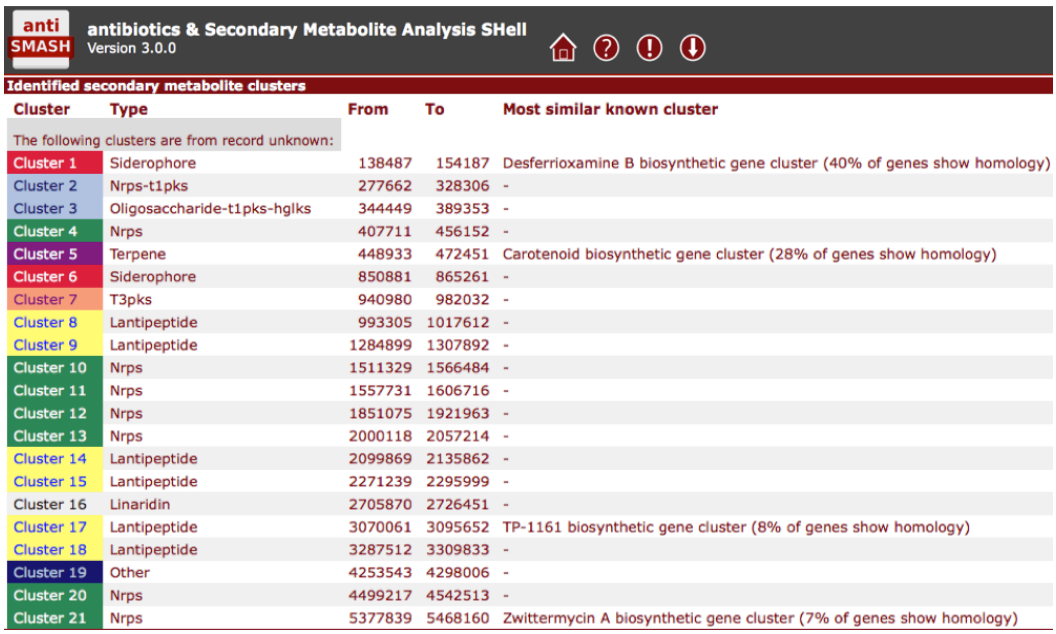
		34	2548	Dehydratase & SpaB C-terminal
		34	2556	Dehydratase & SpaB C-terminal
		36	2693	Dehydratase & SpaB C-terminal
		56	3501	Dehydratase N-terminal & Dehydratase C-terminal
		60	3703	Dehydratase & SpaB C-terminal
	Protein LanM	56	3503	LanM-like

Since the type strain *Pedobacter heparinus* DSM 2366 and *Pedobacter* sp. V48 were identified as potential antimicrobial producers (60, 61), the complete genome of *Pedobacter heparinus* DSM 2366 (accession number CP001681) and the genomic contigs of V48 strain (accession number AWRU00000000) were selected to perform the annotation comparison with NL19 strain secondary metabolites. Similarities were identified between the three bacteria for auxin biosynthesis (4 genes in *P. heparinus* and V48 strain). NL19 strain contained one more gene (aromatic-L-amino-acid decarboxylase). However, the alkaloid biosynthesis was only similar between *Pedobacter heparinus* DSM 2366 and NL19 strain. In the three genomes genes related with the biosynthesis of lanthipeptides were identified. In V48 strain, only a *lanB* gene was identified. *Pedobacter heparinus* DSM 2366 possesses genes encoding a LanC and a truncated LanB synthetase in the same gene cluster. Interestingly, NL19 genome possesses more lanthipeptide-related determinants than these two bacteria. This can possibly be due to the limiting characteristics of the environment from where the strain was isolated, where these type of metabolites can play a key role in the competition for nutrients and survival of bacteria (34, 81).

A more detailed analysis of the NL19 draft genome was performed with the web-based platform antiSMASH 3.0. The platform was specifically developed for the *in silico* prediction of gene clusters associated with the production of secondary metabolites (18). It has been widely used in several studies for mining bacteria without any previous knowledge on their antimicrobial activity. For instance, with the aid of this platform,

several RiPs clusters were recently identified in anaerobic bacteria (46) and clusters of specialized metabolites in *Streptomyces coelicolor* A3(2) (82).

With this approach, 21 clusters putatively encoding secondary metabolites were identified (figure 12), which included nonribosomal peptides (NRPS; 7), lanthipeptides (6), siderophores (2), terpenes (1), linaridins (1), polyketides (PKS; 1) and NRP-PK hybrids (1).



Identified secondary metabolite clusters

Cluster	Type	From	To	Most similar known cluster
The following clusters are from record unknown:				
Cluster 1	Siderophore	138487	154187	Desferrioxamine B biosynthetic gene cluster (40% of genes show homology)
Cluster 2	Nrps-t1pks	277662	328306	-
Cluster 3	Oligosaccharide-t1pks-hglks	344449	389353	-
Cluster 4	Nrps	407711	456152	-
Cluster 5	Terpene	448933	472451	Carotenoid biosynthetic gene cluster (28% of genes show homology)
Cluster 6	Siderophore	850881	865261	-
Cluster 7	T3pks	940980	982032	-
Cluster 8	Lantipeptide	993305	1017612	-
Cluster 9	Lantipeptide	1284899	1307892	-
Cluster 10	Nrps	1511329	1566484	-
Cluster 11	Nrps	1557731	1606716	-
Cluster 12	Nrps	1851075	1921963	-
Cluster 13	Nrps	2000118	2057214	-
Cluster 14	Lantipeptide	2099869	2135862	-
Cluster 15	Lantipeptide	2271239	2295999	-
Cluster 16	Linaridin	2705870	2726451	-
Cluster 17	Lantipeptide	3070061	3095652	TP-1161 biosynthetic gene cluster (8% of genes show homology)
Cluster 18	Lantipeptide	3287512	3309833	-
Cluster 19	Other	4253543	4298006	-
Cluster 20	Nrps	4499217	4542513	-
Cluster 21	Nrps	5377839	5468160	Zwittermycin A biosynthetic gene cluster (7% of genes show homology)

Figure 12| Biosynthetic clusters identified with antiSMASH web-based platform**.**

The identification of these clusters does not necessarily mean that NL19 strain possesses and/or expresses these 21 biosynthetic clusters. It should be highlighted that this number can be overestimated by the use of genomic contigs instead of scaffolds or whole genome. This is especially important for larger clusters such as those encoding NRPS or lanthipeptides.

From the six lanthipeptides clusters identified only four of them were considered as “complete”, meaning by that they contain the precursor peptide, the modification enzymes LanB and LanC, and also the transport enzyme. In the remaining two clusters only the modifying enzyme LanB was found. Therefore, only the four “complete” clusters of lanthipeptides were considered for further characterization in this thesis. Accordingly,

the identified lanthipeptide clusters will be further analysed and discussed in the following sections.

3.5 Description of the clusters encoding the biosynthesis of lanthipeptides in NL19 strain

The identified biosynthetic clusters and their lanthipeptide synthethases were named according to the cluster number attributed by antiSMASH. Therefore, *ped8*, *ped14*, *ped15* and *ped17* (figure 13) correspond to the clusters detected in contigs 17, 34, 36 and 56, respectively. These four clusters show similar features and all of them contain genes encoding the essential proteins for the biosynthesis of lanthipeptides, which include the precursor peptides (encoded by the *pedA* structural genes), dehydratases and cyclases (*pedB* and *pedC*) and a protease/transporter gene (*pedT*).

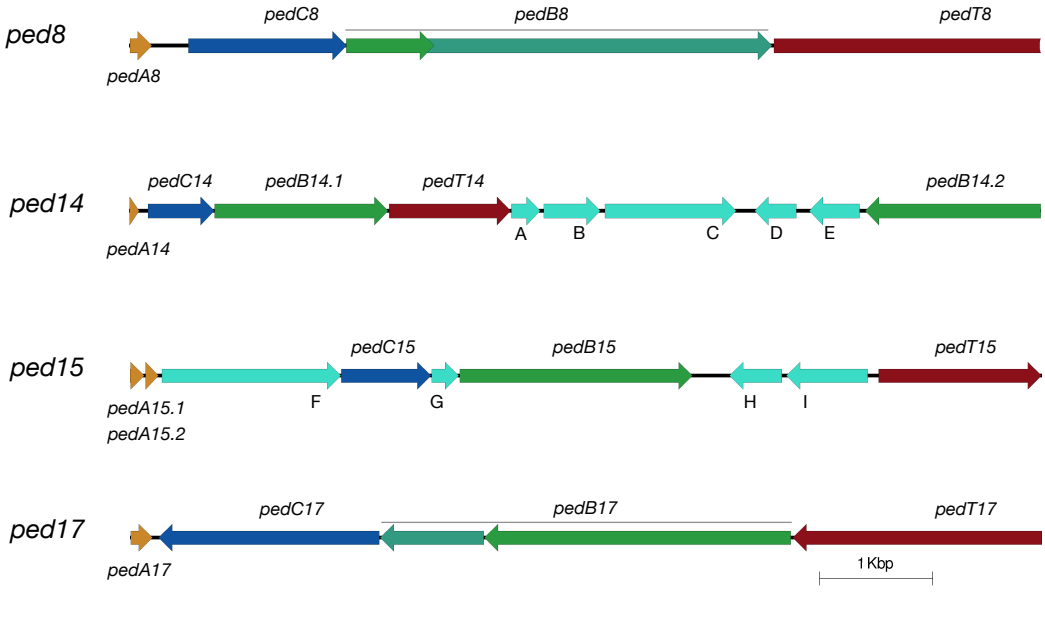


Figure 13| Schematic representation of four lanthipeptide gene clusters identified in the draft genome of *Pedobacter* sp. NL19.

The most recent scheme used to classify the lanthipeptides is based on the enzymes responsible for the dehydration and cyclization reactions. In class I lanthipeptides these reactions are performed by two different proteins: LanB and LanC. Thus, the four clusters of NL19 strain should correspond to this class of lanthipeptides due to the presence of the genes *pedB* and *pedC* that encode LanB and LanC proteins, respectively (figure 13).

3.5.1 The LanBs of *Pedobacter* sp. NL19 - PedBs

The LanB proteins are characterized by the presence of a lanthipeptide dehydratase and a SpaB_C (SpaB C-terminus) domains (44) (figure 14). The predicted PedB14.1, PedB14.2 and PedB15 proteins present these two domains in a single protein. However, in *ped8* and *ped17* clusters, these are encoded in two different ORFs. In the case of *pedB8*, it was confirmed by Sanger reaction that this was due to an error (nucleotide insertion) in the sequence that caused a premature STOP codon. So, in fact, the PedB8 is a single protein composed by the dehydratase and SpaB domains.

Usually, each LanB is responsible for the dehydration of only one precursor peptide. Yet, the *ped14* cluster encodes two distinct PedB dehydratases and only one structural gene.

Figure 14| *Pedobacter* sp. NL19 lanthipeptide synthethases LanB and LanC proteins. The LanB protein comprises the N-terminus dehydratase domain (green and yellow boxes) and C-terminus SpaB_C domain (turquoise). LanCs are represented in blue and their zinc-binding motif is highlighted with green lines in the cyclase domain.

3.5.2 The LanCs of *Pedobacter* sp. NL19 - PedCs

The LanC proteins have a conserved zinc-binding motif (Cys^{421}, Cys^{471}, His^{472}; figure 14 and figure 15), involved in the enzymatic activity and in the activation of cysteine residues for nucleophilic attack. The alignment of the PedC proteins of NL19 with other already characterized LanC enzymes revealed that only PedC8 and PedC14 have this conserved motif (figure 15). These two proteins have also have two other critical conserved residues of LanCs (Asp^{233}, His^{249}; figure 15) involved in the control and correct cyclization. However, PedC15 and PedC17 do not have any of these residues (figure 14 and figure 15). Enzymes without these conserved residues were described as being unable of an accurate cyclization of the precursor peptide (44, 83).

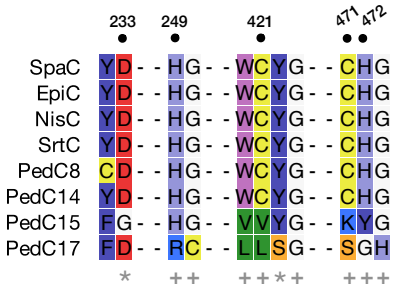


Figure 15| Alignment of critical conserved residues in different LanC enzymes with PedB enzymes. SpaC, subtilin biosynthesis cyclase; EpiC, epidermin biosynthesis cyclase; NisC, nisin biosynthesis cyclase; SrtC, streptin biosynthesis cyclase. The symbol plus (+) represents residues conserved only in LanCs and the asterisk (*) represents residues conserved in LanCs and LanMs.

3.5.3 The LanTs of *Pedobacter* sp. NL19 - PedTs

The removal of the leader sequence is an essential step for the production of active lanthipeptides. In class I lanthipeptides, this reaction is catalyzed by the LanP protease and is then followed by the transport outside from the cell by an ABC transporter (LanT). Interestingly, in all the clusters of NL19 strain, the identified PedTs are homologues to class II lanthipeptides LanTs and not to class I LanPs and LanTs (figure 16).

The LanTs from class II lanthipeptides are bifunctional proteins responsible for: i) the recognition and cleavage of the leader peptide after the double-Gly motif and ii) the transport of the mature peptide outside of the cell (47, 84). Thus, three domains constitute these LanT proteins: the N-terminus peptidase C39 domain followed by an ABC transmembrane domain and a C-terminus ABC ATPase (figure 16).

Clusters containing the association of LanB/LanC enzymes (class I) with class II LanT proteins were very recently described for the first time by Singh and Sareen (84). The study consisted of *in silico* analysis and identified three of these type of clusters, two of them in the proteobacteria *Corallococcus coralloides* DSM 2259 and *Cystobacter fuscus* DSM 2262, and the other in the actinobacteria *Mycobacterium tusciae* JS617 (84). Nevertheless, the biosynthesis of the lanthipeptides encoded in these clusters were not characterized so far. Considering the lack of literature and the possibility of producing

relevant bioactive compounds, further laboratorial experiments are needed to understand the full potential of the compounds encoded by this type of clusters.

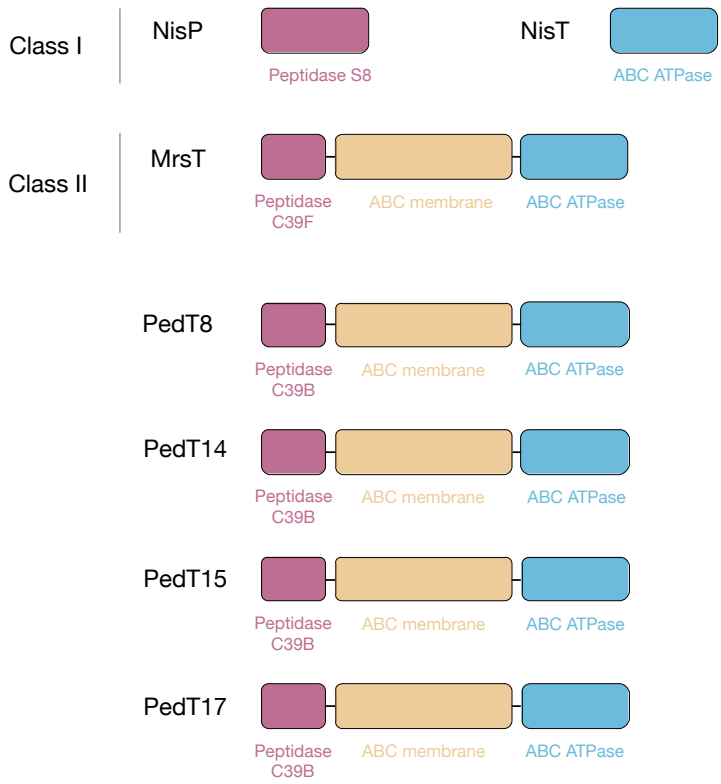


Figure 16| *Pedobacter* sp. NL19 lanthipeptide transporter LanT protein. The bifunctional LanT protein comprises the N-terminus peptidase domain, responsible for the leader peptide cleavage (purple box) and C-terminus ABC transmembrane domains, responsible for the export of the mature peptide.

3.5.4 The LanAs of *Pedobacter* sp. NL19 - PedAs

Each of the *pedA8*, *pedA14* and *pedA17* clusters possesses one structural gene *pedA*. However, cluster *ped15* contains two ORFs that potentially encode lanthipeptides precursor peptides (figure 13). All the PedA peptides have a double-Gly motif that is most probably recognized by the abovementioned PedT proteins (figure 17A). Thus, it is expected that this motif define the end of the leader peptides. The leader peptides of PedA15.1 and PedA15.2 present a high homology, suggesting that their core peptides are both modified by the same PedB15 and PedC15 enzymes (or at least very similar PedB

and PedC). Nevertheless, all leader sequences of PedA peptides share a KBX_nKL (where B= Leu, Val or Ile; figure 17A) consensus sequence. The leader sequences of class I lanthipeptides have the conserved "FXLD" motif, which is recognized by LanB enzymes (figure 17B). This was not identified in PedA leader peptides (figure 17B). Nevertheless, apart from PedA14, all of them have the "LD" sequence (where D is one of the two negatively charged amino acids) in their central regions (figure 17B). The homology of the core peptides of PedA is very low and the number of Cys residues differs between two (PedA8 and PedA17) and three (PedA14, PedA15.1 and PedA15.2), suggesting that the mature peptides will have between two and three Lan/MeLan residues, respectively (figure 17A). It was also observed that the N-terminus of PedA15.1 and PedA15.2 core peptides share some homology.

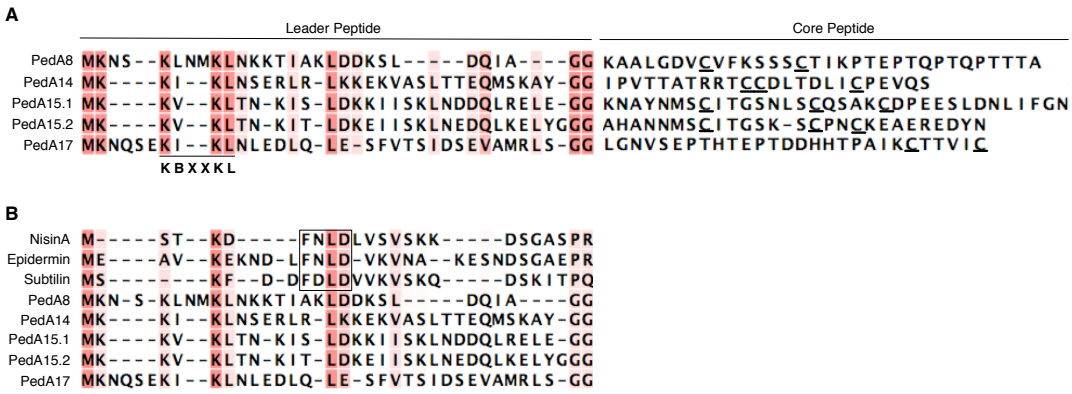


Figure 17| Alignment of *Pedobacter* sp. NL19 putative LanAs (A), with the KBX_nKL consensus sequence and the double-glycine motif in their leader sequence. The Cys residues present in the core peptide are underlined. Alignment of PedA and three known class I lanthipeptides leader sequences, where the class I conserved FNLD motif was highlighted with a black box (B).

3.5.5 Other genes identified in the clusters

Other genes were identified in the clusters *ped14* and *ped15* (figure 13). In *ped14* cluster, these genes encode the following proteins: i) an ABC transporter (*orfA*), ii) an membrane protein associated to a permease of the drug/metabolite transporter (RhaT), iii) a regulatory complex TorS (*orfC*), composed by a signal receiver domain (REC), a histidine kinase ATPase (HATPase) and a histidine kinase A domain (HisKA), iv) an ATPase

involved in chromosome partitioning (*orfD*) and v) an hypothetical protein with an pimeloyl-ACP methyl ester carboxylesterase (MhpC multi-domain). In the *ped15* cluster it was identified (figure 13): i) a gene encoding a protein composed by an outer membrane receptor domain (CirA) and a carboxypeptidase regulatory domain – *orf*F, ii) a gene encoding a hypothetical protein with a PRK11697 multi-domain (*orf*G) and iii) two genes encoding proteins of a regulatory system complex composed by the LytT protein (*orfH*) that has a REC domain and a DNA-binding domain, and a histidine kinase (*orfI*).

Some of these proteins, namely LytT and TorS, function as a two-component signal transduction systems, with the sensor domain detecting a signal and inducing the autophosphorylation of a histidine residue (52, 85). Subsequently, the phosphoryl group is transferred to the receiver domain of the response regulator mediating the cellular response to the stimulus (52, 85). These type of two-component systems were described as regulators of the biosynthesis of some lanthipeptides, namely nisin, subtilin and mersacidin (52). Thus, the presence of these systems on the *ped* clusters can indicate that their production is a regulated process (52, 86).

3.6 Phylogenetic analysis of the lanthipeptide synthetases

The number of genomes available in the public databases (complete and draft) is constantly increasing. Thus, the presence of PedB, PedC and PedT-like proteins in other genomes was investigated and their phylogenies were analysed according to Zhang et al. (2012) (44).

3.6.1 Phylogenetic analysis of PedB enzymes

Five lanthionine dehydratases were identified in four biosynthetic gene clusters (PedB8, PedB14.1, PedB14.2, PedB15 and PedB17). However, PedB17 is encoded in two different ORFs resulting in a disruption between the dehydratase domains and SpaB_C domain (figure 13 and figure 14). Therefore this protein was not considered for the phylogenetic analysis.

The phylogenetic tree of LanB was constructed using the Bayesian MCMC (figure 18) and maximum-likelihood method (appendix 3, figure 26) and resulted in two similar tree topologies Overall, the enzymes were divided in two distinct clades: one consisting of LanB proteins from actinobacteria and the other with LanBs from firmicutes, bacteroidetes and proteobacteria (figure 18).

The enzymes from bacteroidetes and proteobacteria seem to be derived from firmicutes, as already suggested by Zhang et al. (2012) (44). All the dehydratases from *Pedobacter* sp. NL19 grouped within the bacteroidetes clade. PedB14.1 and PedB15 were clustered together, where their most closely related LanBs are from genomes of *Chitinophaga pinensis* DSM 2588 and *Segetibacter koreensis* DSM 18137. The PedB8 and PedB14.2 enzymes share the same clade with the LanBs from proteobacteria and are more closely related to LanBs of *Runella slithyformis* DSM 19594 and *Sphingobacterium* sp. H1ai.

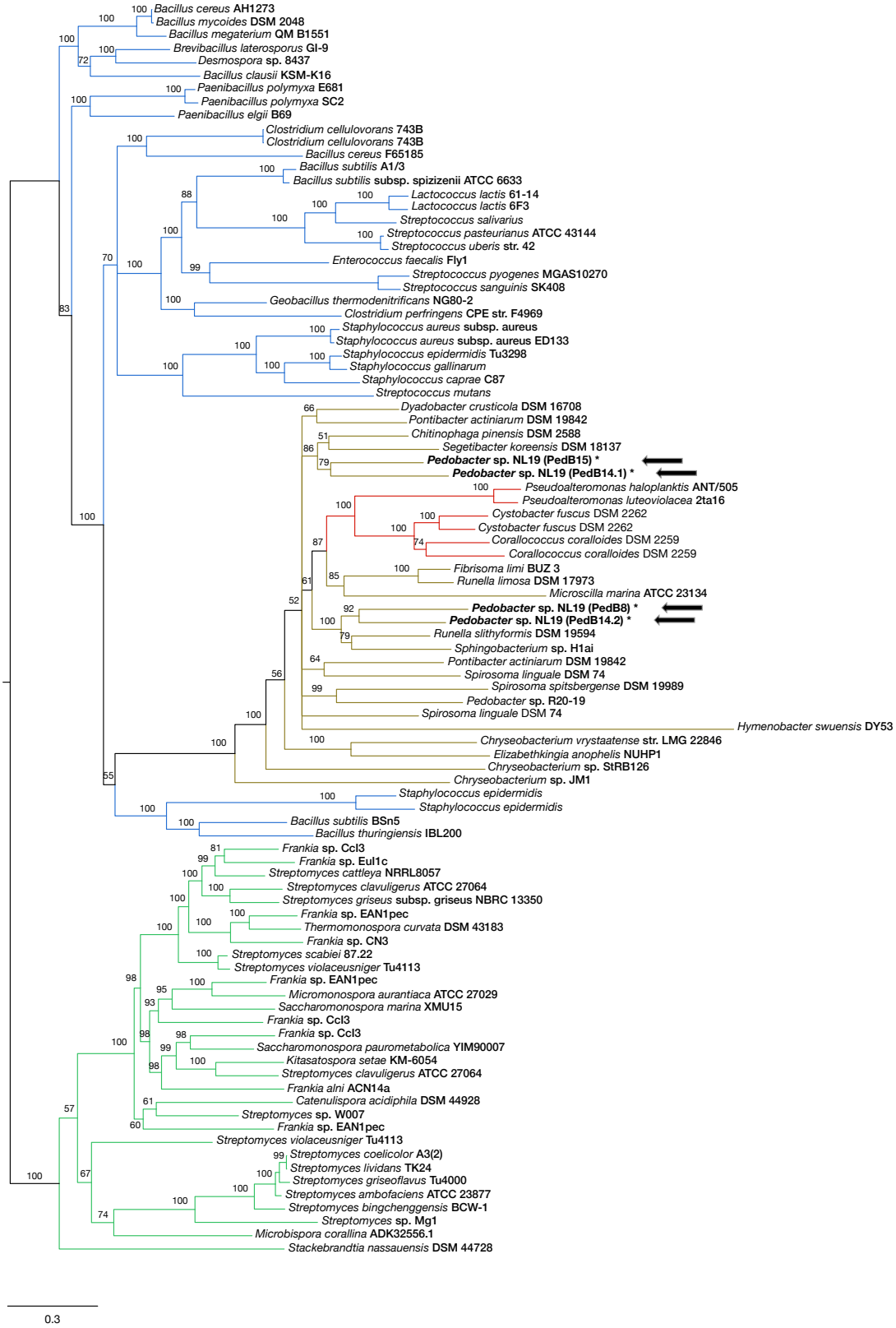


Figure 18| Bayesian MCMC phylogeny of LanB enzymes from different phyla. The Bayesian posterior probability is shown above or below the lines. Bacteroidetes are shown with yellow lines, proteobacteria with red lines, firmicutes with blue lines and actinobacteria with green lines. Black arrows indicate the *Pedobacter* sp. NL19 LanBs position.

3.6.2 Phylogenetic analysis of PedC enzymes

For the phylogenetic analysis of the LanC enzymes, generally only sequences belonging to the same gene cluster as the LanBs included in the previous section were considered. Nonetheless, some LanCs enzymes were used despite their truncated LanB enzyme, such as *Pedobacter heparinus* DSM 2366. Additionally, LanCs from bacteroidetes species without the zinc-binding motif were selected to understand the clustering of the NL19 PedC15 and PedC17. The trees constructed with the Bayesian MCMC method (figure 19) and maximum-likelihood (appendix 4, figure 27), revealed an evolutionary division between the LanC cyclases and the remaining C-terminal LanLs and LanMs enzymes, as described by Zhang et al. (2012) (44). Moreover, the human LanCLs, with unknown functions, were closely related with LanMs and LanLs than the LanCs, as already showed by previous evolutionary studies of these enzymes (44, 87). Thus, the created LanC polyphiletic clade was subdivided in two sub-clades, one comprising bacteroidetes and proteobacteria enzymes and the second with enzymes from firmicutes and actinobacteria. This result is similar a the study conducted before were the bacteroidetes and proteobacteria enzymes formed a unique clade (44, 87). Also, the LanCs from proteobacteria and bacteroidetes seem to have evolved together, as observed for their LanB enzymes.

Focusing on *Pedobacter* sp. NL19, it was found that all the PedC enzymes are part of different polytomies, where PedC14 form a single clade. PedC8 grouped with the LanCs of *Chryseobacterium vrystaatense* LMG 22846 and *Elizabethkingia anophelis* NUHP1. The polytomies represent an analytical problem because in such cases, the evolutionary relationship is not fully resolved. The PedC15 seem to be more closely related to the LanCs of *Chryseobacterium* sp. StRB126 and *Pontibacter actiniarum* DSM 19842 and also to other proteins without the conserved zinc-binding motif. Interestingly, PedC17 shows a higher evolutionary genetic divergence from the other bacteroidetes LanCs. Thus, the PedCs and PedBs from the same gene cluster do not show homology to their counterparts of the same species, raising some questions about their coevolution. So, the improvement of the evolutionary relationship resolution for LanC proteins can further contribute to this issue.

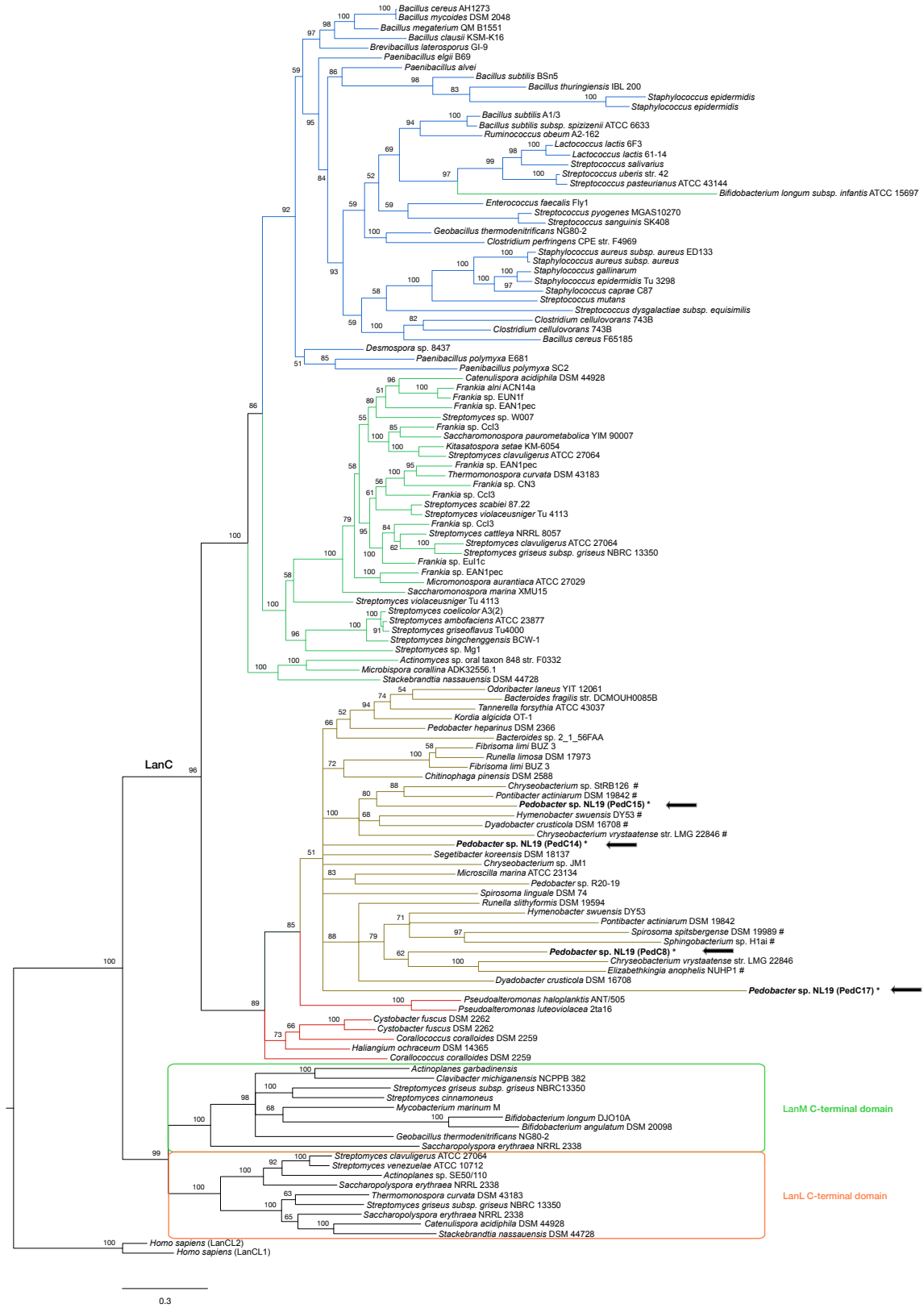


Figure 19| Bayesian MCMC phylogeny of LanC enzymes from different phyla. The Bayesian posterior probability is shown above or below the lines. Bacteroidetes are shown with yellow lines, proteobacteria with red lines, firmicutes with blue lines and actinobacteria with green lines. Black arrows indicate the *Pedobacter* sp. NL19 LanCs position. The symbol cardinal (#) represents LanCs without zinc-binding motif.

3.6.3 Phylogenetic analysis of PedT proteins

The LanT phylogenetic analysis was performed with LanTs found in the same gene clusters as LanCs and LanBs from bacteroidetes and proteobacteria of the previous sections. Some transporters of class I and class II were also included (appendix 1, table 15). The tree obtained with Bayesian MCM method (figure 20) and maximum-likelihood (appendix 5, figure 28) suggests that the LanTs of bacteroidetes and proteobacteria probably evolved from the transporters of class II lanthipeptides from firmicutes.

The PedTs present in NL19 genome are deeply grouped within the bacteroidetes clade. PedT8 and PedT17 are more closely related to each other than with PedT14 and PedT15. Furthermore, PedT15 was clustered with *Hymenobacter swuensis* DY53. Interestingly, PedT14 seem to have evolved together with other *Pedobacter* LanTs (*Pedobacter* sp. R20-19 and *Pedobacter heparinus* DSM 2366). Surprisingly, in the previously analysis, none of the LanBs or LanCs of *Pedobacter* sp. NL19 clustered together with LanBs or LanCs enzymes of the same genus.

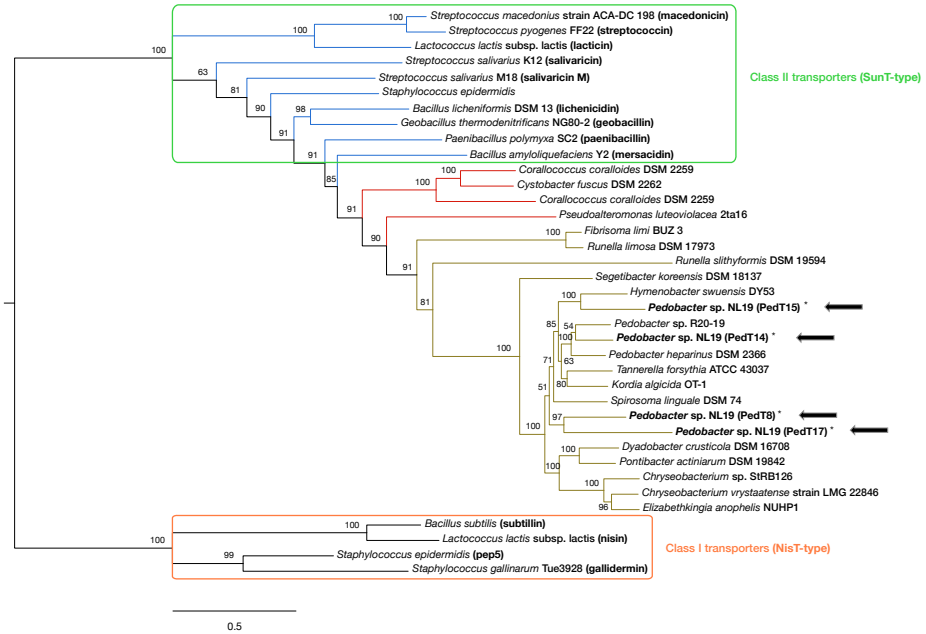


Figure 20| Bayesian MCMC phylogeny of LanT proteins from different phyla. The Bayesian posterior probability is shown above or below the lines. Bacteroidetes are shown with yellow lines, proteobacteria with red lines and firmicutes with blue lines. Black lines represents LanTs of class I. Black arrows indicate the *Pedobacter* sp. NL19 LanTs position.

3.6.4 Analysis of precursor peptides from bacteroidetes

The lanthipeptide gene clusters from *Pedobacter* genus and those encoding LanB and LanC enzymes closely related to PedBs and PedCs were selected to identify the associated structural genes (figure 21).

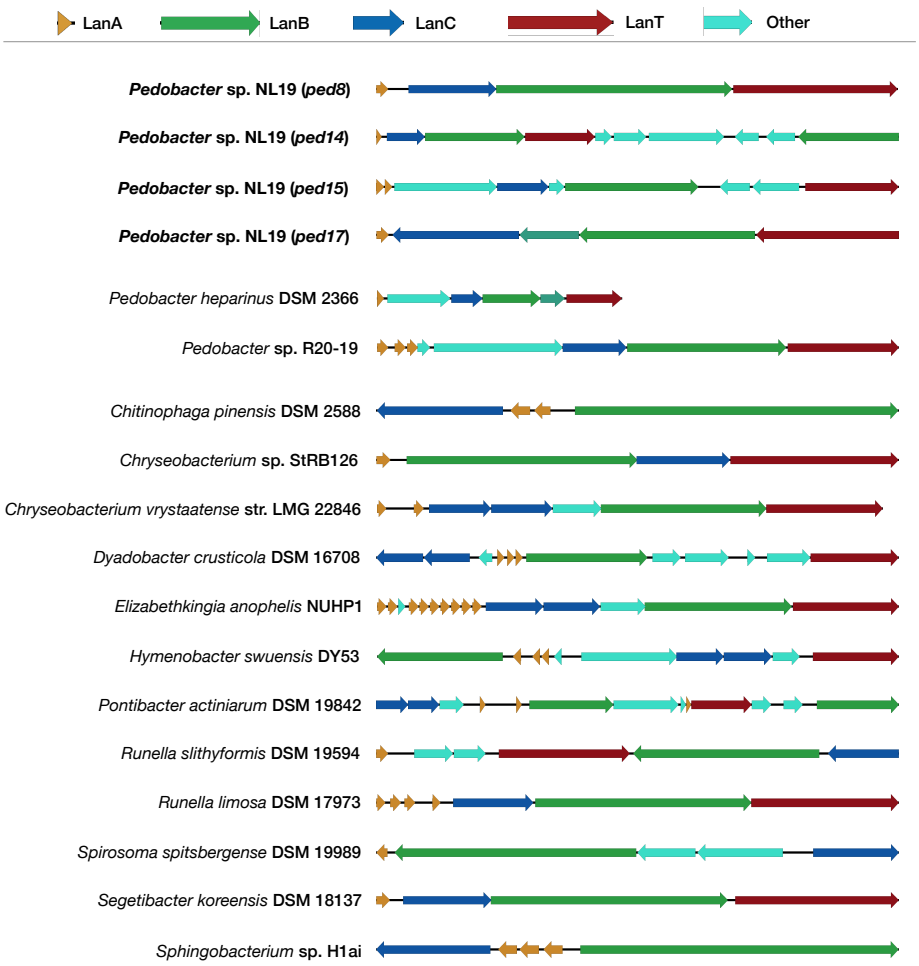


Figure 21| Representation of the lanthipeptide clusters identified in bacteroidetes species.

Like *ped15*, some of the other lanthipeptide clusters present in the genome of bacteroidetes encode more than one potential precursor peptide, including *Pedobacter* sp. R20-19, *C. pinensis* DSM 2588, *C. vrystaatense* LMG 22846, *D. crusticola* DSM 16708, *E. anophelius* NUHP1, *H. swuensis* DY53, *P. actiniarum* DSM 19842 and *Sphingobacterium* sp. H1ai. The sequences of the precursor peptides of the selected bacteroidetes were

compared. This analysis was performed separately for their leader and core peptides, excluding sequences that were equal and from the same cluster (figure 22 and 23).

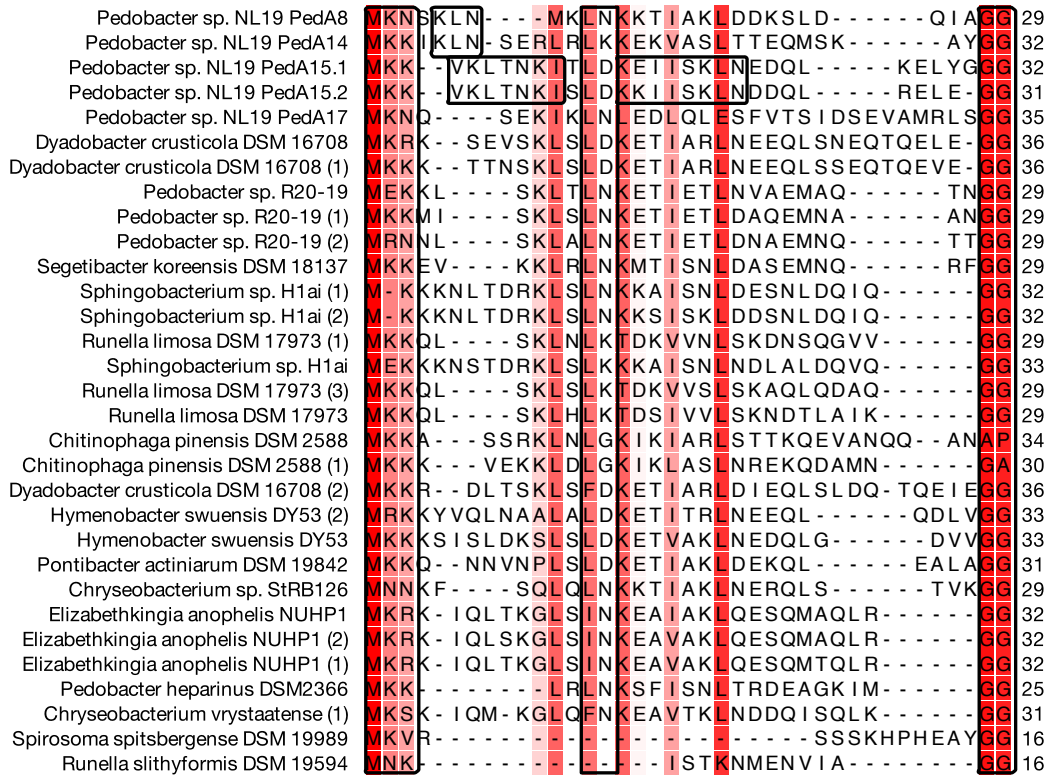


Figure 22| Alignment of the leader peptide sequence of lanthipeptides from bacteroidetes. The color gradient, from light red to red, show residues conserved from 50% to 100%, respectively. Conserved regions were represented with a black box.

The alignment of the leader peptides from the selected bacteroidetes showed that the conserved class II GG motif is a common feature (figure 22). The exception was *Chitinophaga pinensis* DSM 2588 that have different cleavage sites, one with the GA motif (also recognized by LanT from class II lanthipeptides) and the other with a predicted AP cleavage site. The identified KBX_nKL motif previously identified for *Pedobacter* sp. NL19 leader sequences does not occur in all bacteroidetes. However, they seem to share the KLXLXKXXBXXL motif (where B is Ile, Leu ou Val), which can possibly be important for the post-translational modifications that occur in the mature peptides. The leader sequences of different precursor peptides identified in the same cluster present a high degree of homology, as also observed for PedA15.1 and PedA15.2 leaders.

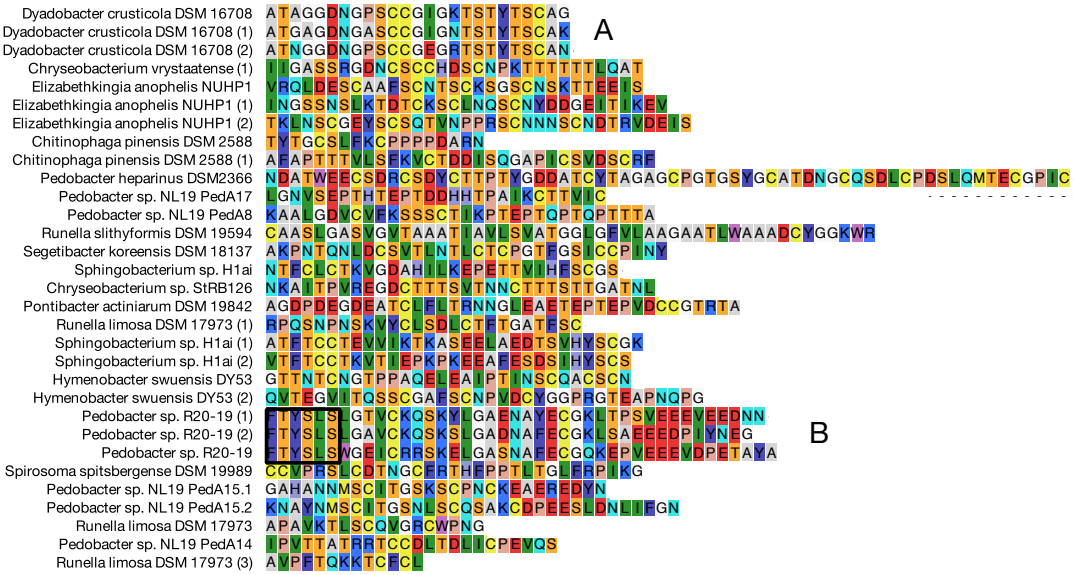


Figure 23| Amino acid sequence of the core peptides of bacteroidetes. The conserved motif of two LanA from the same species and gene cluster are highlighted for *Dyadobacter crusticola* DSM 16708 (A) and *Pedobacter* sp. R20-19 (B).

The alignment of the bacteroidetes core peptides revealed low residue conservation and therefore it is not shown. Conserved regions were not even identified for peptides in the same genus as observed for *Pedobacter* sp. NL19, *Pedobacter* sp. R20-19 and *Pedobacter heparinus* DSM 2366 (figure 23). In fact, the core peptides analysed only share some degree of homology, when encoded in the same gene cluster. For instance, the three core peptides of *Dyadobacter crusticola* DSM 16708 are highly similar (figure 23A). The three core peptides of *Pedobacter* sp. R20-19 share the N-terminus sequence FTYSLS (figure 23B), an identical situation was observed between PedA15.1 and PedA15.2 core peptides. Also, it was found that the number of Cys residues in the core peptides are distinct, ranging from 1 to 8 residues, which indicates that the number of Lan/MeLan residues in the mature peptides will also vary between 1 and 8.

3.7 Analysis of *Pedobacter* sp. NL19 by mass spectrometry

To investigate the production of PedA mature peptides by *Pedobacter* sp. NL19, colonies and culture supernatant were analysed by mass spectrometry. The colonies were tested after 2 and 7 days, and the supernatant after 7 days of growth. The masses corresponding to the predicted mass of PedA8 (M= 3180 Da) and PedA14 (M= 2616 Da) mature peptides were identified in the colonies that were incubated for 7 days and washed with 50% of acetonitrile:water solution (figure 24). However, the identification of these masses is not conclusive due to the intensity of the acquisition and high background noise obtained in the spectra. Furthermore, to confirm if the masses detected could correspond to the PedA8 and PedA14 mature peptides, it will be necessary to acquire a MS/MS spectra.

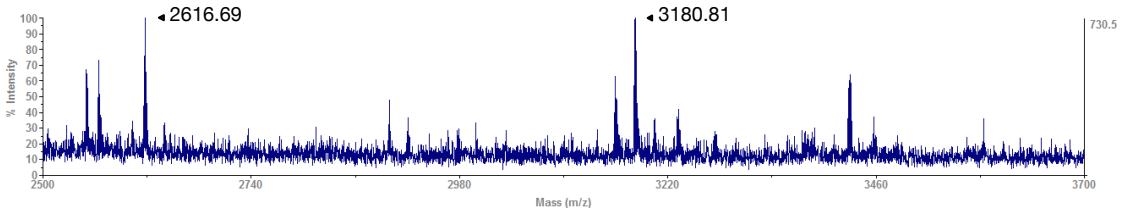


Figure 24| MALDI-TOF MS spectra for *Pedobacter* sp. NL19 colonies grown for 7 days and treated with 50% ACN:dH_2O showing the molecular masses (Da) corresponding to PedA8 and PedA14 mature peptides.

Interestingly, it was possible to identify the masses of two pedopeptins when colonies are grown for 2 days and 7 days following a treatment with a 50% acetonitrile:water solution and analysed with the sinapinic acid (figure 25). These included pedopeptin A (M= 1115 Da) and pedopeptin B (M= 1099 Da) that were described as antibiotics with broad spectrum produced by *Pedobacter* sp. SANK 72003 strain (58). Therefore, *Pedobacter* sp. NL19 is also producing pedopeptins, which might also be contributing to NL19 bioactivity observed activity. Furthermore, these results need to be further clarified with MALDI-TOF MS protocol optimization, such as the

sample concentrations and/or the matrix used, or employing a different approach to achieve better results.

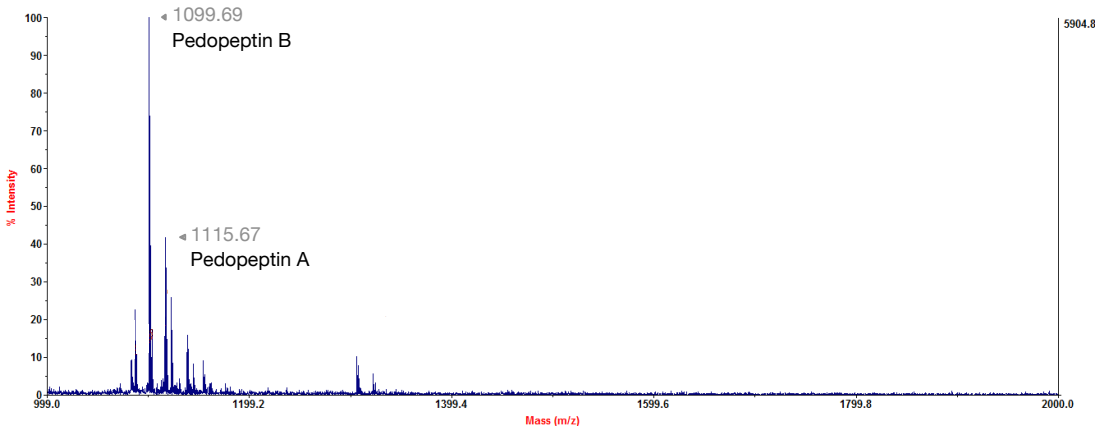


Figure 25| MALDI-TOF MS spectra for *Pedobacter* sp. NL19 colonies with pedopeptin A (M= 1115 Da) and pedopeptin B (M= 1099 Da). The colonies were treated with 50% ACN:dH_2O and the used matrix was sinapinic acid.

Chapter IV. Conclusions

The development of innovative sequencing techniques spurred great advances in genome research of various organisms. Consequently, these techniques have been exploited to search for new biosynthetic clusters (including those for antimicrobials) in a fast and cost effective manner.

Pedobacter sp. NL19 was isolated from an abandon uranium mine, which can be considered as an extreme environment due to its radioactivity and high metal concentrations. Such conditions can require a fast adaptation process from bacteria to allow their survival.

The *in vitro* analysis of NL19 strain showed its activity against Gram-negative and Gram-positive bacteria. Therefore, this bioactivity was the pretext for a finer analysis of the genome of this strain. The draft genome of *Pedobacter* sp. NL19 was obtained and further investigated especially with respect to the presence of secondary metabolite encoding gene clusters, in particular those coding for antimicrobials. The assembled genome generated a genome size of 5,988,703 bp and 201 contigs. The functional analysis performed with Blast2GO program identified 13 biological process, 11 molecular functions and 6 cellular component categories. The most abundant biological process and molecular function were metabolic process and catalytic activity, respectively.

Moreover, RAST automatic annotation of the contigs identified genes encoding for stress response, resistance to metals and also antibiotic resistance. Further analysis of NL19 genome, with the web-based platform antiSMASH, revealed the presence of many clusters encoding the biosynthesis of secondary metabolites. In total, twenty-one biosynthetic clusters were identified. The majority of them are related with the production of peptide families that include several antibacterial compounds like the lanthipeptides and nonribosomal peptides. Thus, the analysis exposed the NL19 potential as a producer of bioactive compounds. However, in this thesis a more detailed characterization was exclusively performed with the clusters of lanthipeptides given the number of clusters identified and also these have characteristics of both class I and class II lanthipeptides. The “hybrid” nature of these clusters is due to the presence of LanB and LanC enzymes, essential for the biosynthesis of class I lanthipeptides, and the bifunctional LanT protein, responsible for the removal of the leader peptide and transport. This

remarkable feature was only recently described in the literature and only for two strains. Moreover, the detailed study of these enzymes revealed the presence of one LanB truncated in *ped17*, probable due to sequencing error similar to PedB8. Additionally, a second LanB was identified in cluster *ped14*. Furthermore, two of the four identified LanC enzymes (LanC15 and LanC17) lacked the zinc-binding motif residues (Cys^{421}, Cys^{471}, His^{472}) and other residues (Asp^{233}, His^{249}) required for the correct cyclization and production of active peptides.

Furthermore, some conserved residues and two conserved motifs were identified among the *Pedobacter* sp. NL19 leader peptides, KBX_nKL and LD, except for PedA14, that contains a LK motif. However, the core peptides were very distinctive between the different gene clusters, but with high homology when present in the same gene cluster, e.g. PedA15.1 and PedA15.2.

Further analysis and genome mining of similar proteins of NL19 allowed the identification of wide distribution of LanBs, LanCs and LanTs in the bacteroidetes phylum. In addition, the presence of hybrid clusters, possessing class I and class II genes was observed, revealing a common and unique feature in the bacteroidetes. Also, by phylogenetic analysis it appears that NL19 and bacteroidetes enzymes evolved from firmicutes.

Considering all the information gathered from the enzymes in each gene cluster, Ped8 seems to be the best candidate to produce a lanthipeptide. Additionally, by MALDI-TOF MS a mass corresponding to the predicted mass of PedA8 was identified, after 7 dehydrations and the formation of two thioether rings.

4.1 Future perspectives

The study performed, largely *in silico*, showed the potential ability of *Pedobacter* sp. NL19 to produce bioactive compounds. Still, a complete genome characterization or improvement of the available assembly would be interesting for future analysis of *Pedobacter* sp. NL19. Further studies and laboratorial analysis are required to understand the implications of the lanthipeptide clusters organization and the activity of these peptides. The absence of immunity genes indicates that probably these peptides do not act as lantibiotics and thus do not show antibacterial activity. However, the isolation and characterization of these compounds will allow a better understanding of their pathway and to confirm or deny this assumption.

Additionally, further studies of the secondary metabolites identified by antiSMASH, especially the nonribosomal peptides, terpene and polyketides, should be performed to fully characterize the potential of this bacterium to produce novel compounds.

Chapter V. References

1. Richards R. The romantic conception of life: science and philosophy in the age of Goethe: The University of Chicago Press; 2002.

2. Brown T. Gene cloning and DNA analysis: an introduction: John Wiley & Sons; 2010.

3. O'Connor C. Isolating Hereditary Material: Frederick Griffith, Oswald Avery, Alfred Hershey, and Martha Chase. Nature Education. 2008;1(1):105.

4. Adams J. DNA sequencing technologies. Nature Education. 2008;1(1):193.

5. Pareek C, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011;52(4):413-35.

6. Loman N, Constantinidou C, Chan J, Halachev M, Sergeant M, Penn C, et al. High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nat Rev Microbiol. 2012;10(9):599-606.

7. Mardis E. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387-402.

8. Mardis E. A decade's perspective on DNA sequencing technology. Nature. 2011;470(7333):198-203.

9. Metzker M. Sequencing technologies - the next generation. Nat Rev Genet. 2009;11(1):31-46.

10. Nyrén P. The History of Pyrosequencing®. Pyrosequencing® Protocols: Springer; 2007. p. 1-13.

11. Lasken R, McLean J. Recent advances in genomic DNA sequencing of microbial species from single cells. Nature Rev Genet. 2014;15(9):577-84.

12. Mikheyev A, Tin M. A first look at the Oxford Nanopore MinION sequencer. Mol Ecol Resour. 2014;14(6):1097-102.

13. Schneider G, Dekker C. DNA sequencing with nanopores. Nature biotechnology. 2012;30(4):326-8.

14. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012.

15. Strickland E. The gene machine and me. Spectrum, IEEE. 2013;50(3):30-59.

16. Caboche S. Biosynthesis: Bioinformatics bolster a renaissance. Nat Chem Biol. 2014;10(10):798-800.

17. Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23(21):2947-8.

18. Blin K, Medema M, Kazempour D, Fischbach M, Breitling R, Takano E, et al. antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Res. 2013:gkt449.

19. Scherlach K, Hertweck C. Triggering cryptic natural product biosynthesis in microorganisms. Org Biomol Chem. 2009;7(9):1753-60.

20. Winter J, Behnken S, Hertweck C. Genomics-inspired discovery of natural products. Curr Opin Chem Biol. 2011;15(1):22-31.

21. Ziemert N, Podell S, Penn K, Badger J, Allen E, Jensen P. The natural product domain seeker NaPDoS: a phylogeny based bioinformatic tool to classify secondary metabolite gene diversity. PLoS One. 2012;7(3):e34064.

22. Dias D, Urban S, Roessner U. A historical overview of natural products in drug discovery. Metabolites. 2012;2(2):303-36.

23. Nunnery J. Investigations of marine cyanobacterial secondary metabolites: isolation, structure elucidation, and synthesis [doctoral's thesis]: University of California; 2012.

24. Cragg G, Newman D. Natural products: a continuing source of novel drug leads. Biochim Biophys Acta. 2013;1830(6):3670-95.

25. Kennedy D, Wightman E. Herbal extracts and phytochemicals: plant secondary metabolites and the enhancement of human brain function. Adv Nutr. 2011;2(1):32-50.

26. Vaishnav P, Demain A. Unexpected applications of secondary metabolites. Biotechnol Adv. 2011;29(2):223-9.

27. Ramakrishna A, Ravishankar G. Influence of abiotic stress signals on secondary metabolites in plants. Plant Signal Behav. 2011;6(11):1720-31.

28. Li J, Vederas J. Drug discovery and natural products: end of an era or an endless frontier? Science. 2009;325(5937):161-5.

29. Pereira A, Pita J. Alexander Fleming (1881-1955), Da descoberta da penicilina (1928) ao Prémio Nobel (1945). Revista da Faculdade de Letras: História Porto. 2005:129-51.

30. Baltz R. Renaissance in antibacterial discovery from actinomycetes. Curr Opin Pharmacol. 2008;8(5):557-63.

31. Harvey A. Natural products in drug discovery. Drug Discov Today. 2008;13(19):894-901.

32. Lam K. New aspects of natural products in drug discovery. Trends Microbiol. 2007;15(6):279-89.

33. Koehn F, Carter G. The evolving role of natural products in drug discovery. Nat Rev Drug Discov. 2005;4(3):206-20.

34. Boman H. Antibacterial peptides: basic facts and emerging concepts. J Intern Med. 2003;254(3):197-215.

35. Arnison P, Bibb M, Bierbaum G, Bowers A, Bugni T, Bulaj G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat Prod Rep. 2013;30(1):108-60.

36. Lata S, Sharma B, Raghava G. Analysis and prediction of antibacterial peptides. BMC Bioinformatics. 2007;8(1):263.

37. Kersten R, Yang Y-L, Xu Y, Cimermancic P, Nam S-J, Fenical W, et al. A mass spectrometry–guided genome mining approach for natural product peptidogenomics. Nat Chem Biol. 2011;7(11):794-802.

38. Civjan N. Natural Products in Chemical Biology. New Jersey: John Wiley & Sons; 2012. 436 p.

39. Caradec T, Pupin M, Vanvlassenbroeck A, Devignes M-D, Smail-Tabbone M, Jacques P, et al. Prediction of monomer isomery in Florine: a workflow dedicated to nonribosomal peptide discovery. PLoS One. 2014;9(1).

40. Kastin A. Handbook of biologically active peptides. 2 ed. Oxford: Academic Press; 2013.

41. Strieker M, Tanović A, Marahiel M. Nonribosomal peptide synthetases: structures and dynamics. Curr Opin Struct Biol. 2010;20(2):234-40.

42. Hahn M, Stachelhaus T. Selective interaction between nonribosomal peptide synthetases is facilitated by short communication-mediating domains. Proc Natl Acad Sci U S A. 2004;101(44):15585-90.

43. Zhang Q, Yang X, Wang H, van der Donk W. High divergence of the precursor peptides in combinatorial lanthipeptide biosynthesis. ACS Chem Biol. 2014;9(11):2686-94.

44. Zhang Q, Yu Y, Vélasquez J, van der Donk W. Evolution of lanthipeptide synthetases. Proc Natl Acad Sci U S A. 2012;109(45):18361-6.

45. Knerr P, van der Donk W. Discovery, biosynthesis, and engineering of lantipeptides. Annu Rev Biochem. 2012;81:479-505.

46. Letzel A-C, Pidot S, Hertweck C. Genome mining for ribosomally synthesized and post-translationally modified peptides (RiPPs) in anaerobic bacteria. BMC Genomics. 2014;15(1):983.

47. van der Donk W, Nair S. Structure and mechanism of lanthipeptide biosynthetic enzymes. Curr Opin Struct Biol. 2014;29:58-66.

48. Müller W, Ensle P, Krawczyk B, Süssmuth R. Leader peptide-directed processing of labyrinthopeptin A2 precursor peptide by the modifying enzyme LabKC. Biochemistry. 2011;50(39):8362-73.

49. Caetano T. Lichenicidin biosynthesis and search for novel antibacterial peptides [doctoral's thesis]: Universidade de Aveiro; 2011.

50. Dischinger J, Chipalu S, Bierbaum G. Lantibiotics: promising candidates for future applications in health care. Int J Med Microbiol. 2014;304(1):51-62.

51. Asaduzzaman S, Sonomoto K. Lantibiotics: diverse activities and unique modes of action. J Biosci Bioeng. 2009;107(5):475-87.

52. Teng K, Zhang J, Zhang X, Ge X, Gao Y, Wang J, et al. Identification of Ligand Specificity Determinants in Lantibiotic Bovicin HJ50 and the Receptor BovK, a Multitransmembrane Histidine Kinase. J Biol Chem. 2014;289(14):9823-32.

53. Draper L, Ross R, Hill C, Cotter P. Lantibiotic immunity. Curr Protein Pept Sci. 2008;9(1):39-49.

54. Chatterjee C, Paul M, Xie L, van der Donk W. Biosynthesis and mode of action of lantibiotics. Chem Rev. 2005;105(2):633-84.

55. Minogue T, Daligault H, Davenport K, Broomall S, Bruce D, Chain P, et al. Complete genome assembly of *Enterococcus faecalis* 29212, a laboratory reference strain. Genome Announc. 2014;2(5):e00968-14.

56. Steyn P, Segers P, Vancanneyt M, Sandra P, Kersters K, Joubert J. Classification of heparinolytic bacteria into a new genus, Pedobacter, comprising four species: *Pedobacter heparinus* comb. nov., *Pedobacter piscium* comb. nov., *Pedobacter africanus* sp. nov. and *Pedobacter saltans* sp. nov. Proposal of the family Sphingobacteriaceae fam. nov. Int J Syst Bacteriol. 1998;48(1):165-77.

57. Euzéby J. List of Bacterial Names with Standing in Nomenclature: a Folder Available on the Internet. Int J Syst Evol Microbiol. 1997;47(2):590-2.

58. Kozuma S, Hirota-Takahata Y, Fukuda D, Kuraya N, Nakajima M, Ando O. Screening and biological activities of pedopeptins, novel inhibitors of LPS produced by soil bacteria. J Antibiot (Tokyo). 2014;67(3):237-42.

59. Wong C, Tam H, Alias S, González M, González-Rocha G, Domínguez-Yévenes M. *Pseudomonas* and *Pedobacter* isolates from King George Island inhibited the growth of foodborne pathogens. Pol Polar Res. 2011;32(1):3-14.

60. Bitzer A, Garbeva P, Silby M. Draft genome sequence of *Pedobacter* sp. strain V48, isolated from a coastal sand dune in the Netherlands. Genome Announc. 2014;2(1):e00094-14.

61. Wyatt M, Lee J, Ahilan Y, Magarvey N. Bioinformatic evaluation of the secondary metabolism of antistaphylococcal environmental bacterial isolates. Can J Microbiol. 2013;59(7):465-71.

62. QIAGEN. CLC Genomics Workbench - User Manual: CLC bio; 2015. Available from: http://clcsupport.com/clcgenomicsworkbench/602/index.php?manual=Introduction_CLC_Genomics_Workbench.html.

63. Overbeek R, Olson R, Pusch G, Olsen G, Davis J, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2013.

64. Lowe T, Eddy S. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25(5):0955-964.

65. Lagesen K, Hallin P, Rødland E, Stærfeldt H-H, Rognes T, Ussery D. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. 2007;35(9):3100-8.

66. Santos T, Cruz A, Caetano T, Covas C, Mendo S. Draft genome sequence of *Pedobacter* sp. strain NL19, a producer of potent antibacterial compounds. Genome Announc. 2015;3(2).

67. Ashburner M, Ball C, Blake J, Botstein D, Butler H, Cherry J, et al. Gene Ontology: tool for the unification of biology. Nature Genet. 2000;25(1):25-9.

68. Conesa A, Götz S, García-Gómez J, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21(18):3674-6.

69. van Heel A, de Jong A, Montalbán-López M, Kok J, Kuipers O. BAGEL3: automated identification of genes encoding bacteriocins and (non-) bactericidal posttranslationally modified peptides. Nucleic Acids Res. 2013;41(W1):W448-W53.

70. Altschul S, Madden T, Schäffer A, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389-402.

71. Thompson J, Higgins D, Gibson T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22(22):4673-80.

72. Miller M, Pfeiffer W, Schwartz T, editors. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. Gateway Computing Environments Workshop (GCE), 2010; 2010: IEEE.

73. Ronquist F, Teslenko M, van der Mark P, Ayres D, Darling A, Höhna S, et al. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012;61(3):539-42.

74. Rambaut A. FigTree v1. 4.2: Tree figure drawing tool [16 May 2015]. Available from: http://tree.bio.ed.ac.uk/software/figtree/.

75. Darriba D, Taboada G, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011;27(8):1164-5.

76. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59(3):307-21.

77. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth B, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic acids Res. 2012;40(15):e115-e.

78. Wieser A, Schneider L, Jung J, Schubert S. MALDI-TOF MS in microbiological diagnostics - identification of microorganisms and beyond (mini review). Appl Microbiol Biotechnol. 2012;93(3):965-74.

79. Seng P, Drancourt M, Gouriet F, La Scola B, Fournier P-E, Rolain J, et al. Ongoing revolution in bacteriology: routine identification of bacteria by matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Clin Infect Dis. 2009;49(4):543-51.

80. Jakubowska A, Nalcacioglu R, Millán-Leiva A, Sanz-Carbonell A, Muratoglu H, Herrero S, et al. In Search of Pathogens: Transcriptome-Based Identification of Viral Sequences from the Pine Processionary Moth (*Thaumetopoea pityocampa*). Viruses. 2015;7(2):456-79.

81. O'Brien J, Wright G. An ecological perspective of microbial secondary metabolism. Curr Opin Biotechnol. 2011;22(4):552-8.

82. Gadd G, Sariaslani S. Advances in applied microbiology. 1 ed: Academic Press; 2014.

83. Li B, van der Donk W. Identification of essential catalytic residues of the cyclase NisC involved in the biosynthesis of nisin. J Biol Chem. 2007;282(29):21169-75.

84. Singh M, Sareen D. Novel LanT associated lantibiotic clusters identified by genome database mining. PloS One. 2014;9(3):e91352.

85. Su MS-W, Gänzle M. Novel two-component regulatory systems play a role in biofilm formation of *Lactobacillus reuteri* rodent isolate 100-23. Microbiology. 2014;160(Pt 4):795-806.

86. Wang J, Gao Y, Teng K, Zhang J, Sun S, Zhong J. Restoration of bioactive lantibiotic suicin from a remnant lan locus of pathogenic *Streptococcus suis* serotype 2. Appl Environ Microbiol. 2014;80(3):1062-71.

87. Yu Y, Zhang Q, Donk W. Insights into the evolution of lanthipeptide biosynthesis. Protein Sci. 2013;22(11):1478-89.

Chapter VI. Appendices

Appendix 1. Lanthipeptide synthetases accession numbers

Table 11| Accession numbers of LanB and LanC enzymes used in this thesis.

Strain	LanB	LanC
Actinomyces sp. oral taxon 848 str. F0332	ZP_06162153.1	ZP_06162154.1
Bacillus cereus AH1273	ZP_04178050.1	ZP_04178052.1
Bacillus cereus F65185	ZP_04205873.1	ZP_04205869.1
Bacillus clausii KSM-K16	YP_177053.1	YP_177052.1
Bacillus megaterium QM B1551	YP_003565953.1	YP_003565951.1
Bacillus mycoides DSM 2048	ZP_04172085.1	ZP_04172087.1
Bacillus subtilis BSn5	YP_004206153.1	YP_004206154.1
Bacillus subtilis subsp. spizizenii ATCC 6633	ZP_06872918.1	ZP_06872916.1
Bacillus thuringiensis IBL 200	ZP_04075567.1	ZP_04075568.1
Bacillus subtilis A1/3	AAL15564.1	AAL15566.1
Bacillus cereus F65185	ZP_04205873.1	ZP_04205869.1
Bacteroides sp. 2_1_56FAA	Partitioned	ZP_08590997.1
Bifidobacterium longum subsp. infantis ATCC 15697	Partitioned	YP_002322012.1
Brevibacillus laterosporus GI-9	CCF16798.1	CCF16796.1
Catenulispora acidiphila DSM 44928	YP_003114660.1	YP_003114661.1
Chitinophaga pinensis DSM 2588	WP_012789109.1	WP_012789112.1
Chryseobacterium sp. JM1	WP_034729687.1	WP_034729684.1
Chryseobacterium sp. StRB126	BAP32675.1	BAP32676.1
Chryseobacterium vrystaatense str. LMG 22846	WP_034747469.1	WP_034747475.1
Chryseobacterium vrystaatense str. LMG 22846	Manually annotated	WP_034747479.1
Clostridium cellulovorans 743B	YP_003845682.1	YP_003845681.1
Clostridium cellulovorans 743B	YP_003845682.1	YP_003845687.1
Clostridium perfringens CPE str. F4969	YP_473415.1	YP_473413.1
Corallococcus coralloides DSM 2259	AFE10374.1	WP_014398021.1
Corallococcus coralloides DSM 2259	WP_014396470.1	WP_014396471.1
Cystobacter fuscus DSM 2262	EPX61619.1	WP_002629116.1
Cystobacter fuscus DSM 2262	WP_002631341.1	WP_002631340.1
Desmospora sp. 8437	ZP_08465087.1	ZP_08465089.1
Dyadobacter crusticola DSM 16708	WP_035333711.1	WP_031529520.1
Dyadobacter crusticola DSM 16708	Manually annotated	WP_031529521.1
Elizabethkingia anophelis NUHP1	WP_024564932.1	WP_024564934.1
Elizabethkingia anophelis NUHP1	Manually annotated	WP_024564935.1
Enterococcus faecalis Fly1	ZP_05578969.1	ZP_05578971.1
Fibrisoma limi BUZ 3	WP_009279746.1	WP_009279749.1
Fibrisoma limi BUZ 3	Manually annotated	CCH51158.1
Frankia sp. CcI3	YP_480236.1	YP_480235.1
Frankia sp. CcI3	YP_480926.1	YP_480927.1
Frankia sp. CcI3	YP_482401.1	YP_482400.1
Frankia sp. CN3	ZP_09165893.1	ZP_09165894.1
Frankia sp. EAN1pec	YP_001507128.1	YP_001507129.1

Frankia sp. EAN1pec	YP_001504430.1	YP_001504429.1
Frankia sp. EAN1pec	YP_001505690.1	YP_001505691.1
Frankia sp. EuI1c	YP_004020784.1	YP_004020783.1
Frankia sp. EUN1f	ZP_06417290.1	ZP_06417291.1
Frankia alni ACN14a	YP_712916.1	YP_712915.1
Geobacillus thermodenitrificans NG80-2	YP_001124395.1	YP_001124397.1
Haliangium ochraceum DSM 14365	Partitioned	YP_003267499
Hymenobacter swuensis DY53	WP_044000691.1	WP_044000684.1
Hymenobacter swuensis DY53	Manually annotated	WP_044000685.1
Kitasatospora setae KM-6054	YP_004907608.1	YP_004907609.1
Kordia algicida OT-1	Partitioned	ZP_02161439.1
Lactococcus lactis 61-14	BAG71480.1	BAG71482.1
Lactococcus lactis 6F3	CAA48381.1	CAA48383.1
Microbispora corallina ADK32556.1	ADK32555.1	ADK32556.1
Microscilla marina ATCC 23134	WP_004155757.1	WP_004155770.1
Micromonospora aurantiaca ATCC 27029	YP_003837207.1	YP_003837208.1
Odoribacter laneus YIT 12061	Partitioned	ZP_09642632.1
Paenibacillus polymyxa SC2	YP_003945785.1	YP_003945783.1
Paenibacillus elgii B69	ZP_09077227.1	ZP_09077225.1
Paenibacillus polymyxa E681	YP_003869832.1	YP_003869828.1
Paenibacillus alvei	Partitioned	ADG29283.1
Pedobacter heparinus DSM 2366	Partitioned	YP_003090840.1
Pedobacter sp. NL19 **(Cluster 8)**	Partitioned	WP_041878755.1
Pedobacter sp. NL19 **(Cluster 14)**	WP_041880930.1	WP_041880928.1
Pedobacter sp. NL19 **(Cluster 14)**	WP_041880943.1	Manually annotated
Pedobacter sp. NL19 **(Cluster 15)**	WP_041881197.1	WP_041881201.1
Pedobacter sp. NL19 **(Cluster 17)**	Partitioned	WP_041882673.1
Pedobacter sp. R20-19	WP_029287207.1	WP_029287206.1
Pontibacter actiniarum DSM 19842	WP_025606750.1	WP_025606771.1
Pontibacter actiniarum DSM 19842	WP_025606763.1	WP_025606773.1
Pseudoalteromonas haloplanktis ANT/505	WP_002962914.1	WP_002962916.1
Pseudoalteromonas luteoviolacea 2ta16	WP_023398949.1	WP_023398948.1
Ruminococcus obeum A2-162	Partitioned	CBL24757.1
Runella limosa DSM 17973	WP_028522718.1	WP_028522717.1
Runella slithyformis DSM 19594	WP_013928182.1	WP_013928181.1
Saccharomonospora marina XMU15	ZP_09744179.1	ZP_09744180.1
Saccharomonospora paurometabolica YIM 90007	ZP_09032097.1	ZP_09032096.1
Segetibacter koreensis DSM 18137	WP_018616399.1	WP_018616400.1
Sphingobacterium sp. H1ai	WP_025141518.1	WP_025141521.1
Spirosoma linguale DSM 74	WP_012929186.1	WP_012929187.1
Spirosoma linguale DSM 74	ADB40692.1	Manually annotated
Spirosoma spitsbergense DSM 19989	WP_020607438.1	WP_020607442.1
Stackebrandtia nassauensis DSM 44728	YP_003514142.1	YP_003514143.1
Staphylococcus caprae C87	ZP_07840598.1	ZP_07840600.1
Staphylococcus epidermidis	CAA90025.1	CAA90026.1
Staphylococcus epidermidis	CAA74350.1	CAA74351.1

Staphylococcus epidermidis Tu 3298	CAA44253.1	CAA44254.1
Staphylococcus gallinarum	ABC94903.1	ABC94904.1
Staphylococcus aureus subsp. aureus ED133	ADI98309.1	ADI98308.1
Staphylococcus aureus subsp. aureus	YP_494457.1	YP_494456.1
Streptomyces clavuligerus ATCC 27064	ZP_08218308.1	ZP_08218309.1
Streptomyces clavuligerus ATCC 27064	ZP_06771940.1	ZP_06771941.1
Streptomyces sp. W007	ZP_09400509.1	ZP_09400510.1
Streptomyces ambofaciens ATCC 23877	CAJ88053.1	CAJ88054.1
Streptomyces bingchenggensis BCW-1	YP_004965200.1	YP_004965201.1
Streptomyces cattleya NRRL 8057	YP_004910077.1	YP_004910078.1
Streptomyces coelicolor A3(2)	NP_624599.1	NP_624600.1
Streptomyces griseoflavus Tu4000	ZP_07315085.1	ZP_07315084.1
Streptomyces griseus subsp. griseus NBRC 13350	YP_001825359.1	YP_001825358.1
Streptomyces lividans TK24	ZP_06533438.1	ZP_06533437.1
Streptomyces scabiei 87.22	YP_003488857.1	YP_003488858.1
Streptomyces sp. Mg1	ZP_04997226.1	ZP_04997227.1
Streptomyces violaceusniger Tu 4113	YP_004811344.1	YP004811345.1
Streptomyces violaceusniger Tu 4113	YP_004815225.1	YP_004815226.1
Streptococcus dysgalactiae subsp. equisimilis	Partitioned	YP_002996025.1
Streptococcus mutans	AAF99579.1	AAF99580.1
Streptococcus pasteurianus ATCC 43144	YP_004559249.1	YP_004559248.1
Streptococcus pyogenes MGAS10270	YP_598531.1	YP_598530.1
Streptococcus salivarius	AEX55164.1	AEX55161.1
Streptococcus sanguinis SK408	EGF18911.1	EGF18912.1
Streptococcus uberis 42	ABA00879.1	ABA00881.1
Thermomonospora curvata DSM 43183	YP_003302206.1	YP_003302207.1
Tannerella forsythia ATCC 43037	Partitioned	YP_005013067.1

Table 12| Accession numbers of LanM enzymes used in this thesis.

Strain	LanM
Actinoplanes garbadinensis	ACR33053.1
Bifidobacterium angulatum DSM20098	ZP_04448254.1
Bifidobacterium longum DJO10A	WP_041920958.1
Clavibacter michiganensis NCPPB 382	YP_001222710.1
Geobacillus thermodenitrificans NG80–2	YP_001126159.1
Mycobacterium marinum M	YP_001849230.1
Saccharopolyspora erythraea NRRL 2338	YP_001106583.1
Streptomyces cinnamoneus	CAD60521.1
Streptomyces griseus subsp. griseus NBRC13350	YP_001826321.1

Table 13| Accession numbers of LanL enzymes used in this thesis.

Strain	LanL
Actinoplanes sp. SE50/110	AEV85426.1
Catenulispora acidiphila DSM 44928	YP_003113256.1
Saccharopolyspora erythraea NRRL 2338	YP_001106807.1
Saccharopolyspora erythraea NRRL 2338	YP_001106221.1
Stackebrandtia nassauensis DSM 44728	YP_003509405.1
Streptomyces clavuligerus ATCC 27064	ZP_05005408.1
Streptomyces griseus subsp. griseus NBRC 13350	YP_001821664.1
Streptomyces venezuelae ATCC 10712	AEA03262.1
Thermomonospora curvata DSM 43183	YP_003298015.1

Table 14| Accession numbers of LanCL enzymes used in this thesis.

Strain	LanCL
Homo sapiens (LanCL1 isoform X1)	XP_005246300.1
Homo sapiens (LanCL2)	NP_061167.1

Table 15| Accession numbers of LanT enzymes used in this thesis.

Strain	LanT
Actinoplanes garbadinensis ATCC 31049	ACR33057.1
Bacillus amyloliquefaciens Y2	WP_014419286.1
Bacillus licheniformis DSM 13	WP_003186373.1
Bacillus sp. HIL-Y85/54728	CAB60262.1
Chryseobacterium sp. StRB126	WP_045499357.1
Chryseobacterium vrystaatense str. LMG 22846	WP_034747466.1
Dyadobacter crusticola DSM 16708	WP_031529510.1
Elizabethkingia anophelis NUHP1	WP_029728189.1
Fibrisoma limi BUZ 3	WP_009279745.1
Geobacillus thermodenitrificans NG80-2	WP_011887656.1
Hymenobacter swuensis DY53	WP_044000683.1
Kordia algicida OT-1	WP_040559592.1
Lactococcus lactis subsp. lactis	AAC72259.1
Microbispora corallina NRRL 30420	ADK32558.1
Paenibacillus polymyxa SC2	WP_013373084.1
Pedobacter heparinus DSM 2366	WP_012780734.1
Pedobacter sp. NL19 **(Cluster 8)**	WP_041878749.1
Pedobacter sp. NL19 **(Cluster 14)**	WP_041880933.1
Pedobacter sp. NL19 **(Cluster 15)**	WP_041881190.1
Pedobacter sp. NL19 **(Cluster 17)**	WP_041882670.1

Pedobacter sp. R20-19	WP_029287209.1
Pontibacter actiniarum DSM 19842	WP_025606756.1
Pseudoalteromonas luteoviolacea 2ta16	WP_023398947.1
Runella limosa DSM 17973	WP_028522719.1
Runella slithyformis DSM 19594	AEI48872.1
Segetibacter koreensis DSM 18137	WP_018616398.1
Spirosoma linguale DSM 74	WP_012929183.1
Streptococcus macedonicus ACA-DC 198	ABI30230.1
Streptococcus pyogenes FF22	AAB92603.1
Streptococcus salivarius K12	ABI63630.1
Streptococcus salivarius M18	WP_004183550.1
Streptomyces cinnamoneus DSM 40005	CAD60523.1
Tannerella forsythia 92A2	WP_041590508.1

Appendix 2. Structural genes accession numbers

Table 16| Accession numbers of the structural genes used in this thesis.

Strain	LanA
Chitinophaga pinensis DSM 2588	ACU58934.1
Chitinophaga pinensis DSM 2588	WP_012789111.1
Chryseobacterium sp. JM1	WP_034729675.1
Chryseobacterium sp. StRB126	WP_045499348.1
Chryseobacterium vrystaatense str. LMG 22846	WP_034747482.1
Chryseobacterium vrystaatense str. LMG 22846	WP_034747485.1
Dyadobacter crusticola DSM 16708	WP_031529516.1
Dyadobacter crusticola DSM 16708	WP_031529517.1
Dyadobacter crusticola DSM 16708	WP_031529518.1
Elizabethkingia anophelis NUHP1	WP_024564936.1
Elizabethkingia anophelis NUHP1	WP_024564937.1
Elizabethkingia anophelis NUHP1	WP_024564938.1
Elizabethkingia anophelis NUHP1	WP_024564939.1
Elizabethkingia anophelis NUHP1	WP_024564940.1
Elizabethkingia anophelis NUHP1	WP_024564941.1
Elizabethkingia anophelis NUHP1	WP_024564942.1
Elizabethkingia anophelis NUHP1	WP_024564943.1
Elizabethkingia anophelis NUHP1	WP_024564944.1
Fibrisoma limi BUZ 3	CCH51157.1
Hymenobacter swuensis DY53	WP_044000688.1
Hymenobacter swuensis DY53	WP_044000689.1
Hymenobacter swuensis DY53	WP_044000690.1
Kordia algicida OT-1	EDP96588.1
Lactococcus lactis 6F3	WP_014570405.1
Microscilla marina ATCC 23134	Manually annotated
Pedobacter heparinus DSM 2366	ACU02776.1
Pedobacter sp. NL19 **(Cluster 8)**	WP_041878759.1
Pedobacter sp. NL19 **(Cluster 14)**	Manually annotated
Pedobacter sp. NL19 **(Cluster 15)**	WP_041881204.1
Pedobacter sp. NL19 **(Cluster 15)**	WP_041881206.1
Pedobacter sp. NL19 **(Cluster 17)**	WP_041882676.1
Pedobacter sp. R20-19	WP_029287195.1
Pedobacter sp. R20-19	WP_029287197.1
Pedobacter sp. R20-19	WP_029287199.1
Pontibacter actiniarum DSM 19842	WP_025606758.1
Pontibacter actiniarum DSM 19842	WP_025606766.1
Pontibacter actiniarum DSM 19842	WP_025606767.1
Pseudoalteromonas haloplanktis ANT/505	EGI71407.1
Pseudoalteromonas luteoviolacea 2ta16	Manually annotated
Runella limosa DSM 17973	WP_028522715.1

Runella limosa DSM 17973	WP_028522716.1
Runella limosa DSM 17973	Manually annotated
Runella slithyformis DSM 19594	WP_013928188.1
Segetibacter koreensis DSM 18137	WP_018616401.1
Sphingobacterium sp. H1ai	WP_025141519.1
Sphingobacterium sp. H1ai	WP_025141520.1
Sphingobacterium sp. H1ai	Manually annotated
Spirosoma linguale DSM 74	ADB40684.1
Spirosoma linguale DSM 74	ADB40686.1
Spirosoma spitsbergense DSM 19989	Manually annotated
Tannerella forsythia ATCC 43037	WP_041590498.1
Tannerella forsythia ATCC 43037	WP_041590501.1

Appendix 3. Statistical support for the PedB enzymes phylogeny

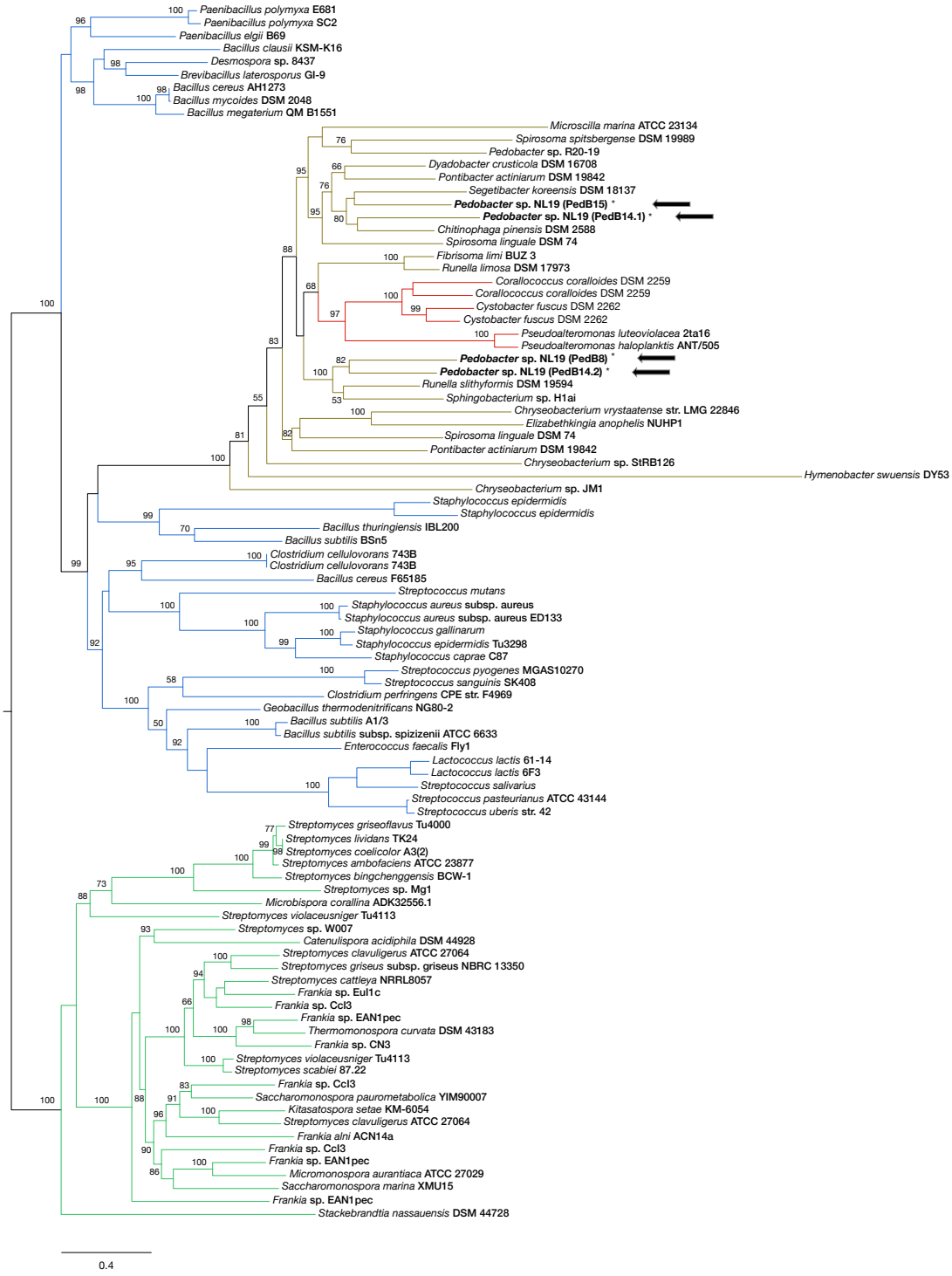


Figure 26| Statistical support for the Bayesian phylogenetic tree of PedB enzymes.
Only support values with aLRT>50% are numbered and shown above or below the lines. Bacteroidetes are shown with yellow lines, proteobacteria with red lines, firmicutes with blue lines and actinobacteria with green lines. Black arrows indicate the *Pedobacter* sp. NL19 LanBs position.

Appendix 4. Statistical support for the PedC enzymes phylogeny

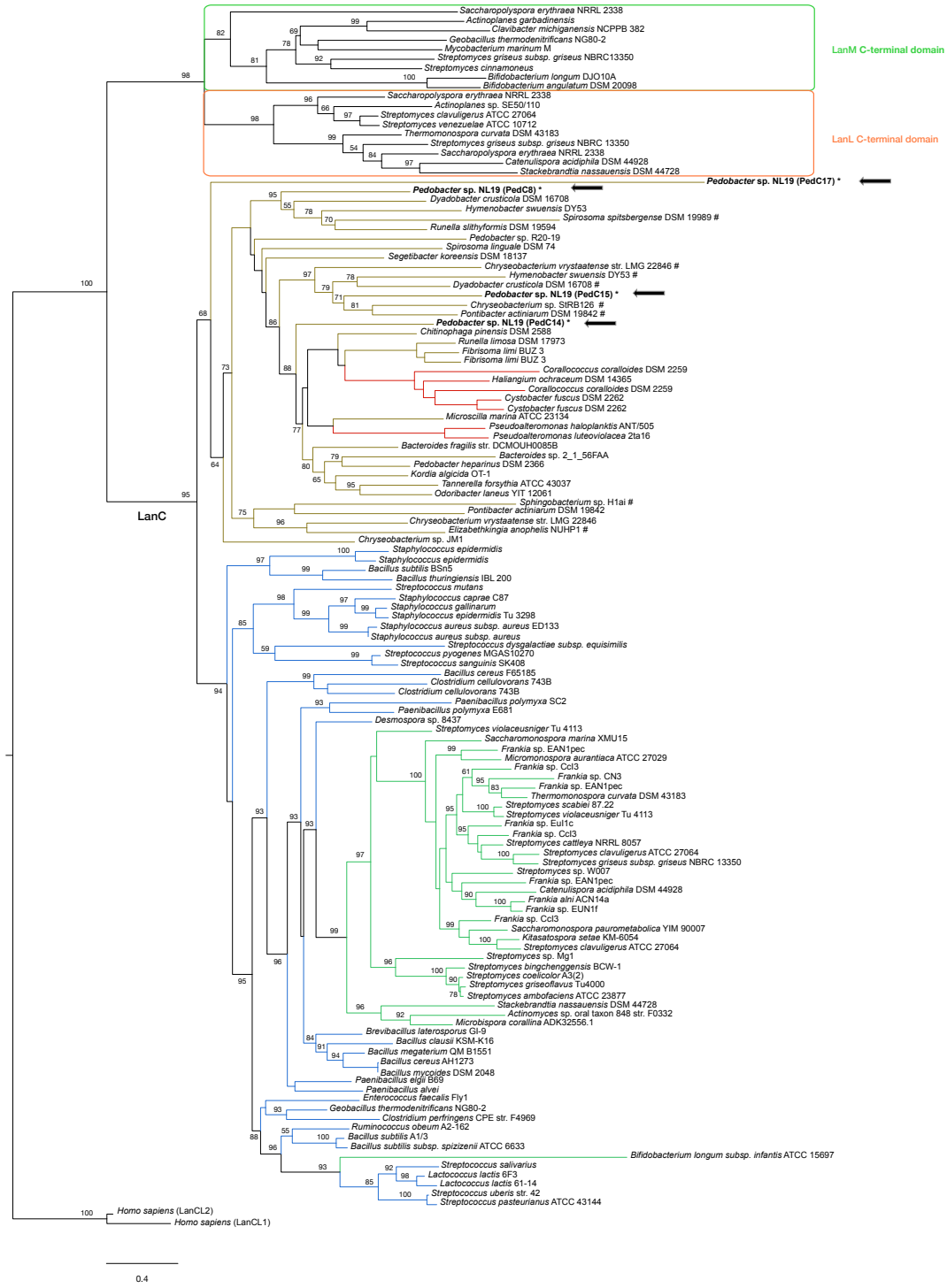


Figure 27| Statistical support for the Bayesian phylogenetic tree of PedC enzymes.
Only support values with aLRT>50% are numbered and shown above or below the lines. Bacteroidetes are shown with yellow lines, proteobacteria with red lines, firmicutes with blue lines and actinobacteria with green lines. Black arrows indicate the *Pedobacter* sp. NL19 LanCs position. The symbol cardinal (#) represents LanCs without zinc-binding motif.

Appendix 5. Statistical support for the PedT proteins phylogeny

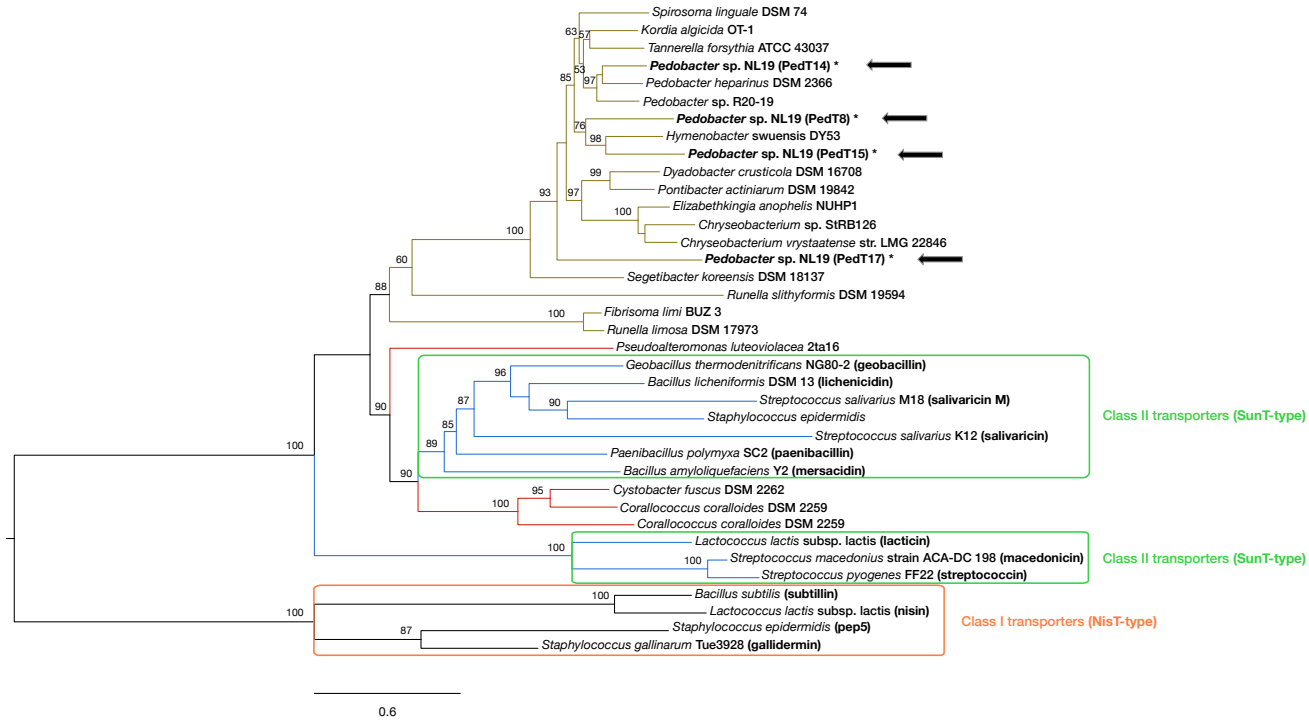


Figure 28| Statistical support for the Bayesian phylogenetic tree of PedT proteins.
Only support values with aLRT>50% are numbered and shown above or below the lines. Bacteroidetes are shown with yellow lines, proteobacteria with red lines, firmicutes with blue lines and actinobacteria with green lines. Black arrows indicate the *Pedobacter* sp. NL19 LanTs position.

Appendix 6. Predicted masses for PedAs

Table 17| Predicted masses for the NL19 mature peptides.

Coloured residues (blue) show the number of dehydrations, correspondent to dH_2O molecules (≈18 Da). Cysteine residues for formation of thioether rings are shown underlined (red).

Core peptide	Monoisotopic mass (Da)	Sequence
PedA8	3270.65	KAALGDVCVFKSSSCTIKPTEPTQPTQPTTTA
	3252.65	KAALGDVCVFKSSSCTIKPTEPTQPTQPTTTA
	3234.65	KAALGDVCVFKSSSCTIKPTEPTQPTQPTTTA
	3216.65	KAALGDVCVFKSSSCTIKPTEPTQPTQPTTTA
	3198.65	KAALGDVCVFKSSSCTIKPTEPTQPTQPTTTA
	3180.65	KAALGDVCVFKSSSCTIKPTEPTQPTQPTTTA
	3162.65	KAALGDVCVFKSSSCTIKPTEPTQPTQPTTTA
	3144.65	KAALGDVCVFKSSSCTIKPTEPTQPTQPTTTA
	3126.65	KAALGDVCVFKSSSCTIKPTEPTQPTQPTTTA
PedA14	2580.28	IPVTTATRRTCCDLTDLICPEVQS
	2562.28	IPVTTATRRTCCDLTDLICPEVQS
	2544.28	IPVTTATRRTCCDLTDLICPEVQS
	2526.28	IPVTTATRRTCCDLTDLICPEVQS
PedA15.1	2974.25	AHANNMSCITGSKSCPNCKEAEREDYN
	2956.25	AHANNMSCITGSKSCPNCKEAEREDYN
PedA15.2	3611.61	KNAYNMSCITGSNLSCQSAKCDPEESLDNLIFGN
	3593.61	KNAYNMSCITGSNLSCQSAKCDPEESLDNLIFGN
	3575.61	KNAYNMSCITGSNLSCQSAKCDPEESLDNLIFGN
	3557.61	KNAYNMSCITGSNLSCQSAKCDPEESLDNLIFGN
PedA17	2979.41	LGNVSEPTHTEPTDDHHTPAIKCTTVIC
	2961.41	LGNVSEPTHTEPTDDHHTPAIKCTTVIC
	2943.41	LGNVSEPTHTEPTDDHHTPAIKCTTVIC
	2925.41	LGNVSEPTHTEPTDDHHTPAIKCTTVIC
	2907.41	LGNVSEPTHTEPTDDHHTPAIKCTTVIC
	2889.41	LGNVSEPTHTEPTDDHHTPAIKCTTVIC

Figure 29| Sequences of the core peptides identified in *Pedobacter* sp. NL19 genome.
The arrow represents the cysteine residues (red) essential for thioether rings formation. The possible dehydrated residues are shown in blue.