UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática

Igor Andrade Figueiredo de Souza

Development of a computational tool to screen antimicrobial peptides in metagenomics data associated to a new vector for high-scale production of lasso peptides

Belo Horizonte
2020

Igor Andrade Figueiredo de Souza

Development of a computational tool to screen antimicrobial peptides in metagenomics data associated to a new vector for high-scale production of lasso peptides

Final Version

Dissertation submitted to the Bioinformatics Graduate Program from Universidade Federal de Minas Gerais as requirement for Master of Science Degree in Bionformatics

Advisor: Dr. TiagoAntônio de Oliveira Mendes
Co-Advisor: Dr. Hilário Cuquetto Mantovani

Belo Horizonte
2020

043 Souza, Igor Andrade Figueiredo de.
Development of a computational tool to screen antimicrobial peptides in metagenomics data associated to a new vector for high-scale production of lasso peptides [manuscrito] / Igor Andrade Figueiredo de Souza. – 2020.
52 f. : il. ; 29,5 cm.

Orientador: Dr. Tiago Antônio de Oliveira Mendes. Co-orientador: Dr. Hilário Cuquetto Mantovani.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa Interunidades de Pós-Graduação em Bioinformática.

1. Biologia Computacional. 2. Anti-Infecciosos. 3. Genômica. 4. Bioprospecção. 5. Peptídeos. I. Mendes, Tiago Antônio de Oliveira. II. Mantovani, Hilário Cuquetto. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Fabiane C M Reis – CRB 6 - 2680

UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DE DEFESA DE DISSERTAÇÃO

IGOR ANDRADE FIGUEIREDO DE SOUZA

Às oito horas do dia **30 de outubro de 2020**, reuniu-se, através do aplicativo Zoom, a Comissão Examinadora de Dissertação, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho do discente Igor Andrade Figueiredo de Souza, intitulado: **"Development of a computational tool to screenantimicrobial peptides in metagenomics dataassociated to a new vector for high-scale production oflasso peptides"**, requisito para obtenção do grau de Mestre em **Bioinformática.** Abrindo a sessão, o Presidente da Comissão, **Dr. Tiago Antonio de Oliveira Mendes**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	**Instituição**	**Indicação**
Dr. Tiago Antonio de Oliveira Mendes	Universidade Federal de Viçosa	Aprovado
Dra. Raquel Cardoso de Melo Minardi	Universidade Federal de Minas Gerais	Aprovado
Dra. Sabrina de Azevedo Silveira	Universidade Federal de Viçosa	Aprovado
Dra. Danielle Biscaro Pedrolli	Universidade Estadual Paulista	Aprovado

Pelas indicações, o candidato foi considerado: **Aprovado**

O resultado final foi comunicado publicamente ao candidato pelo Presidente da Comissão. Nada mais havendo a tratar, o Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 30 de outubro de 2020.

Dr. Tiago Antonio de Oliveira Mendes - Orientador

Dra. Raquel Cardoso de Melo Minardi

Dra. Sabrina de Azevedo Silveira

Dra. Danielle Biscaro Pedrolli

Documento assinado eletronicamente por **Sabrina de Azevedo Silveira**, **Usuário Externo**, em 30/10/2020, às 11:05, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Danielle Biscaro Pedrolli**, **Usuário Externo**, em 30/10/2020, às 11:46, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Tiago Antônio de Oliveira Mendes**, **Usuário Externo**, em 30/10/2020, às 11:46, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Raquel Cardoso de Melo Minardi**, **Professora do Magistério Superior**, em 30/10/2020, às 12:07, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **0342433** e o código CRC **60A92D41**.

Referência: Processo nº 23072.237971/2020-17 SEI nº 0342433

Acknowledgments

I would like to thank both laboratories where I developed this study, Laboratório de Microbiologia Oral e Anaeróbios (UFMG) and Laboratório de Biotecnologia Molecular (UFV), both full of competent coworkers and amazing friends.

I also would like to thank CNPq, CAPES, FAPEMIG, and Bill and Melinda Gates Foundation to provide means to make possible this work.

A special thanks to Professor Tiago Mendes, for guiding me through this journey, not only in pursuit of knowledge but also to improve myself as a professional.

And for last, but not least, my dear friends Natália Guimãraes, Renato Senra and Higor Sette. Not only for taught me protocols but sharing the weight of an academic's life, a terrifying and wonderful new world.

Resumo

O surgimento e o desenvolvimento da resistência antimicrobiana (AMR) contra os antimicrobianos convencionais são os principais temores da medicina moderna e da segurança alimentar, levando ambas as indústrias farmacêutica e alimentícia a renovar seu interesse na descoberta de produtos naturais microbianos com propriedades antimicrobianas. Com o advento de abundante quantidade de dados metagenômicos disponíveis, contendo sequências de espécies cultivadas e não-cultivadas, o campo da descoberta de produtos naturais está se transformando, e novas estratégias, como a mineração de peptídeos, estão sendo desenvolvidas. Aqui, apresentamos a primeira ferramenta de correspondência de padrões criada para prospectar peptídeo laço diretamente a partir de *short reads* de dados metagenômicos, que atualmente é uma restrição para a abordagem clássica de mineração genômica com base em similaridade de sequências. A ferramenta recebe como entrada arquivos de formato FASTQ ou FASTA e os padrões de consulta. Um teste controle foi realizado em uma comunidade simulada, contendo 27 genomas de produtores conhecidos de peptídeos laço e 9 genomas não produtores. Os padrões de consulta foram obtidos a partir de 35 sequências peptídeos laço de 27 genomas distintos. As sequências foram divididas aleatoriamente em grupo treino, contendo 21 sequências e um grupo teste com 14 sequências. Para o grupo reino, uma matriz de distância foi obtida pela técnica de dimensionamento multidimensional e na separação de três grupos. Para cada grupo, uma sequência consenso foi obtida, e um padrão de consulta foi usado para procurar potenciais peptídeos laço em dados metagenômicos de rúmen, levando a descoberta de 3 novos potenciais peptídeos. Para validar os potenciais peptídeos laço, um vetor de expressão para *E. Coli* contendo genes otimizados para produção de peptídeos laço em alto rendimento foi projetado e sintetizado. O vetor foi capaz de produzir o peptídeo laço microcina J25, com a capacidade de inibir bactérias gram-negativas e gram-positivas. O estudo demonstra o potencial de acessar dados de comunidades bacterianas, um recurso potencialmente único para novos peptídeos antimicrobianos.

Keywords: antimicrobianos, genômica, bioprospecção, triagem computacional, peptídeos laço

Abstract

The emergence and development of antimicrobial resistance (AMR) against conventional antimicrobials are the main fears of modern medicine and food security, leading both pharmaceutical and food industries to renew their interest in the discovery of microbial natural products with antimicrobial properties. With the advent of the massive amount and freely available metagenomic data, containing sequences from cultured and uncultured species, the field of natural product discovery is transforming, and new strategies such as peptide mining are being developed. Here, we present the first pattern-matching tool created to prospect lasso peptide directly from short reads of metagenomic data, which is a constraint for the classical genome mining approach based on sequence similarity. The tool receives as input FASTQ or FASTA format files and query patterns. A control test was performed on a mock community containing 27 genomes from known lasso peptide producers and 9 negative genomes. The query patterns were designed based on 35 lasso sequences associated with 27 genomes. The sequences were randomly divided into a group termed training, containing 21 sequences and a testing group with 14 sequences. For the training group, a distance matrix was obtained by multidimensional scaling technique based on the plot of three groups. For each group, a consensus sequence was obtained, and a query pattern was used to screen potential lasso peptides in rumen metagenomics data resulting in 3 new peptides. To validate the potential new lasso peptides, a user-friendly *E. coli* expression vector containing optimized genes to produce lasso peptides in high yield was designed and synthesized. The vector was able to produce the lasso peptide microcin J25 with the ability to inhibit both gram-negative and gram-positive bacteria. The study presents the potential to access data from whole communities, which are a potentially unique resource for novel antimicrobial peptides.

Keywords: antimicrobials, genomics, bioprospecting, computational screening, lasso peptides

LIST OF FIGURES

LIST OF TABLES

TABLE OF CONTENTS

1. INTRODUCTION

The emergence and development of antimicrobial resistance (AMR) against conventional antimicrobials are the main fears of modern medicine and food security, leading both pharmaceutical and food industries to renew their interest in the discovery of microbial natural products with antimicrobial properties (Rehman et al 2018). This concern has provoked a myriad of international entities and governmental institutions to direct efforts in studying the subject and releasing technical reports. It is projected an increase the current deaths associated with AMR from 700,000 thousand per year to 10 million by 2050 (UK Review, 2016). This estimative can be considered underestimated, mainly because, in in-development countries, the AMR deaths associated are not tracked properly. The Figure 1 shows the comparison among the majors' death causes and the projection for AMR-related deaths.

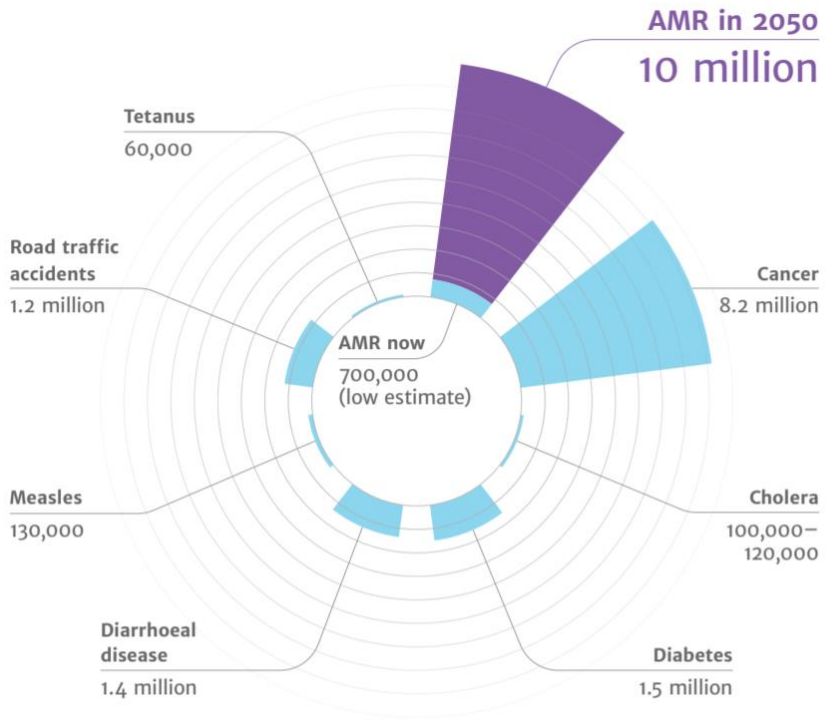


Figure 1 - Estimated deaths worldwide in 2050 and others major causes of deaths. In light blue the current number of deaths per year and in purple the estimated number of AMR deaths. (UK Review, 2016)

The AMR emergence also is going to greatly impact the world economy due to the overload upon healthcare systems and livestock production, leading to losses greater than that caused by the economic recession in 2008, and in this context, the countries in development are going to be the most impacted (World Bank 2017). The World Bank Group projected two possible scenarios, one for high-AMR, a pessimist scenario, and a more optimist scenario with low levels of AMR. In the Figure 2, the livestock production is plotted for both scenarios and comparing between low-income, middle-income, and high-income countries, exposing the bigger impact to the poorest countries.

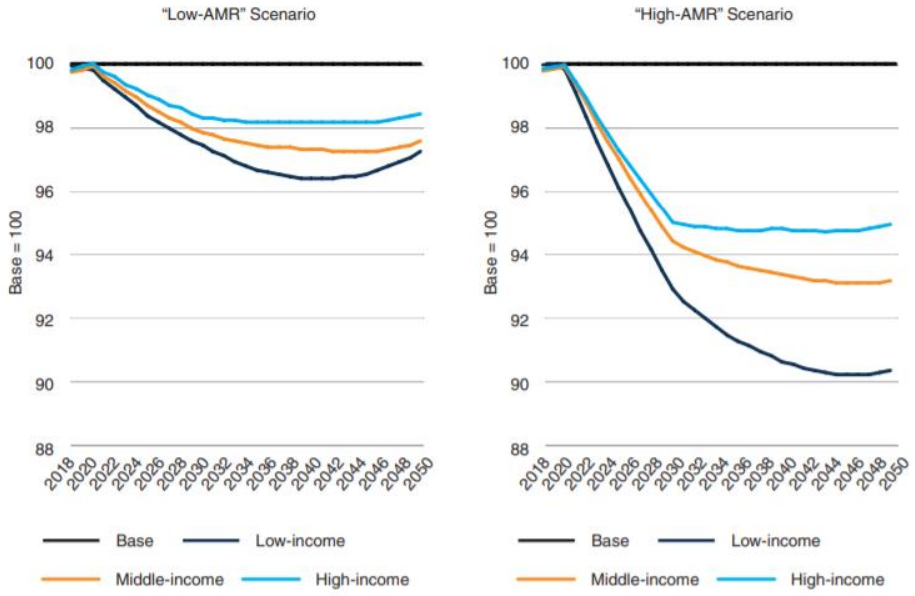


Figure 2 – The decline in livestock production worldwide in two different scenarios of AMR. The Base 100 is the current production. (World Bank 2017)

The graph in the Figure 3 shows the Health-Care costs for both scenarios, optimistic and pessimist. The base line is a scenario without AMR-resistance.

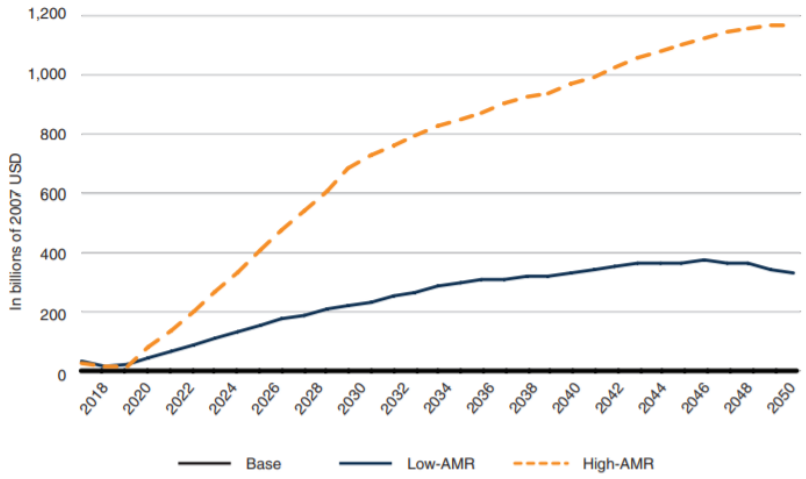


Figure 3 - Health care costs worldwide will increase nearly US$ 1.2 trillion in the 'High-AMR' scenario and US$ 400 billion in the 'Low-AMR' scenario. The Base lien corresponds to the healthcare costs in 2017. (World Bank 2017)

The renewed interest in the discovery of natural products with antimicrobial properties directed the attention to bacteriocins as potential alternatives to currently available antibiotics and chemotherapeutic drugs. Bacteriocins family includes a diversity of proteins and peptides in terms of size, microbial targets, modes of action, and immunity mechanism. Many of them critically differ from traditional antibiotics: they have a relatively narrow killing spectrum and they are only toxic to bacteria closely related to the native producer (Riley, M. and Wertz, E.

2002). This is a critical difference since it has become clear that the administration of broad-spectrum antibiotics can lead to collateral damage to the human commensal bacteria (Cotter, P. et al. 2013). Bacteriocins not only harbor high specificity as an important characteristic, but they also show mechanisms of action that are distinct from current chemotherapeutic drugs, slowing down the development of resistance by the pathogen. Moreover, they are amenable to gene-based engineering due to their proteinaceous nature, which opens new opportunities (Cotter, P. et al. 2013).

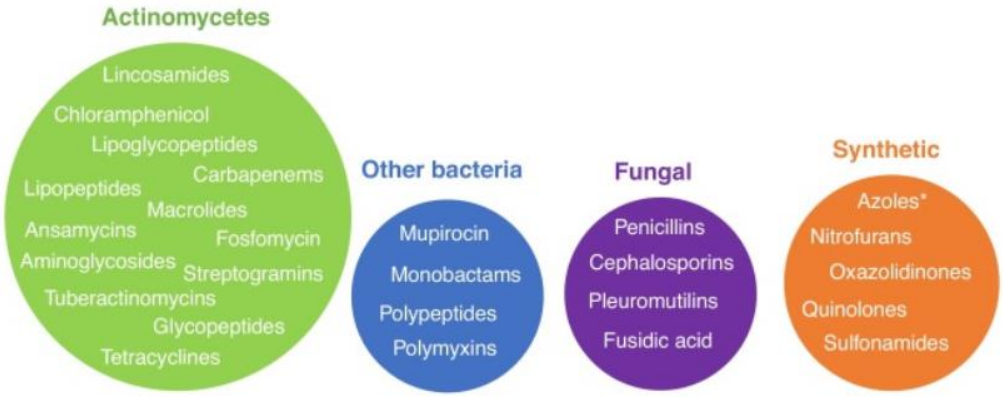


Figure 4 - Most clinically relevant classes of antibiotic are derived from natural products. (Hutchings et al, 2019)

Despite the microbial biosynthesis represent the most important historical source for the development of antibiotics, as exposed above in the Figure 4, the golden age of antibiotic discovery declined after the 1970s (Durand et al 2019). In fact, since the 80s the number of new antibiotics approved by the Food and Drug Administration in the USA declined. As shown in the Figure 5, although the last decade, 15 new antibiotics were approved, 5 of them now has limited accessibility due to the companies which sold them, or they were sold off or declared bankruptcy (McKenna, M. 2020).

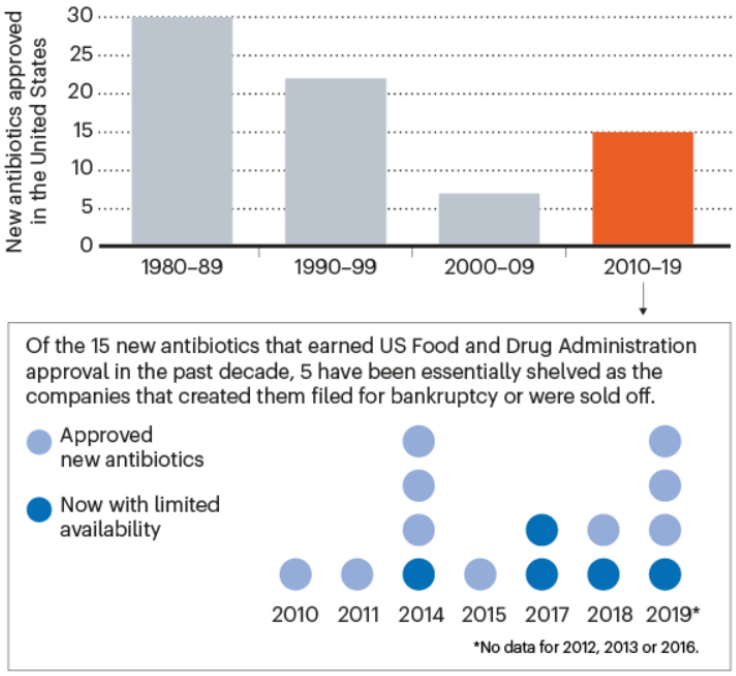


Figure 5 – The number of antibiotics declined over the last decades. (McKenna, M. 2020)

The development of a new antibiotic is an effort estimated at US$ 1.4 bi and with a long path to profitability, as shown in the Figure 5. For almost two decades, the large corporations that once dominated antibiotic discovery have been fleeing the business. For instance, from the 20 pharmaceutical companies that invested in antibiotic discovery in the 1980s, there were only five left by 2015 (Durand, G. et al 2018; McKenna, M. 2020). The investments from venture capital also have declined, and between 2003 and 2013 less than 5% of the venture capital (i.e. around US$ 1.8 bi) invested in the pharmaceutical industry was applied in the development of new antibiotics. The context raises a red flag for innovative ways for the discovery of new drugs and producing them.

The traditional bioassay-guided methods, responsible for the great number of discovered antibiotics in the golden age (in the 1970s) have drastically changed in recent years, with the abundant availability of genomic sequences. One problem of the traditional screening methods is the frequent rediscovery of known compounds, making them increasingly unappealing and economically disadvantageous (Metelev, M. et al 2015; Baltz, H. et al 2006). The critical change in the last years was the use of genomic information to guide the natural product discovery process, greatly reducing the trial and error efforts and increasing the use of computational methods. This change was possible due to reduction of the sequencing cost by high-through technologies developed in the last decades, making possible not only the generation of but the access to abundant freely genomic data (Maksimov, M. et al, 2013; Maksimov, M. et al, 2012a; Cheung-Lee, W. and Link, J. 2019; Wintern J. et al, 2011). Genomic-guided strategies have facilitated prioritization based on predicted novel products and structure elucidation. One of these new strategies is genome-mining, which involves searching in genomes for genes or gene clusters indicative of novel natural products.

Bacteriocins belonging to the group of the ribosomally synthesized and post-translationally modified peptides (RiPPs) are often a group of natural products targeted by genome-mining approaches. They are gene-encoded peptides with unique characteristics that are exploited during genomic identification: their biosynthetic gene clusters (BGCs) are relatively small, RiPPs precursor is typically encoded near the modification enzymes, RiPP precursors have separate sites for enzyme-binding and modification (referred to as the leader and core regions, respectively). The unmodified leader region is proteolytically removed during maturation. This provides a facile route to evolving new natural products (NPs) since RiPP enzymes can be highly selective for a particular leader sequence, but promiscuously process many core sequences (Tietz et al, 2017). A RiPP that gained attention in the last years is the lasso peptides.

Lasso peptides show a singular topology that is reminiscent of a lariat knot. This so-called lasso fold is accomplished by the threading of the linear C-terminal tail through the N-terminal macrolactam ring. The macrolactam ring is formed by condensation of the N-terminal α-amino group with the carboxylic acid side chain of an Asp or Glu residue at positions 7-9 (Hegemann, J. et al, 2019; Martin-Gómez and Tulla-Puche, 2018). The Figure 6 is shown all the topologies lasso peptide assumes until its final lasso fold. The formed topology is predominantly stabilized by steric interactions, accomplished by the presence of bulky residues, also called plug residues, above and below where the C-terminal tail thread the ring. The stabilization also is sometimes, but not necessarily, assisted by the presence of disulfide bridges, and the presence/absence and a number of these divide the lasso peptide group into four classes (Maksimov, M. et al 2012; Hegemann, J. et al 2015; Martin-Gómez, H. and Tulla-Puche, J., 2018).

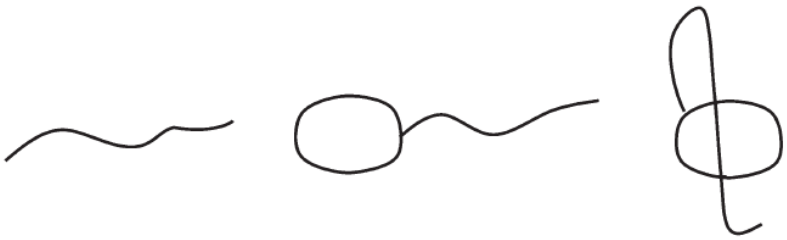


Figure 6 – From left to right: linear peptide, branched-cyclic peptide, and lasso peptide. (Adapted from Martin-Gómez and Tulla-Puche, 2018)

Class I lasso peptides have two disulfide bonds, one involves the N-terminal Cys and the second connects the ring to the tail. Class II presents no disulfide bonds, and the stabilization is only due to the plugs. Class III and Class IV both have only one disulfide bond, in the former connecting the ring to the tail and in the latter, the disulfide bond is present in the tail. (Martin-Gómez and Tulla-Puche, 2018; Hegeman, J. et al, 2019). Figure 7 below shows the structure of the four classes.

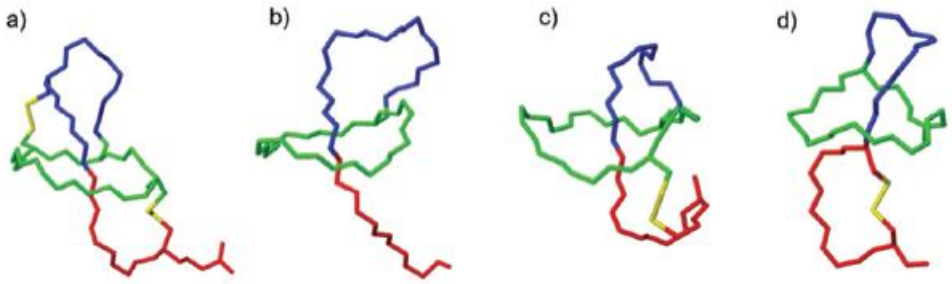


Figure 7 – (a) Class I, (b) Class II, (c) Class III, (d) Class IV. The ring residues are shown in green, amino acids belonging to the loop in blue, and the amino acids in the tail in red. Disulfide bridges are shown in yellow. (Adapted from Martin-Gómez and Tulla-Puche, 2018)

The rigid lasso fold can confer high stability against proteolytic degradation and many lasso peptides are known to withstand prolonged incubation at elevated temperatures (Hegemman, J. et al 2019). Microcin J25 (MccJ25) was the first lasso peptide discovered and considered as an archetype of the group. It was shown to exhibit its antimicrobial activity even after it was autoclaved. For a long time was believed that thermal stability was a universal feature of lasso peptides, but in 2013 was discovered the Caulosegnins (Hagemman, J. et al, 2012), group of lasso peptides with thermal sensitivity. Nevertheless, the general stability of these peptides put them as promising protein scaffolds for developing novel biopharmaceuticals (Knappe et al. 2011; Zhao, N. 2016). Regarding the biological activity of these peptides, a broad spectrum of pharmacological effects has been described. To date, they have been reported to act in the management of diabetes, inhibit HIV replication, and target receptors involved in cancer. Moreover, some of these peptides show antimicrobial and receptor antagonist activity (Martin-Gómez and Tulla-Puche, Hegemman et al 2015, Alvarez-Siero et al 2016, Maksimov and Link 2013, Kersten et al 2012).

Lasso peptide biosynthesis is accomplished in three steps: first, the precursor peptide (termed A) is recognized and bound by a so-called RiPP recognition element (RRE) protein. After, a

cysteine protease with homology to transglutaminases cleaves off the leader peptide and releases the core peptide, thus allowing an ATP-dependent macrocyclase (the C protein) with homology to asparagine synthetases to activate the Asp/Glu carboxylic acid in form of an AMP ester before catalyzing the macrolactam formation by condensation with the α-amino group. The RRE protein and cysteine protease are found either as discrete proteins (B1 and B2 proteins) or fused into a single polypeptide (B protein). In this way, a minimum of three genes is necessary for the lasso peptide BGC – or four genes, if B protein is split (Tietz et al. 2017). The Figure 8 shows schematically the biosynthesis path of lasso peptides.

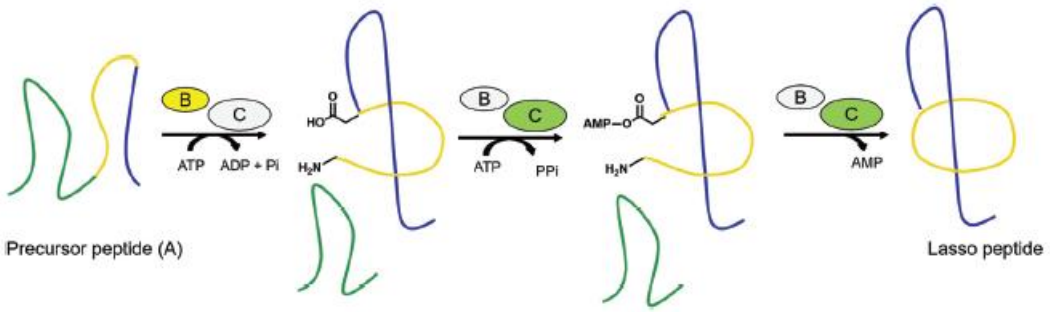
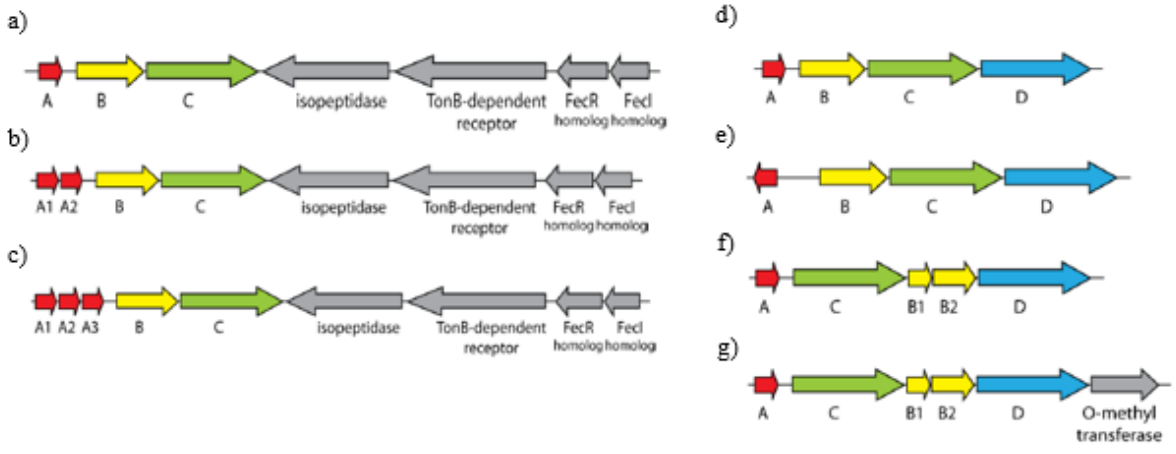


Figure 8 – Proposed mechanism by which lasso peptides mature. Protein B contains the RRE and cysteine protease activity, responsible to cleave off the leader peptide from the core peptide and Protein C catalyzes the isopeptide bonding. (Adapted from Martin-Gómez and Tulla-Puche, 2018).

The first lasso peptide biosynthetic cluster described was for MccJ25, containing the three minimum genes and included a fourth gene encoding an ABC transporter (protein D). The presence of this protein in the BGC is and indicative of antimicrobial activity since this protein confers immunity for the native producer. As novel lasso peptides, BGCs have been discovered, a diversity of biosynthetic clusters was identified. For example, in actinobacteria and firmicutes, splitting-B protein is more present than in proteobacteria. Some clusters also present isopeptidases, and others have kinases. Therefore, it might be possible that the actual ecological roles of some lasso peptides are linked to their genetic surroundings (Hegemann 2015). The Figure 9 presents multiple BGCs for known lasso peptides.



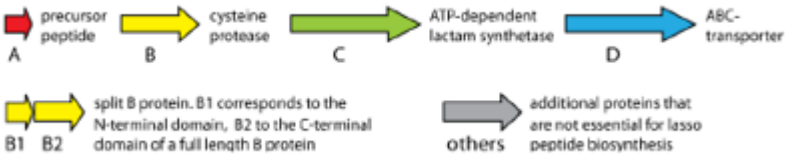


Figure 9 – Clusters organizations for: (a) astexin-1, rhodanodin, rubrivinodin, sphingonodin I, sphingonodin II, sphinpyxin I, sphingopyxin II, syanodin I, xanthomonin III, zucinodin (b) astexins-2/3, caulonidins IV/V, caulonidins VI/VII, xanthomonins (c) caulonodins I-III, caulosegnins I-III (d) capistruin, burhizin (e) microcin J25 (f) lariatin, SRO15-2005, SSV-2083, streptomonomicin (g) lassomycin. (Adapted from Hegemman et al. 2015)

The BCG minimum requirement by the proteins B and C, and the respective homology encountered in these proteins, made possible the shift from the bioassay-guided discovery of lasso peptides to genome-mining guided. Until 2008, all lasso peptides had been isolated using functional screens, activity-driven compound isolations, and purified from culture broths of the native hosts. The discovery of new lasso peptides happened largely by chance. This changed with the report of capistruin, a lasso peptide produced by *Burkholderia thailandensis* that was identified by genome mining. At the time, the microcin J25 cluster was the only know BGC, and the proteins McjB and McjC were used as queries in a BLAST search. Once was found homologous proteins CapB and CapC in the genome sequence from *B. thailandensis*, a manual search was performed to locate putative short ORFs in the flanking regions, following the BCG logic. A 144 bp ORF coding putative lasso peptide precursor protein CapA was located. The putative CapA gene encodes a 47 amino acid precursor, which the first 28 residues were proposed to represent a leader peptide. The core peptide includes the remained 19 amino acids in the C-terminal portion. The 19 residues peptide fragment features a N-terminal Gly and an Asp at position 9, meeting the general primary structure of peptide lassos (Knappe et al. 2008).

The Capistruin discovery established the first strategy (Figure 10) to prospect new lasso peptides, using homology-based genome mining. The sequences of known B and C proteins are used as templates in BLAST searches. Once the results of the queries suggest the putative presence of biosynthetic gene clusters, manually expand the search in the vicinity regions looking for ORFs with the consistent length for lasso peptides and with a primary structure with a fixed Thr in the leader peptide, the presence of an N-terminal Gly in the core peptide and an Asp or Glu in the 7-9 positions. Recently was known that other residues are present in the N-terminal, as Ser, Ala, Cys, Leu (Tietz et al. 2017), and Trp (Koos, J. and Link, A. 2018).

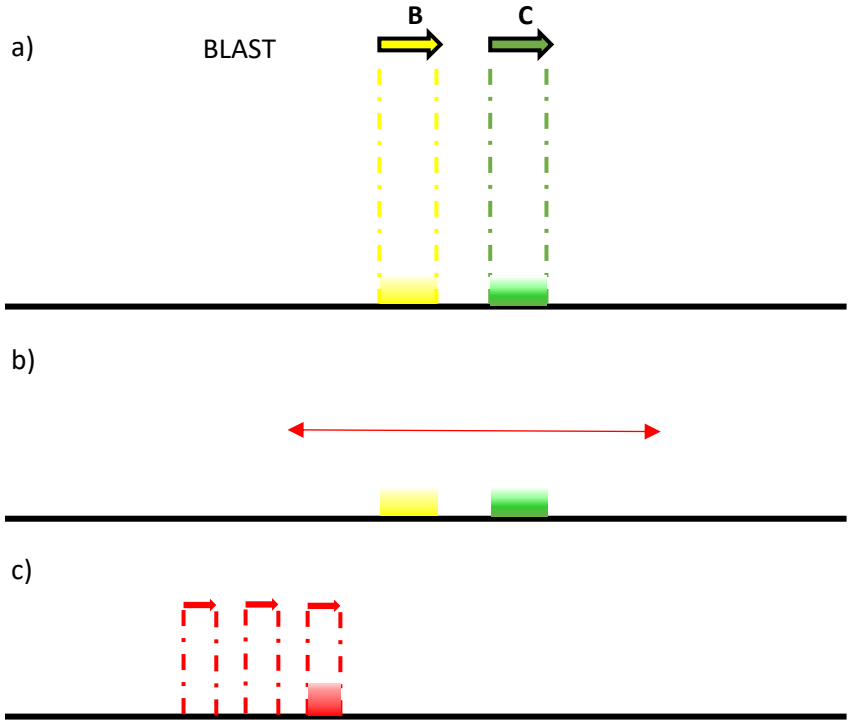


Figure 10 – First genome mining approach applied in the discovery of Capistruin (Knappe et al. 2008). The homology-based approach starts with (a) aligning the sequences of known proteins B and C against a genome, (b) expands the search to vicinity (c) looking for putative precursor sequences.

An alternative strategy based on the precursor-centric genome mining was applied in the discovery of astexin-1, a novel lasso peptide produced by *Asticcacaulis excentricus*. This approach first probes short open reading frames for putative precursor peptides and then surveys the genetic neighborhood for likely maturation enzymes, as shown in Figure 11. Application of such method by (Maksimov et al. 2013) identified 79 putative gene clusters out of 3000 known genomes at the time of the study, distributed across nine bacterial phyla and an archaeal phylum. To be able to efficiently identify precursor genes and differentiating them from false positives is necessary the construction of good precursor patterns targeting the conservation of some residues in the hypervariable sequence, as the conserved Thr in the penultimate position of the leader peptide, a conserved Gly, and the presence of a carboxylic acid residue in the 7-9 positions, along with constraint restrictions on the length of both leader and core peptides.

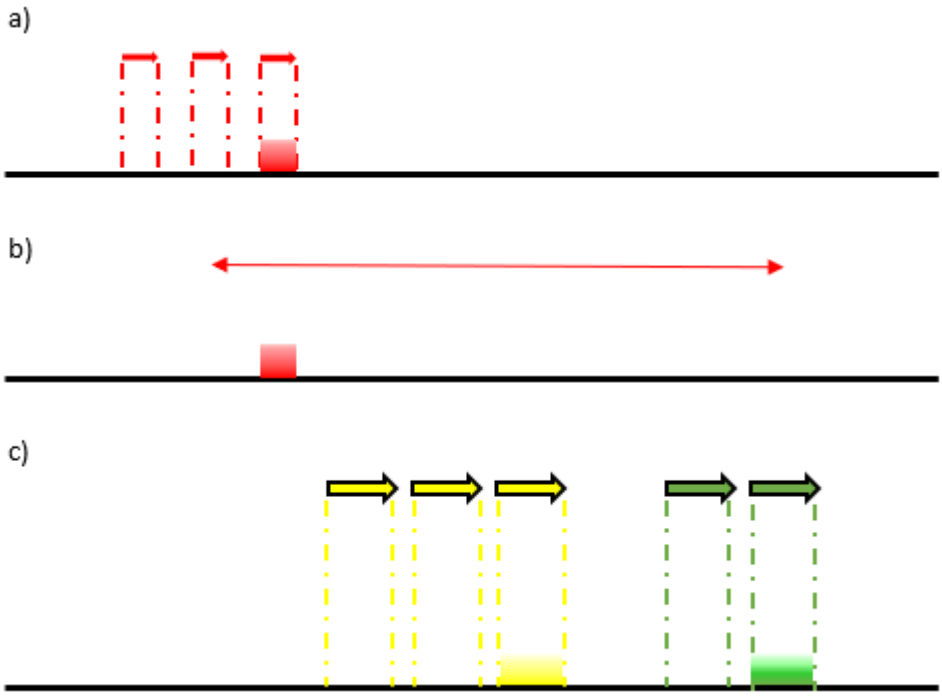


Figure 11 – The precursor-centric approach (Maksimov et al. 2013) looks for (a) putative precursor peptides and then (b) expand the search within the vicinity looking (c) for putative machinery proteins.

More recently, a third approach was developed, leveraging mass spectrometry technology. The goal of this methodology is to connect expressed natural products (chemotype) with their gene clusters (genotype). Figure 12 shows a schematic workflow for this approach. Mass spectrometry-guided genome mining, or natural product peptidogenomics (NPP), uses the data obtained from mass spectrometry, querying the peptide sequences identified against open reading frame translations iteratively while applying the biosynthetic logic of known natural products families (Kersten et al. 2011). This method successfully identified two novel lasso peptides from *Streptomyces* bacteria: one class I (SSV-2083) and the other class II (SRO15-2005) (Maksimov et al 2013, Kersten et al. 2011).

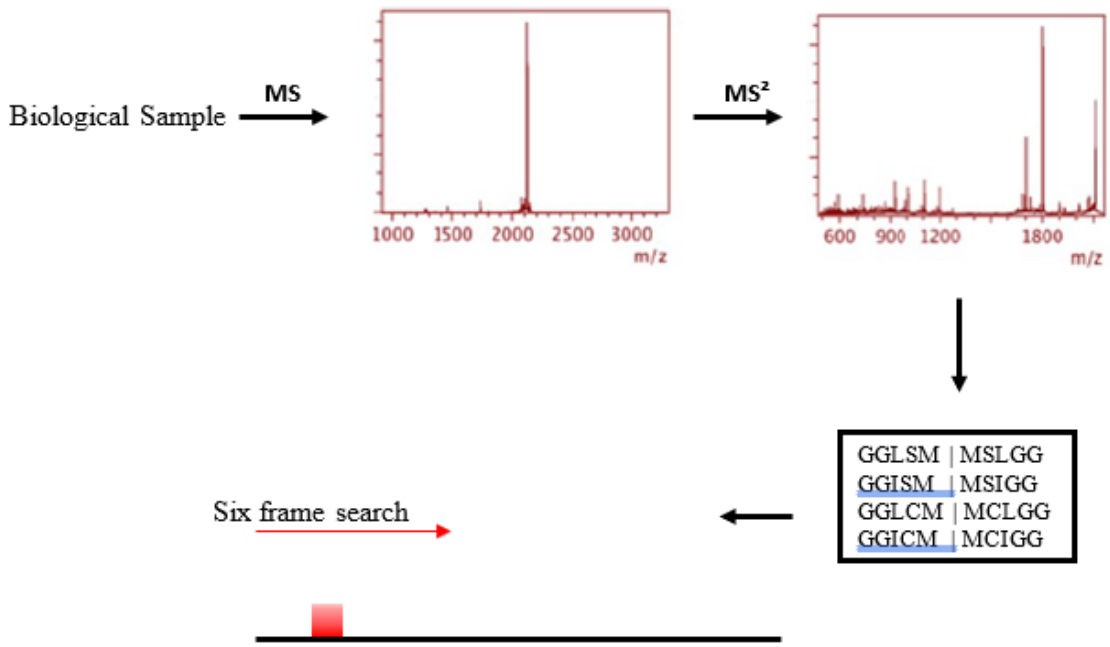


Figure 12 – Mass spectrometry-guided genome mining approach workflow, which uses the information of the peptides fragments to perform alignment queries in the genome context. (Adapted from Maksimov et al. 2013)

It is often the case that the native producers of lasso peptides, identified by genome mining, produce no or exceedingly small amounts of the natural product (Mevaere, 2018; Metelev, M. et al, 2015), necessitating heterologous production. For this, the gene cluster can be modified by the insertion of artificial strong inducible promoters (Knappe, T. et al, 2008; Cheung-Lee, W. and Link, A., 2019). In respect of the heterologous production in *E. coli* of lasso peptides stemming from proteobacteria, it was shown that production could be vastly increased through the exchange of the intergenic region between the genes encoding the precursor peptide and the processing enzyme with an *E. coli* optimized ribosomal binding site (Hegemman et al. 2015).

Natural product discovery including lasso peptide depends on the existence of gene clusters and the accessibility to genetic sequence (Cheung-Lee and Link et al., 2019). However, we are not able to exploit the full biotechnology potential from microbial biosynthesis once most environmental bacteria are as-yet unculturable. On the other hand, an exciting field towards natural product discovery is metagenomics and microbiome mining. In communities bacteriocins play an important ecological role, even not necessarily linked with the growth or development of the host, they are essential components of innate immune defenses, colonization strategy, and intercellular communications (Riley et al., 2002). In specific environments such as bacterial communities under strong competition is common the case of bacteria up-regulation of those genes to kill other bacteria in the community. Therefore, communities constitute unique resources for novel AMP discovery (Oyama et al. 2017; Hutchings et al. 2019).

Metagenomics is a culture-independent approach that seeks to access the biosynthetic capacity of the uncultured majority of bacterial species, by directly capturing DNA from the environment (environmental DNA, eDNA). Subsequently, it is possible to identify, isolate, and express biosynthetic gene clusters in a heterologous host. There are two strategies to explore the biotechnological potential of microbial communities: functional metagenomics and sequence-based metagenomics. The former offers a means of investigating natural products encoded by uncultured bacteria, where DNA extracted directly from environmental samples is cloned into an easily cultured bacterium. The resulting clones are screened for phenotypes associated with

the production of target small molecules. The latter is based on the reconstruction of draft genomes, using genome-resolve methods and profiles them for biosynthetic content to identify high-value targets (Crits-Christoph et al. 2018; Charlop-Powers et al. 2014).

Sequence-based metagenomics imposes a complex problem since its goal is to assembly simultaneously all genomes from the entire mixture of DNA from a different environment, including viral, bacterial, or eukaryotic organisms with different levels of abundance (Ghurye et al. 2016). Less abundant and particularly rare species will yield only incomplete information obtained from DNA fragments from these species (Wooley et al. 2010) making it impossible to obtain a complete genome or even contigs with reasonable size for those species., The screening for biosynthetic gene clusters in this low-quality data is impractical. Another complicating factor is the introduction of artifacts in the assembly process, the confounding effect of repeats is exacerbated by the fact the unrelated genomes may contain nearly identical DNA (inter-genomic repeats). On the other hand, multiple individuals from the same species may harbor small genetic differences – strain variants (Ghurye et al. 2016).

Functional metagenomics provides a complementary approach, isolating clones that are active in heterologous hosts. The DNA extracted directly from an environmental sample is cloned into an easily cultured bacterium. These clones are then examined in detail for the production of clone-specific small molecules (Brady et al. 2009, Iqbal et al. 2016). The size and the heterologous nature of eDNA libraries pose several challenges, also there should be good knowledge about the vectors and type of host required for sustaining the library. Besides, the successful screening is immeasurably dependent on several elemental parameters, including the complexity and composition of the community playing a significant role in building the library, the abundance of genes in the library, which in turns depends on the dominance of specified species in the community, the expression system, cloning host and vector and the pre-expressed and post-translational machinery (Johnson et al., 2017).

Both current strategies to prospect natural products from microbial communities impose their challenges and offer big potentials to the development of novel techniques and tools to exploit the whole biotechnology potential hidden in the communities. As of now, methods for mining lasso peptides include single-input whole-genome analysis (e.g. antiSMASH and BAGEL4), and RiPPquest, a mass spectrometry-based method for connecting BGCs to molecules. However, these tools are not able to directly mine peptides in metagenomics raw data, relying on assembly-based methods, a process that can introduce artifacts in the data or exclude low abundance and rare species from the community. And so, in this study, we propose a new tool to perform lasso peptide mining directly in metagenomic raw data and an expression platform cluster-independent, where one unique biosynthetic machinery is applied to screen bioactive sequences.

2. GOALS

2.3 General Goal

To develop and validate a computational tool and an expression vector to mine lasso peptides directly from raw metagenomic data.

2.4 Specific Goals

2.4.1 Develop a computational tool to prospect lasso peptides in raw, short reads, from metagenome datasets;

2.4.2 Validate the tool against mock communities built from genomes known to encode lasso peptides and genomes known as not producers;

2.4.3 Identify putative lasso peptides in sheep rumen microbiome metagenome data;

2.4.4 To design and validate an expression platform BGC-independent, harboring the biosynthetic machinery of microcin J25 and a flexible region based on Golden Gate Assembly to clone different core sequences.

3. METHODS

3.1 The Algorithm

The algorithm is conceptually simple: its proposed problem is to receive a query pattern and short *reads* from metagenomic sequencing raw data, in either FASTA or FASTQ format. The query pattern is modeled as non-deterministic finite automata (NDFA) represented by the 5-tuple (Q, ∑, δ, q0, F) where Q is a finite set of states, ∑ is the alphabet, composed by a set of finite symbols representing the 20 amino acids and the stop codon, δ it the transition function, q0 is the initial state from where the read is processed, and F is a set of final states, in this case, the only accepted final state is a stop codon. Graphically. the NDFA is represented by a directed graph, where its vertices are the states, the edges represent a transition consuming symbols from the alphabet. The initial state is denoted by an empty incoming edge and the final state is indicated by a double circle, the representation below is the NDFA for a query pattern TXGX[DE]*, where the symbol X indicates the transition happens using any symbol from the alphabet. The amino acids inside the square brackets indicate there are two possible transitions – for Glu or Asp – at that position. The * symbol is the stop codon, the only acceptable transition from a final state. Figure 12 is schematically represented the NDFA for the query pattern cited.

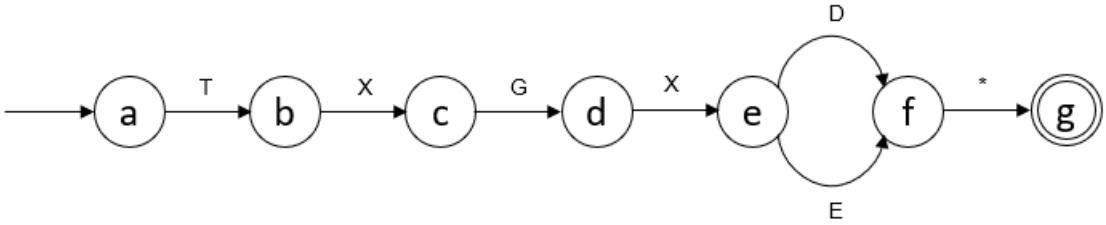


Figure 13 – Schematic finite non-determinist automata representing the pattern T-X-G-X-[DE]-*.

3.2 The Implementation

The algorithm was implemented for both operational systems (OS): Unix-based and Windows 10. The Unix-based implementation is only a command line, and used Python 2.7.3 and the library *multiprocessing*, to have as feature multithreading.

For Windows 10 implementation, a graphic user interface was implemented using IronPython 2.0 and .NET version 4.0 framework. IronPython 2.7.8 is one of a pool of languages capable to run upon the Common Language Runtime. The IronPython script was embedded with the same script used in the Unix-based system implementation. Finally, the IronPython was embedded in a C# script, to make it a Windows Application. Figure 14 below shows schematically the Windows 10 implementation.

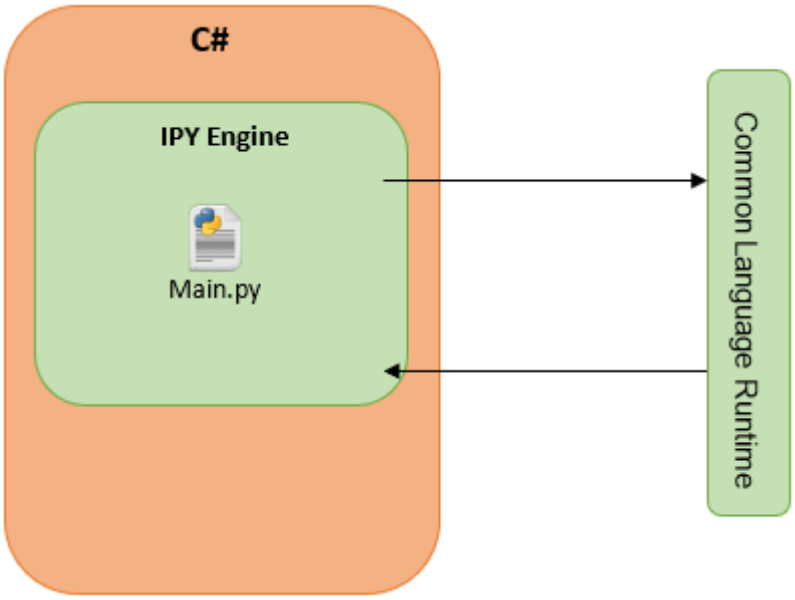


Figure 14 – The embedding process, where the python script interacts with the .NET framework.

3.3 Sequence Logo Generation

Sequence logos were generated using WebLogo 2.8.2 (https://weblogo.berkeley.edu/). The lasso peptide sequences were first aligned manually.

3.4 Mock Files Generation

To validate the ability of the program to recovery true positives lasso sequences, 25 genomes (shown at Table 1) known to encode 35 lasso peptides were used to simulate the mock community along with 9 genomes from bacteria known to not produce lasso peptides (show at Table 2). Using an *in-house* script based on sliding windows all genomes were fragmented in short reads of size 100 pb, and each one tagged with a label in the FASTQ format.

Table 1 – Known genomes to produce lasso peptides and the NCBI accession number.

Positive Genomes	**NCBI Accession Number**
Asticcaulis excentricus CB 48	NC_014816
Phenylobacterium zucineum HLK1	NC_011144
Burkholderia rhizoxinica HKI 45	FR687359
Burkholderia thailandensis E264	NC_007651
Caulobacter sp. K31	NC_010338
Caulobacter segnis ATCC 21756	NZ_CP027850
Escherichia coli strain U44	NZ_LSES00000000
Nocardiopsis alba DSM 43377	NZ_ANAC00000000

Thermobifida fusca	NC_007333
*Rhodococcus jostii K01−B0171**	AB593691
Rubrivivax gelatinosus IL44	NC_017075
Streptomonospora alba strain YIM 90003	NZ_JROO00000000
Sphingobium japonicum UT262	NC_014006
Sphingopyxis alaskensis RB2256	NC_008048
Streptomyces sp. 46	NZ_MTHG00000000
Streptomyces sioyaensis strain DSM 40032	NZ_SDIF00000000
Streptomyces cattleya str. NRRL 8057	FQ859185
Streptomyces davawensis strain JCM 4913	HE971709
Streptomyces griseorubens strain JSD-1	NZ_KL503830
Streptomyces leeuwenhoekii	NZ_LN831790
Streptomyces nodosus strain ATCC 14899	CP009313
Streptomyces mirabilis strain OK461	NZ_FONR00000000
Streptomyces sp. Tue6075	NZ_CP010833
Streptomyces variabilis	CP040941
Xanthomonas gardneri strain ICMP 7383	NZ_CP018731
Streptomyces sp. CC0208	NZ_CP031969
Paenibacillus dendritiformis C454	NZ_AHKH00000000

Table 2 – Known genomes to be not producers of lasso peptides and the NCBI accession number.

Negative Genomes	**NCBI Accession Number**
Acinetobacter baumanii	NZ_CP009257
Brucella abortus	NC_007618
Campylobacter coli	NZ_CP019977
Corynebacterium diphteriae	NZ_LN831026
Leptospira interrogans	NC_004342
Listeria monocytogenes	NC_003210
Salmonela enterica	NC_003197
Staphylococcus aureus	NC_007795
Shigella dysenteriae	NC_007606

3.5 Query Pattern Design

The query pattern building process was performed using 35 different lasso peptide sequences, which were randomly divided into two groups, one termed 'training' with 21 sequences, shown at Table 3, and a 'testing' group with 14 sequences, shown at Table 4, used to assess the possibility of over-fitting. The training group was aligned manually and submitted to EMBOSS Distmat tool (https://www.bioinformatics.nl/cgi-bin/emboss/distmat), using identity was generate the distance matrix and multi-dimensional scaling was performed in the software R Studio version 3.6.1. The resulting cartesian plot was then group up into three different groups. For each group, the sequences were manually multi-aligned and consensus sequences were extracted.

Table 3 – Group of lasso peptides sequences used to build the distance matrix and multidimensional scaling plot.

Genome	**Lasso peptide**	**Reference**
Asticcaulis excentricus CB 48	Astexin-1	Zimmermann et al. 2013
Phenylobacterium zucineum HLK1	Zucinodin	Hagemman et al., 2013
Burkholderia rhizoxinica HKI 45	Burhizin	Hagemman et al., 2013
Caulobacter sp. K31	Caulonodin I	Hagemman et al., 2013
	Caulonodin II	Hagemman et al., 2013
	Caulonodin III	Hagemman et al., 2013
Caulobacter segnis ATCC 21756	Caulosegnin I	Hagemann et al. 2012
	Caulosegnin II	Hagemann et al. 2012
	Caulosegnin III	Hagemann et al. 2012
Nocardiopsis alba DSM 43377	LP-2006	Tietz et al., 2017
Paenibacillus dendritiformis C454	Paeninodin	Zhu et al., 2016
Rhodococcus jostii K01−B0171	Lariatin	Iwatsuki, W. et al., 2006
Rubrivivax gelatinosus IL44	Rubrivinodin	Hagemman et al., 2013
Sphingobium japonicum UT262	Sphingonodin I	Hagemman et al., 2013
	Sphingonodin II	Hagemman et al., 2013
Streptomyces sp. 46	Anantin B	Tietz et al., 2017
Streptomyces cattleya str. NRRL 8057	Moomysin	Tietz et al., 2017
Streptomyces davawensis strain JCM 4913	Citrulassin A	Tietz et al., 2017
Streptomyces leeuwenhoekii	Chaxapeptin	Elsayed et al., 2015
Streptomyces nodosus strain ATCC 14899 and *Streptomyces mirabilis strain OK461*	Syamicin I	Detlefsen, et al. 1995

Xanthomonas gardneri strain ICMP 7383	Xhantomonin II	Hegemann et al., 2014

Table 4 – Groups of lasso peptides sequences classified as 'testing group' to assess over-fitting in the patterns designed using the 'training group'.

Genome	**Lasso peptide**	**Reference**
Asticcaulis excentricus CB 48	Astexin-2	Maksimov et al., 2013
	Astexin-3	Maksimov et al., 2013
Burkholderia thailandensis E264	Capistruin	Knappe et al., 2008
Escherichia coli strain U44	Microcin J25	Salómon R. and Farías. R, 1992
Streptomonospora alba strain YIM 90003	Streptomonomicin	Metelev et al., 2015
Sphingopyxis alaskensis RB2256	Sphingopyxin I	Hagemman et al., 2013
	Sphingopyxin II	Hagemman et al., 2013
Streptomyces sioyaensis strain *DSM 40032*	RES-701-1	Detlefsen et al., 1995
Streptomyces griseorubens strain JSD-1	RP-71955	Frechet et al., 1994
Streptomyces sp. Tue6075	SRO15-2005	Kersten et al, 2011
Streptomyces variabilis	Lagmysin	Tietz et al., 2017
Thermobifida fusca	Fuscanodin	Koos et al., 2018
Xanthomonas gardneri strain ICMP 7383	Xhantomonin I	Hegemann et al., 2014
Streptomyces sp. CC0208	SSV-2083	Kersten et al, 2011

3.6 Metagenome Mining

A total of 94 GB from sheep rumen metagenome data was obtained National Center for Biotechnology Information Sequence Read Archive, under accession no SRA075938, samples SRX445324 and SRX445325 to be mined for putative novel lasso peptides. The raw data was processed by Trimmomatic (Bolger et al., 2014) and a phred value of 30 was set.

3.7 Platform Expression Design

The vector for expression in *E. coli* strains was designed to contain the whole microcin J25 BGC, the three genes responsible for maturation and transport (*mcjBCD*), and the gene *mcjA*. The difference is that mcjA contains the leader sequence but, the core sequence is a blank space. The core region contains two restriction sites for the enzyme BpiI, a restriction enzyme type IIS, with the intent to be clone any sequence in this region (shown in Figure 16) using the same logic as Golden Gate Assembly (Marillonnet and Grutzner, 2020). The BGC synthetic sequence was cloned into the expression vector pET28a(+), Figure 15, into the *NcoI* site, because of this an Ile was substituted by a Val in the second position in the Mcj25 leader sequence. For all

sequences, RBS was optimized for *E. coli* (Hegemann et al., 2013) and a T7 promoter was placed. The coding regions had codon optimization using IDT Optimization Tool (https://www.idtdna.com/CodonOpt) and NEBCutter Tool (http://www.labtools.us/nebcutter-v2-0/) was used to check the presence of BpiI restriction sites, outside from *mcjA* – in the regions where this restriction site was present, the codon was replaced by its synonymous. A 6His-tag was added to the proteins *mcjB, mcjC, mcjD.*

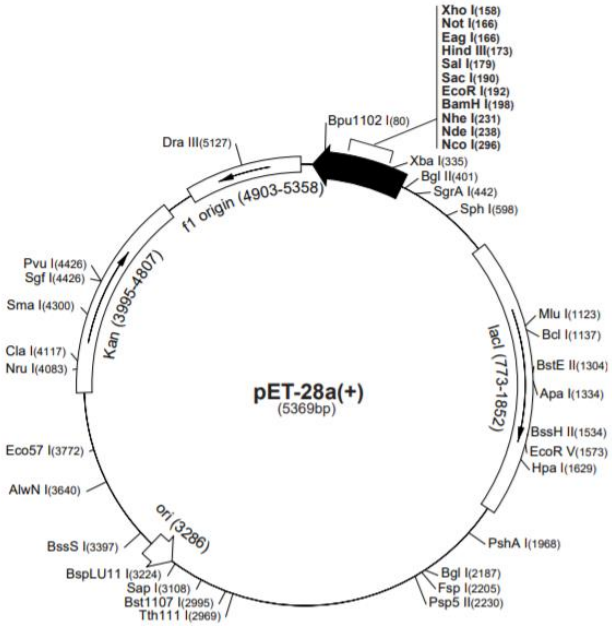


Figure 15 – pET28a(+) vector map, in the cloning region, is where the cluster *mcjABCD* was cloned.

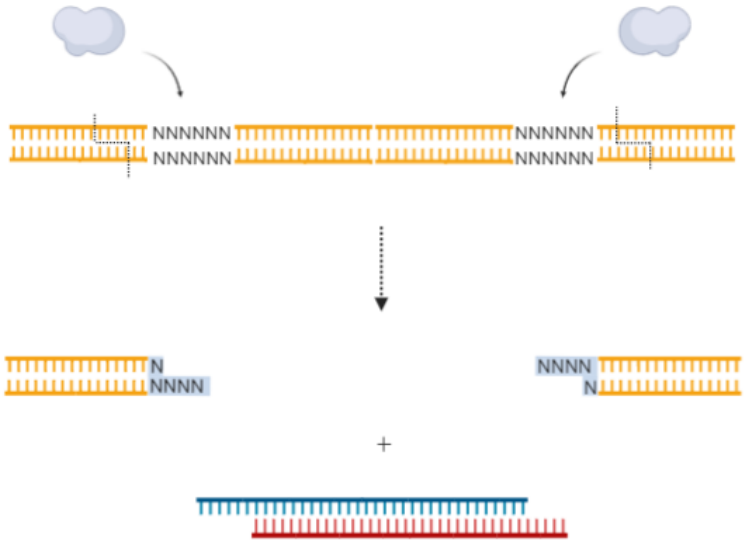


Figure 16 – Diagram showing the two restriction sites type IIs in the *mcjA*. Using the same restriction enzyme, the region is exerted and left behind two overhangs, where the lasso peptide core sequence is cloned.

3.8 Cloning of Microcin J25 Leader Sequence into the Expression Platform

Both synthetic sequences, forward and reverse for Microcin J25 core sequence was purchased as ssDNA, comprising 72 pb each plus 4 pb as overhangs. Each sequence was then phosphorylated with 1 µL of PNK T4 enzyme, 2 µL of Anza Buffer®10x, 500 ng of ssDNA and water nuclease-free in volume to complete 20 µL, the reaction was incubated by 15 minutes at 20 ºC and by 5 minutes at 80 ºC.

The ssDNA phosphorylated were then annealed, to form the dsDNA to be cloned. For the annealing reaction, 4 µL of each dsDNA was added to 2 µL of T4 Buffer Anza 10X and 10 µL of nuclease-free water. The reaction was then incubated for 6 minutes at 95 ºC and cooled down until 25 ºC, at a rate of 1 ºC per minute.

The expression vector pET28(a)+ harboring the microcin J25 BGC was digested using 2 µL of Anza Red Buffer 10X, 1 µL 10 U/µL of BpiI restriction enzyme (Thermofisher), 3 µL of the expression vector, approximately 1500 ng, and nuclease-free water to complete a final reaction volume of 20 µL. The reaction was incubated for 2 hours at 37 ºC. Agarose gel electrophoresis 1% m/m was used to purify the digested plasmid, by 30 minutes at 75 V, and purified using Cellco Agarose Gel Extraction Kit, conforming with the fabricant's instructions.

The 50 ng of purified digested vector were then added to 100 ng of microcin J25 core sequence dsDNA, 2 µL of T4 Buffer 10x, 1 µL of T4 Ligase enzyme and water nuclease-free to a final reaction volume of 20 µL, and incubated at 4ºC overnight. To confirm the cloning, standard PCR was used, to a final reaction volume of 25 µL, 1 µL of each appropriated primer at 10 mM was added to 1 µL of dNTP at 10 mM, 0.25 µL of U/µL Taq DNA polymerase (MARCA), 10-100 ng of template, 1 µL of Buffer 10X and water nuclease-free to complete the reaction volume. The annealing temperature was set to 53 ºC and extension was performed for 30 seconds at 72 ºC.

3.9 Calcium Chloride Heat-Shock Transformation

E. coli strains DH5α and Artic Express were grown in a 5 mL LB culture overnight at 37 ºC and 100 rpm. The cells were harvested by centrifugation (1 x 15 min at 5000 rpm) and washed twice with 1 mL of CaCl2 solution 0.1 M. Afterwards, the pellet was resuspended in 200 µL of CaCl2 solution 0.1 M and 10-50 ng of plasmid added and incubated on ice by 40 minutes. After the incubation period, the cells were incubated at 42ºC for 45 seconds and immersed immediately on ice, staying there for 5 minutes. A total of 500 µL LB medium or SOB was added, and the cells incubated at 37 ºC for 2 h. 100 µL of cells were spread into LB agar plates containing 50 µg/mL of kanamycin and incubated overnight at 37 ºC.

3.10 Heterologous Expression of the Microcin J25 BGC

For the heterologous expression, the expression vector cloned with microcin J25 core sequence was transformed into *E. coli* Artic Express by calcium chloride heat-shock method. One transformed colony was inoculated in 40 mL of LB medium, without antibiotic, and incubated at 37 ºC until OD600 of ~0.4 and then a 1 mL was sampled (Time t0) and the expression was induced by addition of IPTG to a final concentration of 0.04 mM and then cooled to 12 ºC and incubated for 12 h. After this period the cells were harvested by centrifugation (1 x 15 min at 5000 rpm), and the supernatant and pellet recovered and stored.

3.11 Western Blotting

Western Blot was performed to verify the expression of the proteins of the biosynthetic gene cluster machinery, McjB, McjC, and McjD. The pellet harvested in the heterologous expression step was resuspended in PBS buffer. 20 μL of Buffer was added to an aliquot of 60 μL of resuspended pellet and incubated at 98ºC by 5 minutes to lysis the cells. The proteins were electrophoresed by 14% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) by 2 hours at 100V. The proteins separated by SDS-PAGE were transferred using a Semi-Dry Blotter Horitz to a nitrocellulose membrane (60 V, 400 mA, 1h). The membrane was then blocked with BSA/PBS-T solution 3% m/m overnight. Subsequently, the membrane was washed in 3 cycles with TBS-T (10 min per cycle) and then immunoblotted with anti-poly-histidine peroxidase monoclonal antibody, in a concentration of 1:5000 to TBS-T, by 2 hours. Afterward, the membrane was again washed in three cycles, and mice anti-IgG conjugated with peroxidase added to the membrane by 2 hours. To reveal the membrane the colorimetric method was used (50 mM of This_HCL pH 7.6, 10 mg of 3,3'-Diaminobenzidine tetrahydrochloride hydrate, and 10 μL of H2O2 30% m/v).

3.12 Antimicrobial Activity Assay

A single colony of each test bacterial strain was grown in LB overnight at 37 ºC under agitation. Bacteria culture of each strain was added to LB agarose plate 1% m/m. The supernatant obtained from heterologous expression was concentrated in Speed Vac, from an initial volume of 2 mL to 20 μL and added to 3 mm wells punched in the agarose plate. After being incubated overnight at 37 ºC, the diameter of each clean zone of growth inhibition was measured. For negative control *E. coli* strain *Artic Express* wild type was incubated in the same conditions as the strain used to heterologous expression.

4. RESULTS

A total of 29 genomes from three phyla – *Proteobacteria*, *Actinobacteria*, and *Firmicutes* – known to encode 35 different lasso peptides (shown in Panel 1) was used to simulate short *reads* from high-throughput sequencing data, composing mock .sra files. The lasso peptide sequences were selected following the criteria: availability of genomic sequence in public repositories and elucidated structure by NMR or MS.

Panel 1 – All lasso peptide sequences used in the studied.

Astexin-2	MTKRTTIAARRVGLIDLGKATRQTKGLTQIQALDSVSGQFRDQLGLSAD
Astexin-1	MHTPIISETVQPLTAGLIVLGKASAETRGLSQGVEPDIGQTYFEESRINQD
Astexin-3	MRTYNRSLPARAGLTDLGKVTTHTKGPTPMVGLDSVSGQYWDQHAPLAD
Zucinodin	MTRLLNLMSVRLLGFGSAKAATNGGIGGDFEDLNKPFDV
Burhizin	MNKQQQESGLLLAEESLMELCASSETLGGAGQYKEVEAGRWSDRISDDE
Capistruin	MVRLLAKLLRSTIHGSNGVSLDAVSSTHGTPGFQTPDARVISRFGFN
Caulonodin I	MERIEDHIDDELIDLGAASVETQGDVLNAPEPGIGREPTGLSRD

Caulonodin II	MQRIIDETTDGLIELGAASVETQGDVLFAPEPGVGRPPMGLSED
Caulonodin III	MEFEGIPSPDARIDLGLASEETCGQIYDHPEVGIGAYGCEGLQR
Caulosegnin I	MTKKNATQAPRLVRVGDAHRLTQGAFVGQPEAVNPLGREIQG
Caulosegnin II	MTKTHRLIRLGDAQRLTQGTLTPGLPEDFLPGHYMPG
Caulosegnin III	MTSRFQLLRLGKADRLTRGALVGLLLEDITVARYDPM
Microcin J25	MIKHFHFNKLSSGKKNNVPSPAKGVIQIKKSASQLTKGGAGHVPEYFVGIGTPISFYG
LP-2006	MDEEKIGNEISAYETPTVTEMGAFSEVTLGRPNWGFENDWSCVRVC
Paeninodin	MKKQYSKPSLEVLDVHQTMAGPGTSTPDAFQPDPDEDVHYDS
Lariatin	MTSQPSKKTYNAPSLVQRGKFARTTAGSQLVYREWVGHSNVIKPGP
Rubrivinodin	MKEFAMDEELELEIVDLGDAKELTQGAPSLINSEDNPAFPQRV
Streptomonomicin	MSAYEIPTLTRIGKFKDVTKSLGSSPYNDILGYPALIVIYP
Sphingonodin I	MERDNDVIELGAVSVETKGPGGITGDVGLGENNFGLSDD
Sphingonodin II	MDRHDNSEVDEIIDLGTASAVTQGMGSGSTDQNGQPKNLIGGISDD
Sphingopyxin I	MKDFNELIDLGAISVETRGIEPLGPVDEDQGEHYLFAGGITADD
Sphingopyxin II	MERTEVIEEVIDLGKASVETKGEALIDQDVGGGRQQFLTGIAQD
Anantin B	MDEQIELTTAEPYAPPTLTEVGEFNEDTLGFIGWGNDIFGHYSGGF
RES-701-1*	GNWHGTAPDWFFNYYW
Moomysin	MTESIEAYEPPMLVEVGSFAELTRSYHWGDYHDWHHGWYGWWD
Citrulassin A	MKKAYEAPTLVRLGSFRKQTGLLGLAGNDRLVLSKN
RP-71955	MTAIYEPPALQEIGDFDELTKCLGIGSCNDFAGCGYAVVCFW
Chaxapeptin	MEPQMTELQPEAYEAPSLIEVGEFSEDTLGFGSKPLDSFGLNFF
Syamicin I	MSAIYEPPMLQEVGDFEELTKCLGVGSCNDFAGCGYAIVCFW
SRO15-2005	MKQQKQQKKAYVKPSMFQQGDFSKKTAGYFVGSYKEYWSRRII
Lagmysin	MAYERPTLTKVGDFQKVGDFQKVTGLAGQGSPDLLGGHSLL
Fuscanodin	MEKKKYTAPQLAKVGEFKEATGWYTAEWGLELIFVFPRFI
Xhantomonin I	MNSNDTTHSDASNEITVLGVASTDTKGGPLAGEEIGGFNVPGISEE
Xhantomonin II	MDTSNNDARTTALDQDLIVLGVASLDTQGGPLAGEEMIGGITTLGISQD
SSV-2083	MLISTTNGQGTPMTSTDELYEAPELIEIGDYAELTRCVWGGDCTDFLGCGTAWICV

The program receives as input a file containing the reads and a query pattern. In the first attempt to determine a query pattern, to discover novel lasso peptides, all 35 sequences were multi-aligned manually and the logo generated is shown in Figure 17. Figure 18 shows the logo and the consensus sequence for the leader peptide for proteobacteria (Hagemman et al.,

2015).

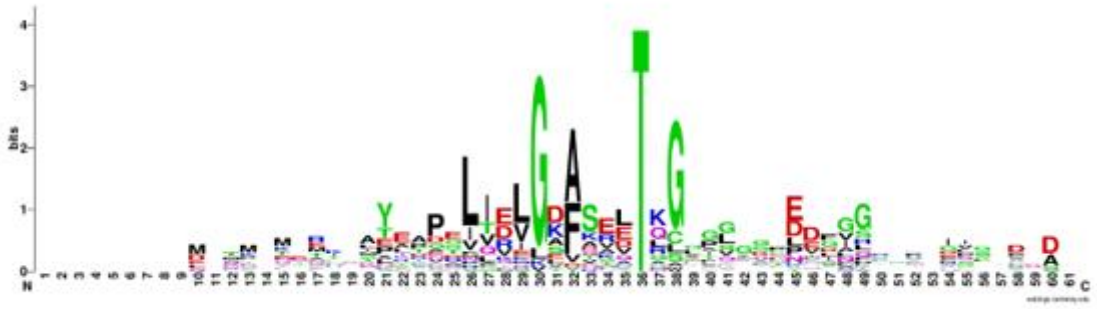


Figure 17 – Logo generated for the 35 lasso peptide sequences. It is notable the high variability within the sequence, where only one residue is conserved in all sequences, a Thr at position -2, in the leader sequence.

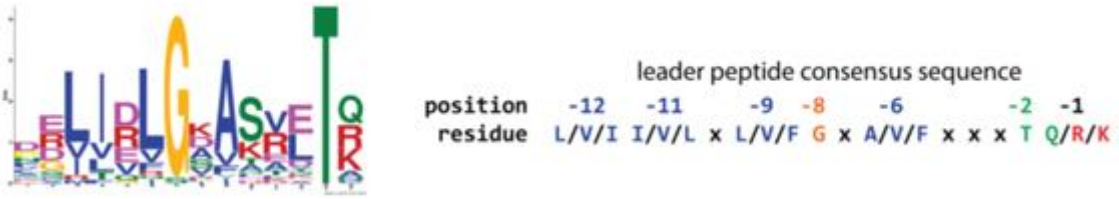


Figure 18 – Logo and consensus sequence obtained for the leader peptide for proteobacteria, by Hagemman (2015). The Thr at position -2 is still conserved.

Based on both logos, the following query pattern was extracted (Figure 19):

-12 -8 -6 -2 +1

Pattern [LVI]-X-X-X-G-X-[AVF]-X-X-X-T-X-G-X{5,7}-[DE]

Figure 19 – Query pattern obtained by the multi-alignment of the 35 lasso peptide sequences.

For the N-terminal position of the core sequence (+1), the Gly residue was fixed, although there are described lasso peptides sequences having a Cys, Ala, Ser, Leu, or Ile at this position. Also, from the logos is possible to visualize variability at positions -12, -8, and -6. For the program to be able to capture this variability, wild cards were introduced (i.e. the number of mismatches permitted) in three mining rounds: i) using no wild card, ii) using one wild card, and iii) using two wild cards. At positions, 7 to 9 are permitted both Glu and Asp, since these residues are necessary to the ring formation. A final constrain is the presence of a stop codon at position 16 or 27, ranging from a minimum length of 15 and a maximum of 26.

In the first round, where no wild card was used, in other words, no mismatch allowed, the program identified little more than 50% of all 35 lasso peptide sequences. Using one wild

card (one mismatch allowed), the rate grew to 90% and using two wild cards (two mismatches allowed), 100% of the 35 lasso sequences were correctly identified. However, using one and two wild cards, the number of sequences identified was in the order of thousands.

To lower down the number of false positives, new patterns were designed based on grouping up the lasso sequences by similarity (using a distance matrix), where sequences more alike would be together and hence the variability would be lower. From the total of sequences, randomly 23 sequences were selected and then used to generate a distance matrix. The remaining 12 sequences were separated and further used to verify over-fitting. Figure 20 shows the graph of multidimensional scaling of the distance matrix.

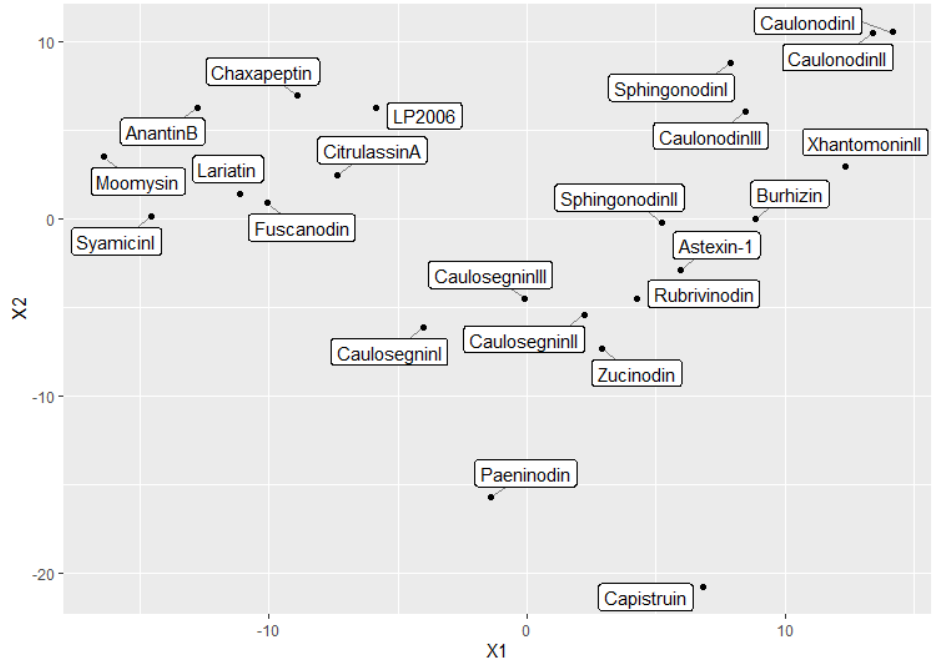


Figure 20 – Multidimensional scaling plot of the distance matrix for the ‘training’ group.

From the MDS plotting shown above, three groups were delineated based on the absolute distance and named Groups: L, O, and N. Figure 21 presents these groups.

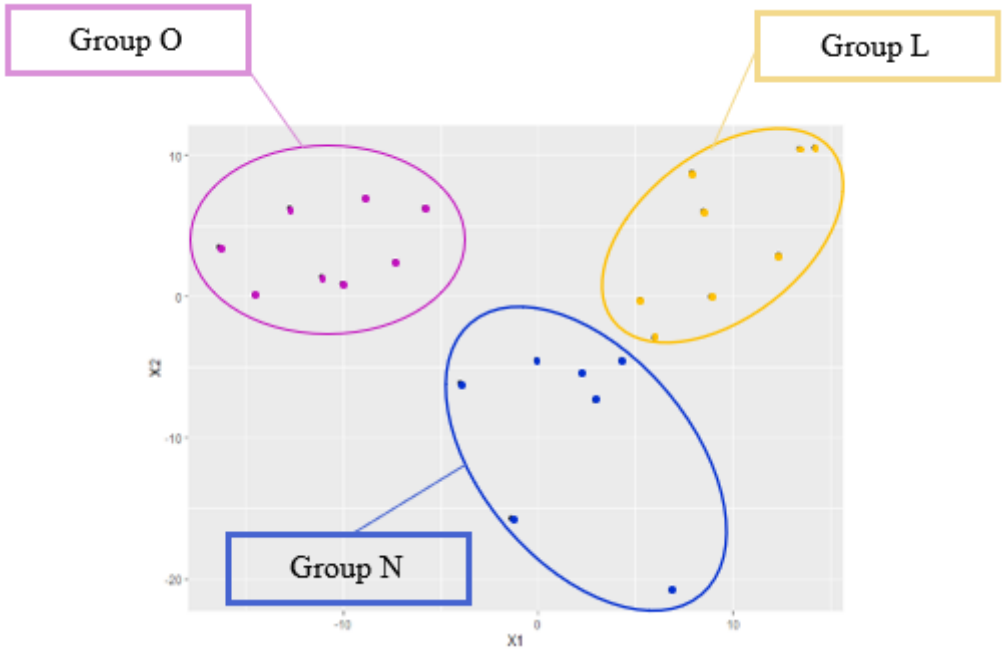


Figure 21 – The three determined groups based on the absolute distance among the sequences.

The sequences from each group were then multi-aligned manually and a single consensus sequence was proposed for groups L, N, and O. Figures 21, 22, and 23 show the result from the alignment for groups L, N, and O, respectively, and the logos.

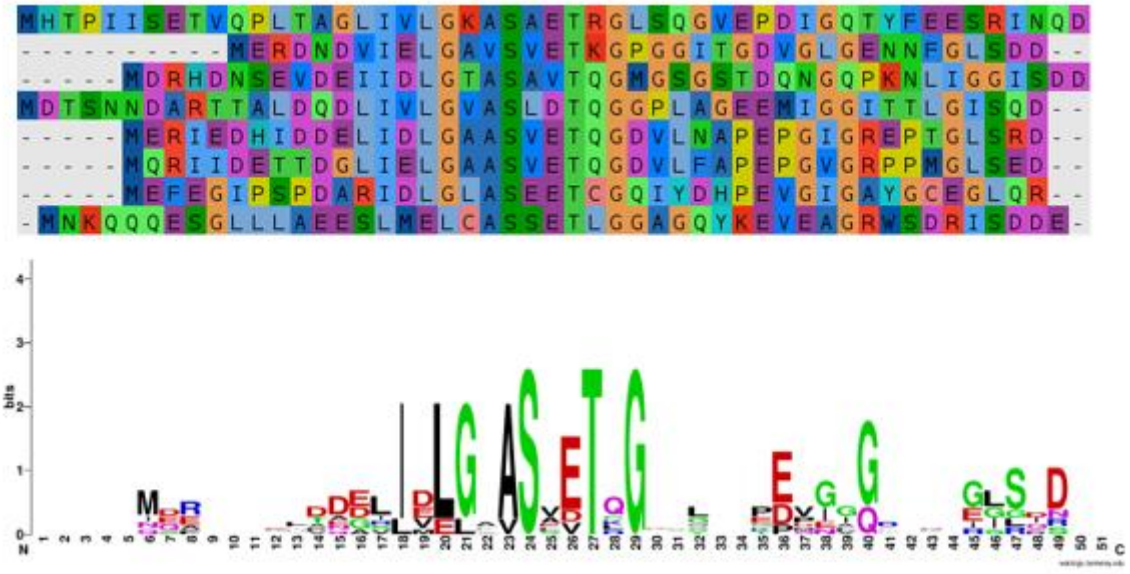


Figure 21 – Multi-alignment for group L and the resulting logo sequence.

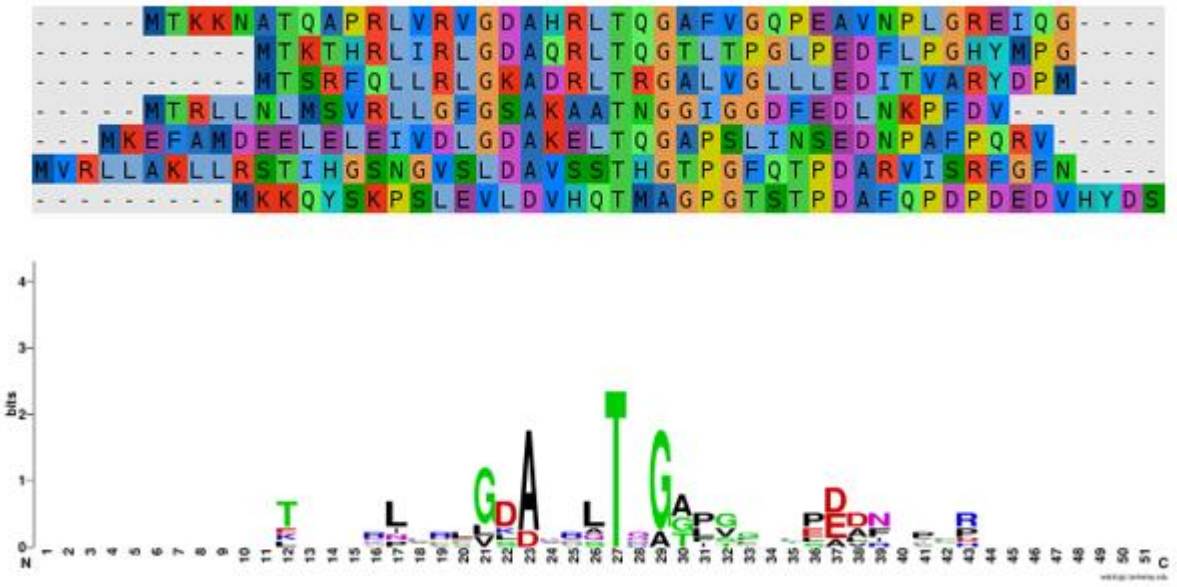


Figure 22 – Multi-alignment for group N and the resulting logo sequence.

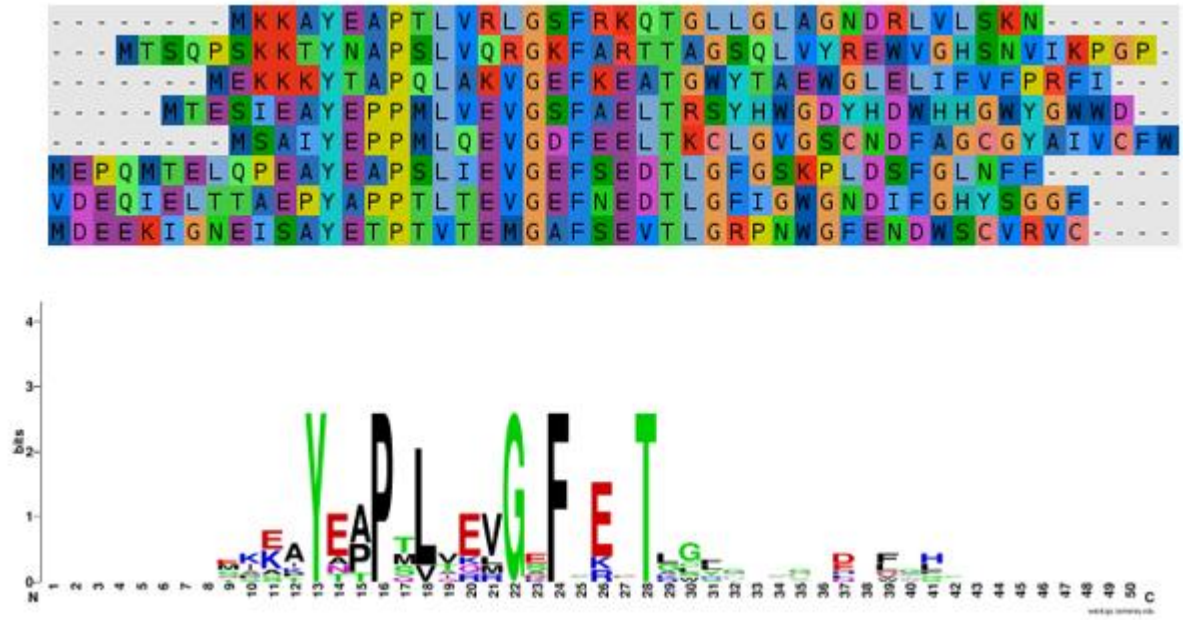


Figure 23 – Multi-alignment for group O and the resulting logo sequence.

For each group, based on the logo, a query pattern was proposed (shown in Panel 2). Getting a single query pattern for each group resulted in more conserved sequences, as we can in the logos where before only position -2 was conserved along with all lasso peptide sequences, and now multiple positions are conserved, within the group. The query patterns were then used as input for two new mining rounds, now including all genomes. The same length constraint was used, a minimum of 15 and a maximum of 26 residues.

Panel 2 – Query patterns determined for the groups L, N, and O.

Group L	[IL]-X-[LE]-[GL]-X-[AV]-X-X-T-X-G-X{5,7}-[DE]
Group N	[LINP]-X-X-X-[GLV]-X-[AD]-X-X-[LAQS]-T-X-[GA]-[AGT]-X{4,6}-[DE]
Group O	Y-X-X-P-X-[LV]-X-X-X-G-X-F-X-[EKR]-X-T-X-[GCLSW]-X{5,7}-[DE]

Tables 6 and 7 show the respective results of the new two mining rounds using the new pattern queries. Table 6 are the result using one wild card and Table 7 using none wild card. In each table is presented the genomes, the number of lasso peptides encoded by the respective genome, and in the last three columns the number of sequences (or sequence pool) identified as lasso peptides, by each query pattern (L, O, and N). For each sequence pool, the number within parentheses represents the known lasso peptides encoded by that genome identified. So, if this number is equal to the number in the second column, this means all lasso peptides were identified, if the sequence pool is higher than that, it means there are false positives in the pool.

Table 6 – Results of the mining round using one wild card.

Genome	Lasso peptides	Nº sequences retrieve to pattern L	Nº sequences retrieve to pattern O	Nº sequences retrieve to pattern N
Asticcaulis excentricus	3	7 (3)	0	60 (2)
Phenylobacterium zucineum HLK1	1	1	0	82 (1)
Burkholderia rhizoxinica HKI 45	1	3 (1)	0	107
Burkholderia thailandensis E264	1	1	0	65 (1)
Caulobacter sp. K31	3	3 (3)	0	123 (2)
Caulobacter segnis ATCC 21756	3	1	1 (1)	127 (3)
Escherichia coli strain U44	1	0	0	81 (1)
Nocardiopsis alba DSM 43377	1	1 (1)	1 (1)	134
Thermobifida fusca	1	0	1 (1)	108
Rhodococcus jostii K01−B0171	1	0	1 (1)	1
Rubrivivax gelatinosus IL44	1	0	0	74 (1)
Streptomonospora alba strain YIM 90003	1	0	1 (1)	125

Sphingobium japonicum UT262	2	2 (2)	0	86 (1)
Sphingopyxis alaskensis RB2256	2	2 (2)	0	83
Streptomyces sp. 46	1	0	1 (1)	185
Streptomyces sioyaensis strain DSM 40032	1	0	2 (1)	181
Streptomyces cattleya str. NRRL 8057	1	0	1 (1)	139
Streptomyces davawensis strain JCM 4913	1	0	1 (1)	202
Streptomyces griseorubens strain JSD-1	1	0	1 (1)	167
Streptomyces leeuwenhoekii	1	0	1 (1)	4
Streptomyces nodosus strain ATCC 14899	1	1	1 (1)	168
Streptomyces sp. Tue6075	1	1	1 (1)	156
Streptomyces variabilis	3	0	3 (3)	131
Xanthomonas gardneri strain ICMP7383	2	2 (2)	0	122 (2)
Streptomyces sp. CC0208	1	0	2 (1)	217

Table 7 – Results of the mining round using none wild card.

Genome	**Lasso peptides**	**Nº sequences retrieve to pattern L**	**Nº sequences retrieve to pattern O**	**Nº sequences retrieve to pattern N**
Asticcaulis excentricus	3	5 (1)	0	0
Phenylobacterium zucineum HLK1	1	0	0	3 (1)
Burkholderia rhizoxinica HKI 45	1	3 (1)	0	2
Burkholderia thailandensis E264	1	0	0	3 (1)
Caulobacter sp. K31	3	3 (3)	0	2
Caulobacter segnis ATCC 21756	3	0	0	5 (3)
Escherichia coli strain U44	1	0	0	2

Nocardiopsis alba DSM 43377	1	0	1 (1)	4
Thermobifida fusca	1	0	1 (1)	1
Rhodococcus jostii K01−B0171	1	0	1 (1)	0
Rubrivivax gelatinosus IL44	1	0	0	3 (1)
Streptomonospora alba strain YIM 90003	1	0	0	5
Sphingobium japonicum UT262	2	2 (2)	0	1
Sphingopyxis alaskensis RB2256	2	1 (1)	0	3
Streptomyces sp. 46	1	0	1 (1)	6
Streptomyces sioyaensis strain DSM 40032	1	0	2 (1)	8
Streptomyces cattleya str. NRRL 8057	1	0	1 (1)	3
Streptomyces davawensis strain JCM 4913	1	0	1 (1)	4
Streptomyces griseorubens strain JSD-1	1	0	1 (1)	5
Streptomyces leeuwenhoekii	1	0	1 (1)	0
Streptomyces nodosus strain ATCC 14899	1	1	1 (1)	4
Streptomyces sp. Tue6075	1	0	0	5
Streptomyces variabilis	3	0	3 (3)	5
Xanthomonas gardneri strain ICMP7383	2	0	0	4 (2)
Streptomyces sp. CC0208	1	0	0	5

In the mining round using one wild card (Table 6), all 35 lasso sequences were identified, but for pattern N larger pools were retrieved, when compared with pattern L and O, which in most cases the sequence pool retrieved corresponds only to the lasso peptides sequences. Table 7 shows the result using no wild card, where 27 of 35 lasso peptide sequences were identified but, the false positive rate decreased significantly for pattern N.

During all mining rounds, 9 negative genomes (shown in Table 8) were also included in the mock community. Only pattern N retrieved few sequences when using one wild card, for the genomes of *S. dysenteriae* and *A. baumanii*.

Table 8 – Negative genomes introduced in the mock community.

Genome	**NCBI Accession**
Acinetobacter baumanii	NZ_CP009257
Brucella abortus	NC_007618
Campylobacter coli	NZ_CP019977
Corynebacterium diphteriae	NZ_LN831026
Leptospira interrogans	NC_004342
Listeria monocytogenes	NC_003210
Salmonela enterica	NC_003197
Staphylococcus aureus	NC_007795
Shigella dysenteriae	NC_007606

For the results in tables 7 and 8, all sequences in the pool not being the true lasso peptides were characterized as non-informative sequences (or true negatives) for the metrics shown below. The mining dataset was composed of millions of reads and few lasso peptides, this unbalance within the data made all query patterns have a specificity very near to 1. In this way, the true positive rate (sensitivity) and positive predictive value (PPV) were used to measure the performance of each query pattern in the respective mining rounds.

Query pattern L retrieved 14 true lasso peptides and 11 false-positive sequences when using one wild card. In the dataset, we have 35 lasso peptide sequences, and so in these conditions, we have a sensitivity of 0.40 and a PPV of 0.70. Using no wild card, the sensitivity went down to 0.22 (a reduction of almost 50%) and the PPV of 0.53. Tables 9 and 10 summarizes these results.

Table 9 – Metrics for query pattern L, using 1 wild card.

1 wild card	Lasso Peptide Sequence	Non-informative sequences
Positive	14	6
Negative	21	1000000
Sensitivity: 0.40	**Specificity: ~1.00**	**PPV: 0.70**

Table 10 – Metrics for query pattern L, using no wild card.

0 wild card	Lasso Peptide Sequence	Non-informative sequences
Positive	8	6
Negative	27	1000000
Sensitivity: 0.22	**Specificity: ~1.00**	**PPV: 0.53**

Query pattern O retrieved 17 true lasso peptides and 2 false-positive sequences when using one wild card. In the dataset, we have 35 lasso peptide sequences, and so in these conditions, we have a sensitivity of 0.48 and a PPV of 0.89. Using no wild card, the sensitivity went down to 0.34 (a reduction of almost 30%) and the PPV of 0.92. Tables 11 and 12 summarizes these results.

Table 11 – Metrics for query pattern O, using 1 wild card.

1 wild card	Lasso Peptide Sequence	Non-informative sequences
Positive	17	2
Negative	18	1000000
Sensitivity: 0.48	**Specificity: ~1.00**	**PPV: 0.89**

Table 12 – Metrics for query pattern O, using no wild card.

0 wild card	Lasso Peptide Sequence	Non-informative sequences
Positive	13	1
Negative	23	1000000
Sensitivity: 0.34	**Specificity: ~1.00**	**PPV: 0.92**

Query pattern N retrieved 14 true lasso peptides and 2928 false-positive sequences when using one wild card. In the dataset, we have 35 lasso peptide sequences, and so in these conditions, we have a sensitivity of 0.4 and a PPV of 0.004. Using no wild card, the sensitivity went down to 0.22 (a reduction of almost 50%) and the PPV of 0.10. Tables 13 and 14 summarizes these results.

Table 13 – Metrics for query pattern N, using 1 wild card.

1 wild card	Lasso Peptide Sequence	Non-informative sequences
Positive	14	2928
Negative	21	1000000
Sensitivity: 0.4	**Specificity: ~1.00**	**PPV: 0.004**

Table 14 – Metrics for query pattern N, using no wild card.

0 wild card	Lasso Peptide Sequence	Non-informative sequences
Positive	8	83
Negative	27	1000000
Sensitivity: 0.22	**Specificity: ~1.00**	**PPV: 0.10**

Separately the query patterns do not have a high sensitivity, and since the design of each query was targeting a specific group, a final evaluation was made uniting the results of queries L and O, summing the unique lasso peptides recovered with pattern L and the unique lasso peptides recovered with pattern O. Here the pattern is excluded because its poor results, having the lowest PPV. The results are summarized in Tables 14 and 15.

Table 15 – Metrics for query pattern L and O, using one wild card.

1 wild card	Lasso Peptide Sequence	Non-informative sequences
Positive	29	8
Negative	6	1000000
Sensitivity: 0.82	**Specificity: ~1.00**	**PPV: 0.78**

Table 15 – Metrics for query pattern L and O, using no wild card.

0 wild card	Lasso Peptide Sequence	Non-informative sequences
Positive	21	7
Negative	14	1000000
Sensitivity: 0.60	**Specificity: ~1.00**	**PPV: 0.75**

One concern about mining large datasets is the time necessary to process the data. In this way, the program presented here has asymptotic complexity linear O(*mn*), and the performance test was conducted. Different mock .sra files were built with different and increasing sizes, ranging from 0.4 GB to 4.6 GB. Using the query patterns N, L, and O, 8 threads, and 10 GB of RAM, the following performance graphic was obtained, confirming its linear complexity. The linear behavior can be seen in Figure 24.

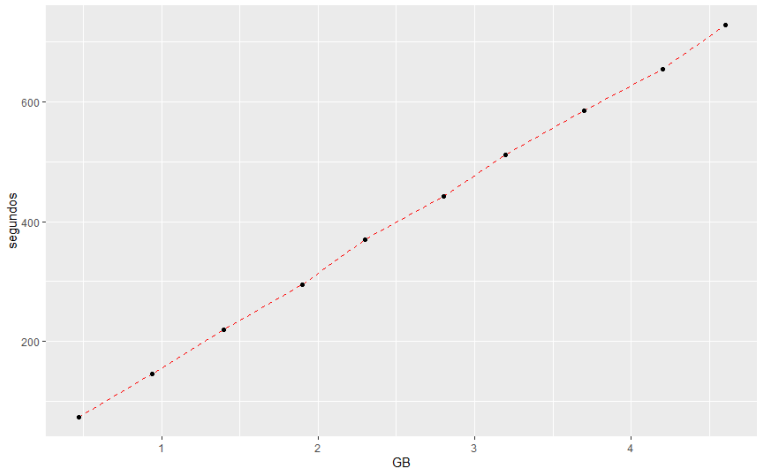


Figure 24 – Test performed using different amount of data and the time needed to process it, showing a linear behavior in agreement with the theoretical asymptotic complexity.

Using the three patterns and one wild card, 90 GB of data obtained from sheep rumen metagenome previously trimmed were used to find novel core lasso peptide sequences. For pattern L and O, respectively pool sizes with 29 and 19 different sequences were retrieved and a total of 1495 sequences for pattern N. To assist the decision making on selecting putative sequences, a heuristic approach was introduced after the mining process. Using a scoring table from RODEO, all sequences within the pools were scored based on the presence or absence and counting of determining residues, and the top-ranked sequences were then selected as putative core sequences. For pattern N, a third filter was applied due to the large number of sequences

recovery. The third filter consists of a regular expression looking for the presence of a Gly at positions 1, 2, and 4 and a Glu at position 8, the same residues found in microcin J25.

Once the program was tested and used to discover putative lasso peptide in raw metagenomics, the expression vector designed was evaluated.

In the expression vector designed, the core sequence for microcin J25 was cloned, and the heterologous expression induced. The sequence cloned is shown in Panel 3, and it was confirmed by standard PCR using as a forward primer the T7 promoter of the pET28a(+) and as a reverse primer, the own negative strand of the core microcin J25. Figure 25 shows the result of the PCR, confirming the cloning.

Panel 3 – Microcin J25 core sequence. The nucleotides marked are complementary to the overhang sequence after the BbsI cleavage.

F	5'**ACTA**AGGGGGGGGCTGGGCACGTTCCTGAGTATTTTGTGGGCATTGGCACGCCGATCTCTTTTTACGGTTAA3'
R	3'TCCCCCCCCCGACCCGTGCAAGGACTCATAAAACACCCGTAACCGTGCGGCTAGAGAAAAATGCCAATT**CAAA**5'

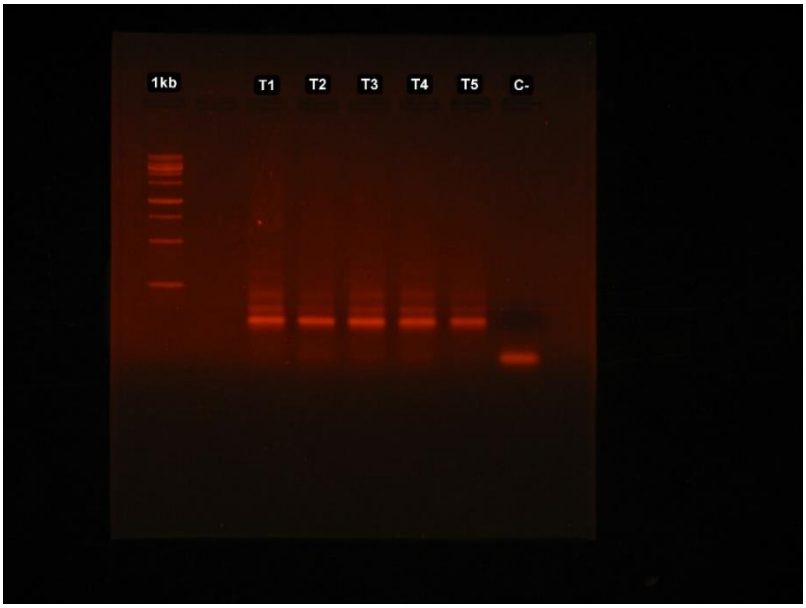


Figure 25 – Agarose gel electrophoresis confirming the cloning of Microcin J25 core sequence.

The expression of the machinery proteins *mcjB*, *mcjC* and, *mcjD* was confirmed using Western Blot since these three genes possess a 6His-tag in the C-terminal end. Figure 26 shows the membrane, at time 0 (0h) already is possible to see the expression of the proteins and 24h after we see multiple bands, one between 45kDa and 60kDa and a strong one near 25kDa. The *mcjC* and *mcjD* have 58.7 kDa and 65kDa respectively and the *mcjB* 24.6 kDa.

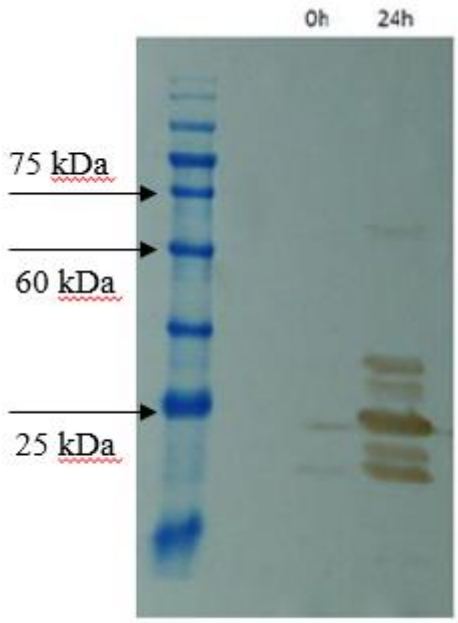


Figure 26 – Western Blot from the harvested cells after 24h of expression induction, showing the bands for proteins *mcjB, mcjC, mcjD,* and also non-specifics bands.

Since the *mcjA* gene containing the cloned core sequence for mccJ25 does not have a 6His-tag, its expression was evaluated by activity bioassay against two bacteria: *Escherichia coli* and *Listeria innocua*. Both pellet and supernatant were used from the induced culture, the negative control was *E. coli Artic Express* wild type.

In the plates were made three wells: in well number 1 was used the cells harvested (pellet) after 24h of induction, in well number 2 was used the supernatant at time 0h and in well 3 was also used the supernatant, but after 24h of induction.

Figure 27 shows the activity test against *E. coli*. The negative control is shown in the left plate, where pellet and supernatants from *E. coli Artic Express* before transformation with the expression vector was used (wild type). The right plate was used supernatant, and pellet of the *E. coli Artic Express* transformed with the expression vector.

Figure 27 – In the wells number 1 was used harvested cells after 24h (pellets), wells number 2 culture supernatant at time 0h, and wells number 3 culture supernatant after 24h. The left plate is the negative control where *E. coli Artic Express* wild type was used against *E. coli*. In the right plate was used the induced *E. coli Artic Express* transformed with the expression vector, against *L. innocua*.

The pellet in both cases did not present any inhibitory activity. However, the supernatants at time 0h and 24h, for both plates presented inhibitory activity, although as shown in Table 16 for *E. coli Artic Express* transformed with the expression vector (column "Diameter mcj25") has shown up to 3x the inhibition diameter when compared to the negative control.

Table 16 – The inhibition zone diameters in mm for each well in both plates, the negative control (column "Diameter WT"), and test (column "Diameter mcj25") against *E. coli*.

Well Number	Diameter WT (mm)	Diameter mcj25 (mm)
1	0	0
2	24.9	68.0
3	41.6	86.0

Figure 28 shows the activity test against *L. innocua*. The same negative control was used, shown in the left plate and the right plate is the bioassay activity with the *E. coli Artic Express* transformed with the expression vector.

Figure 28 – In the wells number 1 was used harvested cells after 24h (pellets), wells number 2 culture supernatant at time 0h, and wells number 3 culture supernatant after 24h. The left plate is the negative control where *E. coli Artic Express* wild type was used against *L.innocua*. In the right plate was used the induced *E. coli Artic Express* transformed with the expression vector, against *L. innocua*.

The pellet in both cases did not present any inhibitory activity. The negative control has not shown any inhibitory activity against *L. innocua*, whilst the transformed strain has shown small, but visible inhibitory activity at time 0h and time 24h. Table 17 summarize the results.

Table 17 – The inhibition zone diameters in mm for each well in both plates, the negative control (column "Diameter WT"), and test (column "Diameter mcj25") against *L. innocua*.

Well Number	Diameter WT (mm)	Diameter mcj25 (mm)
1	0	0
2	0	16.5
3	0	28.0

5. DISCUSSION

Lasso peptides sequences feature high variability, both in the leader and core sequence, as evidenced by the logos generated by the multi-alignment. These peptides present few conservative residues in specific regions, as the ubiquitous Thr at the penultimate position of the leader peptide. Pan and colleagues (2012) conduct mutagenesis analyses, where is proposed the Thr residue bind to a shape-selective pocket within the maturation proteins, and although other residues shape-like Thr is also recognized by the maturation proteins, Thr is the optimal residue and prevalent in all lasso peptides so far discovered (Pan et al., 2012).

The high variability within leader sequences can be explained by the results of Cheung and colleagues (2010), which tested different truncation variants in the leader sequence of the lasso peptide microcin J25. The author cloned different leader sequences wherein each variant was

removed 5 residues, starting from the N-terminal end. For each, the expression level was evaluated. The study found the minimum requirement of eight residues immediately preceding the cleavage site of McjA, in other words, only the last 8 residues of the leader sequence are required to produce normal levels of microcin J25. The last 8 residues are responsible in assist the docking of the precursor peptide into the maturation proteins. This indicates that the variability in the rest of the leader peptide does not impact the normal production of lasso peptides, hence there is no positive selection to maintaining conservation of specific residues at the first positions.

The last 8 residues of the leader sequence along with the presence of the Thr at penultimate position are responsible for the docking in the maturation proteins, this finding explains why for the expression vector we can maintain fixed the Mcj25 leader sequence fixed, whilst the core sequence can be cloned with new lasso peptides, through metagenome mining.

The 8 last residues of the leader sequence, where is expected higher conservation since is responsible for the recognition by maturation proteins, altogether with the maximum length constraint of the core sequence, of 26 residues, represents a total of 34 residues or 102 pb. For most of short reads high-throughput sequencing technologies is feasible the mining of these peptides in a directly from the reads.

The high variability imposes challenges in the designing of patterns since the consensus sequences carry few conserved residues. Here instead of using a unique consensus sequence, was proposed the subdivision of lasso peptides to improve the predictive positive value. But as shown, the query patterns alone do not possess high values of PPV and sensitivity, but when used combined it is possible to significantly increase the sensitivity.

The use of wild cards does not improve the PPV, and still lowers down the sensitivity for the pattern L and O. This indicates that it is possible to capture the variability in consensus sequences through the strategy of subdividing the lasso peptide sequences into different groups. The query pattern N showed poor performance, and hence it is needed future improvements.

The query patterns were used to mine lasso peptides in 9 genomes known to not encode and without any prediction by antiSMASH v5.0 (Blin et al., 2019). The patterns L and O did not retrieve any sequences, only pattern N retrieved sequences from Shigella and Acinetobacter. And so, the assumption of the non-informative sequences retrieved by pattern L and O are true negatives, needs to be further investigated.

Many bacteria are known to encode more than one lasso peptide within the same biosynthetic gene cluster, as Caulonodins (Hegemann et al., 2013), Caulosegnins (Hegemann et al., 2012) and even having more than one BGC, the case of Astexins (Maksimov et al, 2012). Here we report the presence of Lagmysin, a lasso peptide described by Mitchell and colleagues (2017) in a strain of *Streptomyces variabilis*, along with Citrulassin and Aborycin. Using antiSMASH only the two latter peptides have the BGC's predicted: the Citrulassin's BGC at 1,639,468 – 1,661,306 nt. (total: 21,839 nt) and the Aborycin's BGC at d 3,201,531 - 3,224,024 nt. (total: 22,493).

Many bacteria are known to encode more than one lasso peptide within the same biosynthetic gene cluster, as Caulonodins (Hegemann et al., 2013), Caulosegnins (Hegemann et al., 2012) and even having more than one BGC, the case of Astexins (Maksimov et al, 2012). Here we report the presence of Lagmysin, a lasso peptide described by Mitchell and colleagues (2017), in a strain of *Streptomyces variabilis. The Lagmysin found in this strain does not have a predicted BGC by antiSMASH. The prediction points to the presence of* Citrulassin's BGC at 1,639,468 – 1,661,306 nt. (total: 21,839 nt) and the Aborycin's BGC at d 3,201,531 - 3,224,024 nt. (total: 22,493).

A future investigation is a possibility of orphan lasso peptides within a genome, that are still produced but leveraging the protein machinery from other BGC within the same genome. In other words, peptides sequences (gene A) without a biosynthetic gene clusters context, but, expressed due to machinery promiscuity, as observed for the Microcin J25 gene cluster (Maksimov et al 2012, Pavlova et al, 2008). If we assume this possibility, new methods for prospecting lasso peptides independent from genomic contexts, like the one this study provides, will represent milestones in the discovery of novel lasso peptides.

Microcin J25 machinery proteins show a significant degree of substrate promiscuity. Broad mutation analysis of various core peptides revealed the incredibly relaxed specificity of the *mcjC* towards single amino acid substitutions (Zyubko, 2019). Also, insertions e deletions were introduced in the core peptide and did not prevent maturation by the *mcjC*. These findings support the use of the Microcin J25 biosynthetic gene cluster as a platform expression for novel peptides.

The three patterns designed were used against metagenome short reads data from the rumen microbiome. Rumens are an example of such highly competitive communities with very complex microbiomes, under this pressure bacteria develop competitive strategies as producing antimicrobial toxins, which facilitates contest competition with other species (Oyama et al., 2017; Hibbing, 2009). Also, environmental conditions as higher temperature and lower pH, introduce positive selection into protein stability, a physic-chemical property desirable for biotechnological applications.

The *reads* were first trimmed to only retain sequences with high quality, since the program treats each *read* as unique in that dataset, anyone with sequencing error could represent a false positive. For patterns L and O few sequences with the match pattern were retrieved, similar behavior was observed against the positive genomes. For pattern N significant amount of sequences were retrieved. All sequences were passed through a heuristic scoring, to rank them. For pattern L and O the sequences with the high score were selected as putative lasso peptides and for pattern N an additional filter was included. The additional filter was to select the sequences with the pattern **G**-**G**-X{5}-**E**. This pattern is from the Microcin J25 core sequence and was used to select Microcin-like putative peptides and reduce the space of putative peptides.

To validate the expression vector, the core sequence of MccJ25 was cloned and its expression evaluates through activity bioassay. Different from the proteins *mcjB*, *mcjC*, *mcjD* in which a 6His-tag was added during the design, the core sequence of lasso peptide (gene A) wasn't tagged, due to the small size of the mature peptide (21 aa.). Microcin J25 is known to inhibit the growth of a wide range of Gram-negative pathogens including its native producer, *Escherichia coli*. Here the activity against a Gram-positive bacterium, *Listeria monocytogenes*, was also tested even though in the literature there is no report of MccJ25 showing activity against Gram-positive bacteria (Semenova et al., 2005).

The bioassay against the same strain of *E. coli* as the transformant presented inhibition for both: the negative control, where supernatant from *E. coli* wild type was applied, and the test, where was used the supernatant from the *E. coli* transformed. In the first case, *E. coli* is a well-known producer of bacteriocins, fitting in the classical definition, where targets close phylogenetic strains (Riley et al., 2002). Nevertheless, in the test plate, the inhibition zone was larger, which indicates a difference in the composition of the supernatant, which could be the actual Microcin J25. When tested against *L. innocua* the negative control presented any inhibition zone, opposed to the test plate. It was an interesting resulting because if we assumed the presence of *Microcin J25*, the range of activity for this lasso peptide would be expanded to Gram-positive strains.

The machinery proteins *mcjB*, *mcjC* and *mcjD* were tagged with a poly-histidine tag at C-terminal end, and therefore their expression was evaluated using Western Blot. The expected bands appeared along with non-specific ones, which could be the result of degradation and different conformations in the SDS-PAGE electrophoresis. The indication of the activity of microcin J25 in this vector also supports that all proteins are being correctly expressed.

6. CONCLUSION

Here we present an implementation of finite non-deterministic automata as one solution for the pattern matching problem in DNA sequences. Using this algorithm, we were able to directly mine peptides from metagenome raw data, offering the possibility to explore the unique resources which are microbial communities. Although the high variability among the lasso peptide sequences constitutes a limitation to design a single consensus sequence, here we have shown that using intra-groups is a successful approach to address this limitation. The new computation tool independent from genomic context makes it suitable to prospect peptides from low abundant and rare species in a certain community. Alongside a user-friendly expression vector, where the promiscuity of the biosynthetic machinery of lasso peptides is exploited, allows a quick cycle of metagenomic mining, design of oligonucleotides, cloning, heterologous expression, and bioactivity evaluation.

7. FUTURE RESEARCH

Using more lasso peptides sequences is possible to get a set of features from these peptides, which could improve clusterization resulting in better query patterns. The putative peptides prospected in the metagenome mining are going to be submitted to the cycle of cloning, heterologous expression, and bioactivity evaluation, as well as application in food science.

8. REFERENCES

ALVAREZ-SIEIRO, P; MONTALBÁN-LÓPEZ, M. Bacteriocins of Lactic Acid Bacteria: Extending the Family. Appl. Microbiol. Biotechnol., v. 100, p. 2939-2951, 2016.

BALTZ, H. Marcel Faber Roundtable: Is out Antibiotic Pipeline Unproductive Because of Starvation, Constipation or Lack of Inspiration?. J. Ind. Microbiol. Biotechnol., v. 33, p. 507-513, 2006.

BLIN, K. et al. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res., v. 47, n. W1, p. 81-87, 2019.

BOLGER, A. M., LOHSE, M., & USADEL, B. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, v. 30, n. 15, p. 2114-2120, 2014.

BRADY, S. et al. Metagenomic Approaches to Natural Products Discovery from Free-living and Symbiotic Organisms. Nat. Prod. Rep., v. 26, p. 1488-1503, 2009.

CHARLOP-POWERS, Z.; MILSHTEYN, A.; BRADY, S. Metagenomic Small Molecule Discovery Methods. Curr. Opin. Microbiol., v. 19, p. 70-75, 2014.

CHEUNG-LEE, W.; LINK, A. Genome Mining for Lasso Peptides: Past, Present and Future. J. Soc. Ind. Microbiol. Biotechnol., 2019.

CHEUNG, W. et al. Much of the Microcin J25 Leader Peptide is Dispensable. J. Am. Chem. Soc., v. 132, p. 2514-2515, 2010.

COTTER, P.; ROSS, P. HILL, C. Bacteriocins – a viable alternative to antibiotics?. Nat. Rev. Microbiol., v. 11, n. 2, p. 95-105, 2013.

RITS-CHRISTOPH, A. et al. Novel Soil Bacteria Possess Diverse Genes for Secondary Metabolite Biosynthesis. Nature, v. 558, p. 440-444, 2018.

DETLEFSEN, D. et al. Siamycins I and II, new anti-HIV-1 peptides: II. Sequence Analysis and Structure Determination of Siamycin I. J. Antibiot., v. 48, n. 12, p. 1515-1517, 1995.

DURAND, G.; RAOUL, D.; DUBOURG, G. Antibiotic Discovery: history, methods and perspectives. Int. J. Antimicrob. Agents., v. 53, p. 371-382, 2019.

ELSAYED, S. et al. Chaxapeptin, a Lasso Peptide from Extremotolerant *Streptomyces leeuwhenhoekii* strain C58 from the Hyperarid Atacama Desert. J. Org. Chem., v. 80, p. 10252-10260, 2015.

FRÉCHET, D. et al. Solution Structure of RP 71955, a New 21 Amino Acid Tryciclic Peptide Active against HIV-1 Virus. Biochemistry, v. 33, p. 42-50, 1994.

GHURYE, J.; CEPEDA-ESPINOZA, V.; POP, M. Metagenomic Assembly: Overview, Challenges, and Applications. Yale J. Biol. Med., v. 89, p. 353-362, 2016.

HAGEMANN, J. et al. Caulosegnins I-III: A Highly Diverse Group of Lasso Peptides Derived from a Single Biosynthetic Gene Cluster. J. Am. Chem. Soc., v. 135, p. 210-222, 2012.

HAGEMANN, J. et al. Lasso Peptides from Proteobacteria: Genome Mining Employing Heterologous Expression and Mass Spectrometry. Biopolymers, v. 100, n. 5, p. 527-542, 2013.

HEGEMANN, J. et al. Xanthomonins I-III: A New Class of Lasso Peptides with a Seven-Residue Macrolactam Ring. Angew. Chem. Int. Ed., v. 53, p. 2230-2234, 2014.

HAGEMANN, J. et al. Lasso Peptides: An Intriguing Class of Bacterial Natural Products. Acc. Chem. Res. v. 48, p. 1909-1919, 2015.

HAGEMMANN, J. Factors Governing the Thermal Stability of Lasso Peptides. ChemBioChem., v. 21, n. 1, p. 7-18, 2019.

HIBBING, M. et al. Bacterial Competition: Surviving and Thriving in the Microbial Jungle. Nat. Rev. Microbiol., v. 8, 2010.

HUTCHINGS, M. et al. Antibiotics: Past, Present and Future. Curr. Opin. Microbiol., v. 51, p. 72-80, 2019.

IQBAL, H. et al. Natural Products Discovery Through Improved Functional Metagenomics in Streptomyces. J. Am. Chem. Soc., v. 138, n. 30, p. 9341-9344, 2016.

IWATSUKI, M. et al. Lariatins, Antimycobaterial Peptides Produced by *Rhodococcus sp. K01-B0171*, Have a Lasso Structure., J. Am. Chem. Soc., v. 128, p. 7486-7491, 2006.

JOHNSON, J.; JAIN, K.; MADAMWAR, D. Functional Metagenomics: Exploring Nature's Gold Mine. In: PANDEY, A. et al. Current Developments In Biotechnology And Bioengineering. Netherlands: Elsevier, 2017. p. 27-43.

KERSTEN, R. et al. A Mass Spectrometry-Guided Genome Mining Approach for Natural Product Peptidogenomics. Nat. Chem. Biol. v. 7, n. 11, p. 794-802, 2012.

KNAPPE, T. et al. Introducing Lasso Peptides as Molecular Scaffolds for Drug Design: Engineering of an Integrin Antagonist. Angew. Chem. Int., v. 50, p. 8714-8717, 2011.

KNAPPE, T. et al. Isolation and Structural Characterization of Capistruin, a Lasso Peptide Predicted from the Genome Sequence of Burkholderia thailandensis E264. J. Am. Chem. Soc., v. 130, p. 11446-11454, 2008.

KOOS, J. and LINK, A. Heterologous and in Vitro Reconstitution of Fuscanodin, a Lasso Peptide from Thermobifida fusca. J. Am. Chem. Soc., v. 141, n. 2, p. 928-935, 2018.

LING, L. et al. A New Antibiotics Kills Pathogens Without Detectable Resistance. Nature., v. 517, 2015.

MAKSIMOV, M. and LINK, A. Prospecting genomes for Lasso Peptides. J. Ind. Microbiol. Biotechnol., v. 41, n. 2, p. 333-344, 2013.

MAKSIMOV, M. et al. Precursor-centric Genome-mining Approach for Lasso Peptide Discovery. PNAS, 2012a.

MAKSIMOV, M. et al. Lasso Peptides: Structure, Function, Biosynthesis, and Engineering. Nat. Prod. Rep., 2012b.

MARILLONNET, S., & GRUTZNER, R. Synthetic DNA Assembly using Golden Gate Cloning and the Hierarchical Modular Cloning Pipeline. Curr. Protoc. Mol. Biol. V. 130, n. e115, p 1-33, 2020.

MARTIN-GÓMEZ, H.; TULLA-PUCHE, J. Lasso Peptides: chemical approaches and structural elucidation. Org. Biomol. Chem., 2018, v. 16, p. 5065-5080, 2018.

MAVAERE, J. Lasso Peptides from Actinobacteria – Chemical Diversity and Ecological Role. Biomolecules. Université Pierre et Marie Curie, Paris VI, 2016.

MCKENNA, M. The Antibiotic Paradox: Why Companies can't Afford to Create Life-saving Drugs. Nature, v. 584, p. 338-341, 2020. Accessed September 9th at: https://www.nature.com/articles/d41586-020-02418-x?WT.ec_id=NATURE-20200820&utm_source=nature_etoc&utm_medium=email&utm_campaign=20200820&sap-outbound-id=E0398125C2E4708945FE3FBEDA4CB53C3A396F46

METELEV, M. et al. Structure, Bioactivity, and Resistance Mechanism of Streptomonomicin, an Unusual Lasso Peptide from an Understudied Halophilic Actinomycete. Chem. Biol., v. 22, n. 2, p. 241-250, 2015.

NAIMI, S. et al. Fate and Biological Activity of the Antimicrobial Lasso Peptide Microcin J25 Under Gastrointestinal Tract Conditions. Front. Microbiol., v. 9, n. 1764, 2018.

OYAMA, L. et al. The Rumen Microbiome: an Underexplored Resource for Novel Antimicrobial Discovery. NPJ Biofilms. Microbiol., v. 33, 2017.

PAN, S. et al. The Role of a Conserved Threonine Residue in the Leader Peptide of lasso Peptide Precursors. Chem. Commun., v. 48, p. 1880-1882, 2012.

PAVLOVA, O. et al. Systematic Structure-Activity Analysis of Microcin J25. J. Biol. Chem., v. 283, p. 25589-25595, 2008.

REHMAN, Z. et al. Genome Sequence Analysis of Zooshikella gangwensis strain VG4 and its Potential for the Synthesis of Antimicrobial Metabolites. Biotechnol. Rep., v. 19, 2018.

RILEY, M.; WERTZ, J. Bacteriocins: Evolution, Ecology, and Application. Annu. Rev. Microbiol., v. 56, p.117-137, 2002.

SALAMÓN, R. and FARÍAS, R. Microcin J25, a Novel Antimicrobial Peptide Produced by *Escherichia coli*. J. Bacteriol., v. 174, n. 22, 1992.

SEMENOVA, E. et al. Structure-Activity Analysis of Microcin J25: Distinct Parts of Threaded Lasso Molecule Are Responsible for Interaction with Bacterial RNA Polymerase. Journal of Bacteriology, v. 187, n. 11, 2005.

TIETZ, J. et al. A new Genome-mining Tool Redefines the Lasso Peptide Biosynthetic Landscape. Nat. Chem. Biol., v. 13, n. 5, p. 470-478, 2017.

UK REVIEW ON ANTIMICROBIAL RESISTANCE. (chaired by Jim O'Neill). Tackling drug-resistant infections globally: final report and recommendations. 2016. https://amr-review.org/Publications.html.

VAN HEEL, A. et al. BAGEL4: a user-friendly web server to thoroughly mine RiPPs and bacteriocins. Nucleic Acids Res., v. 46, n. W1, p. 278-281.

WINTER, J.; BEHNKEN, S.; HERTWECK, C. Genomic-inspired Discovery of Natural Products. Curr. Opin. Chem. Biol., v. 15, p. 22-31, 2011.

WOOLEY, J. and YE, Y. Metagenomics: Facts and Artifacts, and Computational Challenges. J. Comput. Sci. Technol., v. 25, n. 1, p. 71-80, 2010.

World Bank. Drug resistant infections: a threat to our economic future (vol. 2): final report. 2017. Report Number 114679. http://documents.worldbank.org/curated/en/323311493396993758/pdf/114679-REVISED-v2-Drug-Resistant-Infections-Final-Report.pdf.

ZHAO, N. et al. Lasso Peptide, a Highly Stable Structure and Designable Multifunctional Backbone. Amino Acids. 2016.

ZIMMERMANN, M. et al. The Astexin-1 Lasso Peptides: Biosynthesis, Stability, and Structural Studies. Chemistry & Biology, v. 20, p. 558-569, 2013.

ZHU, S. et al. Insights into the Unique Prosphorylation of the Lasso Peptide Paeninodin. J. Biol. Chem., v. 291, p. 13662-13678, 2016.

ZYUBKO, T. et al. Efficient *in vivo* synthesis of lasso peptide pseudomycoidin proceeds in the absence of leader and leader peptidase. Chem. Sci., v. 10, n. 42, p. 9699-9707, 2019.