Norwegian University of Life Sciences

Master's Thesis 2018 60 ECTS

Faculty of Chemistry, Biotechnology and Food Science
Lars-Gustav Snipen

Identifying Coevolution of Class IIa Bacteriocin and their Immunity Protein

Identifisering av Koevolusjon mellom Klasse IIa Bakterieosin og deres Immunitetsprotein

Gard Kroken

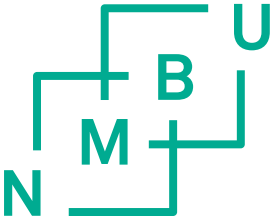
Master of Bioinformatics
Faculty of Chemistry, Biotechnology and Food Science

Norwegian University of Life Sciences (NMBU)
Faculty of Chemistry, Biotechnology and Food Science

Identifying Coevolution of Class IIa Bacteriocin and their Immunity Protein

Gard Kroken

Master's Thesis



Norwegian University
of Life Sciences

Norway
May 2018

Supervisors:
Lars-Gustav Snipen
Dzung Bao Diep
Morten Kjos

Abstract

Class IIa bacteriocins are Antimicrobial peptides found in many lactic acid bacteria. Here they serve as a weapon against competing bacteria by causing cell leakage through pore-formation in the membrane. An immunity protein is produced along with the bacteriocins, which confers immunity for the producer, while foreign bacteria remain susceptible. We hypothesise that these two proteins have coevolved. To test this, we have gathered hundreds of bacteriocin and immunity protein sequences through an iterative genome-mining procedure using HMMER3. We then extract two subsets of the results and apply two forms of MirrorTree methods to each set. A technique was devised to investigate which regions of the bacteriocin correlated the highest with the immunity protein by cropping the bacteriocin from either end. Lastly, we perform a PCA on both datasets as an explorative measure.

The iterative genome mining procedure revealed 59 class IIa bacteriocin-carrying species after filtering, several of which are new additions to the already known bacteriocin producers.

Both the conventional MirrorTree method and the pMirrorTree method showed high levels of coevolution between the sequences, and the set consisting of 16 species in the *Streptococcus* genus showed a very significant coevolution, while the set consisting of sequences from *E. faecium* and *E. faecalis* failed to show significant coevolution through the pMirrortree method.

The results from the cropping method indicates that the C-terminal half of the bacteriocins hold more influence on the coevolution than the N-terminal half.

Lastly, the PCA showed that the bacteriocin and immunity protein were highly correlated in both sets, and interestingly it showed that the bacteriocin and immunity protein in the *Enterococcus* set did not follow the same evolutionary pattern as more conserved sequences did.

We conclude that the class IIa bacteriocins and immunity proteins are coevolved.

Sammendrag

Klasse IIa bakteriosiner er antimikrobielle peptider som finnes i mange melkesyrebakterier. De opererer som vpen mot konkurrerende bakterier ved forrsake cellelekkasje gjennom poreformasjon i membranen. Et immunitetsprotein produseres sammen med bakteriosinene, som gjr produsenten immun mot bakteriosinet, en luksus de fremmede bakteriene ikke fr ta del av. Hypotesen som testes er om bakteriosinet og immunitets proteinet er koevolverte. For teste dette har vi samlet hundrevis av bakteriocin- og immunitetsproteinsekvenser gjennom en iterativ genomscan prosedyre ved bruk av HMMER3. Vi trekker deretter ut to sett med bakteriosiner fra resultatene og bruker to former for MirrorTree-metoder for beregne korrelasjon, og dermed koevolusjon. En teknikk ble utformet for underske hvilke regioner av bakteriosinet som korrelerte hyest med immunitetsproteinet ved trimme aminosyrer fra hver ende av bacteriosinet. Til slutt utfrer vi en PCA p begge datasettene som et eksplorativt tiltak.

Den iterative genomscan prosedyren avslrte 59 klasse IIa bakteriosinbrende arter etter filtrering, hvorav flere er nye treff til de allerede kjente bakteriosinprodusentene.

Bde den konvensjonelle MirrorTree-metoden og pMirrorTree-metoden viste hye niver av koevolusjon mellom sekvensene, og settet bestende av 16 arter i *Streptococcus*-slekten viste en meget signifikant koevolusjon, mens settet bestende av sekvenser fra *E. faecium* og *E. faecalis* mislyktes med vise signifikant koevolusjon gjennom pMirrortree-metoden.

Resultatene fra trimmemetoden indikerer at den C-terminale halvparten av bakteriosinet har strre innflytelse p koevolusjonen enn den N-terminale halvparten.

Til slutt viste PCA at bakteriosin- og immunitetsproteinet var sterkt korrelert i begge settene, og spesifikt i *Enterococcus*-settet viste det at bakteriocin- og immunitetsproteinet ikke fulgte det samme evolusjonre mnsteret som mer konserverte sekvenser gjorde.

Vi konkluderer med at klasse IIa bakteriociner og immunitetsproteiner er koevolverte.

Acknowledgements

I would like to express immense gratitude towards my chief advisor Lars-Gustav Snipen for sticking with me through this project. Your advice has been invaluable, and I am confident that without your help this thesis would not be half of what it became.

I want to thank Dzung Bao Diep and Morten Kjos for allowing me to write a thesis on a subject I find highly interesting, and for always taking time out of your day to answer any question I might have.

I would like to thank Yngve Mardal Moe, Espen Snneland, Magne Bugten, Trym Teigene, Oliver Edwin, and Cecilie Meinich for your friendship, support, and sound advice both before and during this thesis.

Lastly, I would like to thank my family for their love and support through my five years at the University.

Contents

Abstract i

Sammendrag ii

Acknowledgements iii

List of Figures vi

List of Tables ix

Abbreviations x

1 Introduction 1
- 1.1 Bacteriocins . . . 2
  - 1.1.1 Class IIa: Pediocin-like bacteriocins . . . 4
  - 1.1.2 Interraction with the Mannose Phosphotransferase System . . . 5
  - 1.1.3 Immunity Proteins . . . 6
- 1.2 coevolution . . . 7
- 1.3 The goal of this study . . . 9

2 Methods 10
- 2.1 Sequence Alignments . . . 10
  - 2.1.1 Pairwise Sequence Alignments . . . 10
  - 2.1.2 BLAST . . . 12
  - 2.1.3 Multiple sequence alignments and MUSCLE . . . 13
  - 2.1.4 HMMER and Hidden Markow Models . . . 14
- 2.2 Locating bacteriocins with HMMER3 . . . 16
  - 2.2.1 Choice of sequences . . . 16
  - 2.2.2 The NCBI database . . . 17
  - 2.2.3 Iterative Hmmsearch . . . 17
- 2.3 Locating immunity proteins . . . 18
- 2.4 Measurement of co-evolution . . . 18
  - 2.4.1 Mirrortree . . . 20
  - 2.4.2 P-MirrorTree . . . 20
  - 2.4.3 Measuring correlation while cutting off amino-acids from either terminal . . . 23
  - 2.4.4 Principal Component Analysis of the distances . . . 24

3 Results **25**
3.1 Seed sequences . . . 25
3.2 HMMsearch . . . 25
3.3 MirrorTree . . . 27
3.4 pmirrortree . . . 29
3.4.1 *Enterococcus* . . . 29
3.4.2 *Streptococcus* . . . 31
3.5 Exploratory analysis to identify correlating regions . . . 32
3.5.1 *Enterococcus* . . . 33
3.5.2 *Streptococcus* . . . 34
3.6 PCA . . . 38
3.6.1 *Enterococcus* . . . 38
3.6.2 *Streptococcus* . . . 40

4 Discussion **42**
4.1 HMMER search of all genomes . . . 42
4.1.1 Comparing hits to the BAGEL4 database . . . 42
4.1.2 HMMER over BLAST+ . . . 43
4.1.3 The hmmsearch pipeline . . . 43
4.1.4 The missing YGNGV-motiv . . . 43
4.2 Mirrortree server . . . 44
4.3 pMT . . . 45
4.3.1 Cropping method . . . 46
4.4 Principle Component Analysis . . . 47
4.5 Future research . . . 48

5 Conclusion **50**

A **52**

Bibliography **59**

List of Figures

1.1 Overview of the different classes of bacteriocins produced by LAB (Alvarez-Sieiro et al., 2016) 4
1.2 The three dimensional structure of the membrane bound class IIa bacteriocin Sakacin P. Moving left to right, the N-terminal β-sheet domain is marked in pink ahead of the Cystein disulfide bridge. Between the two Cys residues is the hinge-region connecting the two separate domains. The C-terminal domain α-helix is marked in *cyan*. (PDB ID: 1OG7) 5
1.3 The complete pipeline of a class IIa bacteriocins life-cycle, from synthesis to attachment to the Man-PTS. (Ennahar et al., 2000) 6
1.4 Layout of bacteriocin operon. The structural bacteriocin is marked *Green*, directly following this is the immunity protein (*Blue*). Following the immunity protein we find the accessory protein (*purple*) and transport proteins (*red*). The order of these last two tend to vary. In some bacteriocin operons there are regulatory genes incorporated into the operon. (*yellow*) (Kjos, 2012) 7
1.5 "Proposed model of the mode of action (A) and immunity (B) in class IIa bacteriocin systems. The class IIa bacteriocin is depicted in red. (A) The N-terminal -sheet containing part of the bacteriocin initially interacts with an extracellular loop (highlighted) of the man-PTS IIC protein. The helix-containing C-terminal part then engages in specific interactions with transmembrane helices of the IIC and/or IID proteins to cause conformational changes which, in turn, lead to pore formation and eventually cell death. (B) In immune cells, the bacteriocin mediates the same conformational changes, but the pore is blocked by a specific immunity protein (blue) which binds tightly to man-PTS." - Kjos et al. (2011) . 8

2.1 The Hidden Markow Model of the occasionally dishonest casino. 15
2.2 The MSA of the original 23 sequences used. The sequences are a subset of the bacteriocins listed in the review article by J. Nissen-Meyer et al. Nissen-Meyer et al. (2009). 16
2.3 Implementation of itereative search process 19
2.4 The altered implementation of the pMirrorTree method 23

3.1 The resulting mirrortree from the MirrorTree server (Ochoa and Pazos, 2010) using the *Enterococcus* data set. *E. faecium* and *E. faecalis* (*red* box) are visually distinct from one another forming two clear groups in the tree, both in the bacteriocin tree (top) and the immunity protein tree (bottom). 27
3.2 The resulting mirrortree from the MirrorTree server (Ochoa and Pazos, 2010) using the *Streptococcus* data set. The top tree is built from the bacteriocins, while the bottom tree is built from the Immunity proteins 28

3.3 The distribution of correlations in the *Enterococcus* genomes analyzed. The two marked lines indicate the two correlations calculated for the bacteriocin and immunity protein. The first using the whole bacteriocin (*green*), the second using the cropped sequence found in section 3.5 (*red*) 29
3.4 A histogram of the mean distance for each protein in the *Enterococcus* set. The mean distance for the bacteriocin (*red*) and immunity protein (*blue*) are marked as vertical lines. 30
3.5 The correlation distribution of the *Enterococcus* proteins after the shuffling procedure has been carried out. The two target correlations have been marked as in figure 3.3. 30
3.6 The distribution of correlation scores derived from the distance matrix before the swapping procedure had been applied to the data. The correlation between the bacteriocins and immunity proteins are marked at two locations, one for the score of the entire bacteriocin is used for the calculation (*green*), and one where only the optimal cut-off from the N-terminal was used (*red*) 31
3.7 A histogram of the mean distance for each protein in the *Streptococcus* set. The mean distance for the bacteriocin (*red*) and immunity protein (*blue*) are marked as vertical lines. 32
3.8 The distribution of correlation scores derived from the shuffled distance frame. The correlation between the bacteriocins and immunity proteins are marked at two locations, one for the score of the entire bacteriocin is used for the calculation (*green*), and one where only the optimal cut-off from the N-terminal was used (*red*) 32
3.9 A plot of the resulting correlation scores for the bacteriocins in the *Enterococcus* set using the cropping method from the N-terminal. There is an optimal cut off point at position 28, which corresponds to right after the GG-motif in the MSA. 33
3.10 Results from the window-method cutting of from the C-terminal. The two major dips in correlation correspond to the first cystein (*red*) downstream from YGNG-motif, and the second corresponds to the tyrosine (*blue*) of the YGNG-motif. . . 34
3.11 The correlations of the window-method as it is applied to the N-terminal of the bacteriocins found in the *Streptococcus* data-sets. The highest correlation is found at position 28 (*red*), which is the start of the common leader peptide. Another peak is found when calculating correlation from position 39 in the MSA, which corresponds to a variable amino acid just upstream from the Gly-Gly processing cite (*blue*). After this point the correlation drops until position 45 (*green*), corresposinding to the first glysine in the Gly-Gly processing cite. Lastly a new peak is reached at position 53 (*purple*), corresponding to a position just ahead of the pediocin -box motif. 35
3.12 Subset of figure 3.11, showing the fluctuations between position 1 to 60. The three marked sections are the start of the leader peptide (*red*), a variable amino acid just upstream from the double Glysine-motif (*blue*), the first glysine of the GG-motif (*green*), and the beginning of the pediocin-box motif (*purple*) 36
3.13 The window method applied to the C-terminal of the bacteriocin MSA, using the optimal cut-off from the N-terminal (position 28) found in figure 3.12. Two points are marked where a major drop in correlation has occurred, at position 64 (*red*) and 73 (*blue*) from the C-terminal. These correspond to one of the Cysteine needed to form the disulfide bridge, and the pediocin box motif. 37

3.14 The cumulative explained variance in the principal component analysis of the evolutionary distance matrix for the *Enterococcus* data-set. Only the first 10 of the 766 PCs are shown. Using three PCs to plot the variance, over 80% of the variance is explained. 38
3.15 Plot of the top three principal components. The bacteriocin (*blue*) and immunity protein (*green*) are visually distinct from the CGFs (*red*), however always in close proximity to each other. 39
3.16 The cumulative explained variance in the principal component analysis of the evolutionary distance matrix for the *Streptococcus* data-set. Only the first 10 of the 528 PCs are shown. By using the first three Pricipal Components to plot the variance in figur 3.17, over 60% of the variance is explained. 40
3.17 Plot of the top three principal components. The bacteriocin (*red*) and immunity protein (*blue*) are in close proximity to each other in every PC, while remaining close to the center of the CGFs (*green*). 41

A.1 Logo representation of original profile HMM. For each position the relative probability of having a set number. The logo was built using the Skylign tool at *skylign.org* Accessed: 03-04-2018 (J Wheeler et al., 2014). 55
A.2 Profile HMMs used to run the second (left) and third (right) iteration of the HMMsearch. The logos was built using the Skylign tool at *skylign.org* Accessed: 06-04-2018 (J Wheeler et al., 2014). 56
A.3 A histogram showing the correlations found between the bacteriocin proteins and the proteins found by clustering in the *Enterococcus* data set. 57
A.4 Correlation distribution between bacteriocin and the clustered proteins in the *Streptococcus* data set 58

List of Tables

2.1 E-value thresholds used to filter the results from the hmmsearch in each iteration for each for the three genome databases. The remaining sequences are then aligned in order to create the next pHMM. The third iteration marks the end of the search and are the thresholds set to filter the sequences for the result in section 3.2 18

3.1 A tabular representation of how many bacteriocin-like sequences were found in all genomes across the three databases. Note that at least 600 sequences are from *Enterococcus Faecium* in every iteration. 25
3.2 List of organisms that were unexpected according to personal communication with Dzung B. Diep (Diep, 2018). The table also lists if the sequences seem to be contaminations, and if so which species they are contaminated by. 26
3.3 The rate (%) of bacteriocin-carrying species compared to all species within the genera 26

A.1 A comparison of the class IIa bacteriocin carrying species we found and the ones that are listed in the Bagel4 database (*http://bagel4.molgenrug.nl/* Accessed: 2018-05-12) 54

Abbreviations

LAB	**L**actic **A**cid **B**acteria
Man-PTS	**M**annose **P**hosphotransferase **S**ystem
aa	**A**mino **A**cid
DNA	**D**eoxyribo**N**ucleic **A**cid
MSA	**M**ultiple **S**equence **A**lignment
HMM	**H**idden **M**arkow **M**odel
pHMM	**p**rofile **H**idden **M**arkow **M**odel
BLAST	**B**asic **L**ocal **A**lignment **S**earch **T**ool
MUSCLE	**MU**ltiple **S**equence **C**omparison by **L**og-**E**xpectation
PCA	**P**rincipal **C**omponent **A**nalysis
pMT	**pM**irror**T**ree
ORF	**O**pen **R**eading **F**rame

G	Glycine	Gly	P	Proline	Pro
A	Alanine	Ala	V	Valine	Val
L	Leucine	Leu	I	Isoleucine	Ile
M	Methionine	Met	C	Cysteine	Cys
F	Phenylalanine	Phe	Y	Tyrosine	Tyr
W	Tryptophan	Trp	H	Histidine	His
K	Lysine	Lys	R	Arginine	Arg
Q	Glutamine	Gln	N	Asparagine	Asn
E	Glutamic Acid	Glu	D	Aspartic Acid	Asp
S	Serine	Ser	T	Threonine	Thr

Chapter 1

Introduction

Antibiotic resistance has become one of the most substantial threats to human health since the introduction of penicillin to the scientific community by Alexander Fleming (1929). Although antibiotics are one of the greatest boons we have discovered, our over-reliance and usage of this substrate have put high selection pressure on bacteria. A pressure that forces the bacteria to either die out or develop a resistance to the antibiotic agent.

However, resistance has never been an unknown side-effect of antibiotic application, as even in the same year as it's first human trial, Abraham and Chain showed that certain bacteria are capable of exuding an enzyme, named penicillinase, capable of destroying penicillin (Abraham and Chain, 1940). Two years after this discovery was made, penicillin-resistant *Staphylococcus aureus* were found in hospitalised patients, from where it would spread to communities. In the decades to follow resistance became a widespread threat to the medical world, with new resistances arising in a multitude of bacteria. Realizing the severity of the situation, advice to adopt new policies regarding the use of antibiotics was submitted while at the same time emphasising the need for new forms of antibiotics (Neu, 1992).

The problem remains however, as each new antibiotic introduced have not staved off the underlying problem. The selection pressure is just too high for the bacteria to stay susceptible and to low too kill them outright (Ventola, 2015). In essence, we have to lower this selection pressure somehow. One way of combating this is by continually introducing new antibiotics to the playing field. However, until recently there had been a drought for new antibiotics, alerting us to how fragile such a method could potentially be (Conly and Johnston, 2005).

Luckily there are a few other ways of solving this. As mentioned, what forced us into this corner was our heavy reliance on antibiotics as our primary source of protection against infections.

However, this protection is a double-edged sword for the bacteria, as should we decrease our usage of antibiotics these resistant bacteria will be left with a defence mechanism that protects them from very little (Zaman et al., 2017). This would merely be a waste of energy compared to their unprotected brethren, who would thus out-compete the resistance carrying strains. This turn of events would then leave us with a more manageable set of resistant bacteria.

The problem with this method is that it would require a massive cooperative effort to implement regulations, forcing through decisions that would limit the use of antibiotics to only the most severe infections (Zaman et al., 2017). On top of this, the personal care industry would have to be upended, and the food industry would see significant losses in food, as they would have less functional ways of combating food spoilage.

Thus a more agreeable solution is to find alternatives. Until now most of the Anti Microbial Peptides (AMP) we have seen until now are substrates of fungi, but a lot of research has also been put into a different substrate - Bacteriocins.

1.1 Bacteriocins

Most antibiotics that are in use today are either a semi-synthetic variant of AMPs produced by certain genera of fungi, like penicillin, or entirely synthetic constructs. They arose in nature as a way to fend off bacteria, either from competing for resources or disrupting parasitic behaviour. Often they function on a broad array of bacteria and will try to hinder or slow down biological processes in the target cell. For example, tetracycline is a widely used antibiotic that prevents association of the aminoacyl-tRNA with the bacterial ribosome by binding to the A-seat in the 30S ribosome, hence hindering the protein synthesis which in turn results in a weakened or dying cell (Chopra and Roberts, 2001).

However, bacteria are not only a problem for lifeforms more advanced than themselves. Just as all other lifeforms compete for nourishment with their equals, so do bacteria. Hence the production of antimicrobial agents is not foreign to them. These AMPs tend to be more potent than their eukaryotic counterparts, at the cost of having a much more narrow spectrum of antibiotic activity (Diep et al., 2007). This essentially means that they are a product made to fend off bacteria competing for the same resources as the producer. Seeing how these bacteriocins are ubiquitous, they have risen to high popularity in the last few decades as we might use them to combat bacteria that are either immune or developing immunity to standard antibiotics (Cotter et al., 2012).

Most bacteria produce some form of bacteriocins, be they gram-positive or gram-negative (Jack et al., 1995), though in this study we will focus on the former. More specifically we focus on bacteriocins produced by Lactic Acid Bacteria (LAB) as they are by far the most popular organisms when it comes to studying bacteriocins. The reason for this is that LAB are generally recognised as safe (GRAS) organisms, which means they are less regulated than other species (König and Fröhlich, 2017). And now we seek to harness them in the fight against antibiotic resistance (Drider et al., 2006).

Bacteriocins from gram-positive bacteria are usually grouped into three main classes, class I, II, and III (Rea et al., 2011). Class I bacteriocins consist of short bacteriocin (16-28 amino acids long) known as lantibiotics. These bacteriocins are characterised by a high rate of post-translational modification, through which they obtain special amino acids like the meso-lanthionine and β-methyllanthionine before they become active AMPs. Several different sets of subclasses have been proposed for the the Class I bacteriocins based on shape (Jung, 1991), sequence similarities (Piper et al., 2009), structure similarity (Heng et al., 2007), post-translatory modifications and antimicrobial activity (Willey and van der Donk, 2007), as well as proposing to not divide the class into any subclasses (R., 1993).

Next up, class II bacteriocins. This class consists of bacteriocins of a more variable length (36-49 aa), usually with a mass below 10 kDa. Other than mass, they differ from class I bacteriocins by not undergoing any post-translational modification. There is a higher scientific consensus for the classification of the class II bacteriocins, and most agree on the two major groups, class IIa and class IIb. Class IIa is named the Pediocin-like bacteriocins, and consists of, as the name suggests, bacteriocins like Pediocin Pa-1. Class IIb encompasses the two-peptide bacteriocins, class IIc contains the circular bacteriocins, and class IId is for any class II bacteriocins that do not fit into any of the other categories. The last two sub-classes were proposed by Cotter et al. (2005), and have since remained in use (Nissen-Meyer et al., 2009).

Class III bacteriocins are large (<30 kDa) heat-labile proteins, and similar to class II they do not undergo any post-translational modification. The classification of these bacteriocins has been debated within the scientific community, as certain members of this family operate as lytic enzymes. In general, the class is divided in two. Class IIIa is known as bacteriolysins as they cause cell lysis and class IIIb for the bacteriocins that do not. (Joerger and Klaenhammer, 1986).

Online databses exist specifically for bacteriocins, for example the BAGEL databases (van Heel et al., 2013) and BactiBase (Hammami et al., 2010). Both of these servers allow for genome-mining for bacteriocins, and to download profiles needed to quickly annotate bacteriocins in a

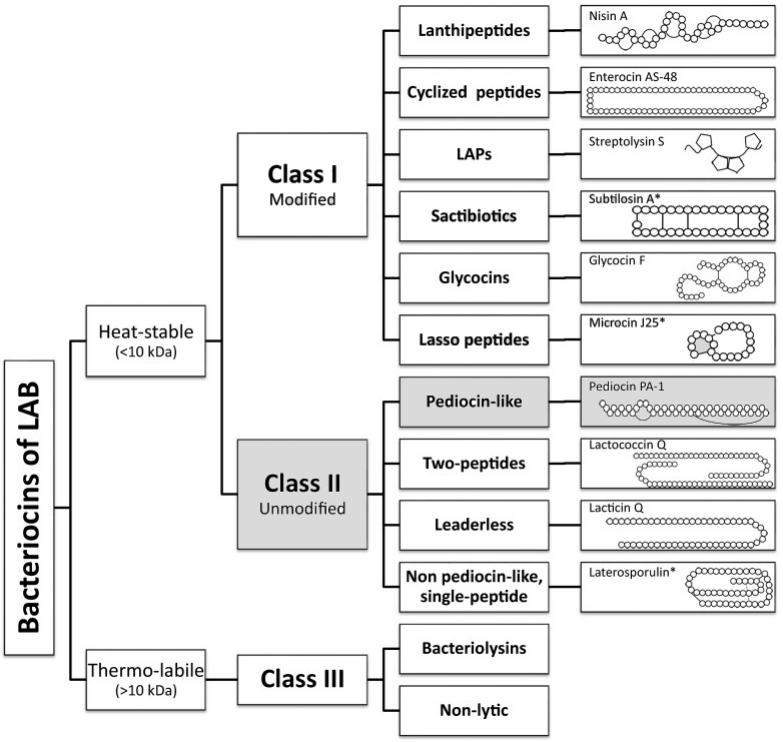


FIGURE 1.1: Overview of the different classes of bacteriocins produced by LAB (Alvarez-Sieiro et al., 2016)

genome.

1.1.1 Class IIa: Pediocin-like bacteriocins

The focus of this study is as the title suggests class IIa bacteriocins. This is a group of bacteriocins labelled Pediocin-like bacteriocins, due to their likeness to the first bacteriocin discovered within this class; Pediocin PA-1 (Biswas et al., 1991). These bacteriocins have shown to be active against *Listeria*, *Enterococcus*, *Carnobacterium*, *Lactobacillus*, *Leuconostoc*, *Pediococcus* and *Clostridium* species (Eijsink et al., 1998). Usually, these bacteriocins are separated into two domains, divided by a flexible hinge structure (Haugen et al., 2005). The N-terminus domain consists of a heavily conserved YGNGV/L-motif called the "Pediocin box" (Ennahar et al., 2000), which usually takes on the structure of an anti-parallel β-sheet. This structure is stabilised by a conserved Cys-Cys disulfide bridge. The second domain is located at the C-terminus structured as one or two α-helices that form a hairpin or a functionally equivalent helixhingehelix structure (Kjos et al., 2011). This region is usually where we see the most variability within the class, and has been used to further divide the class IIa bacteriocins into subgroups (Nissen-Meyer et al., 2009). This hairpin structure penetrates into the hydrophobic parts of the target membrane, causing cytosolic leakage eventually leading to cell death. In some bacteriocins, this

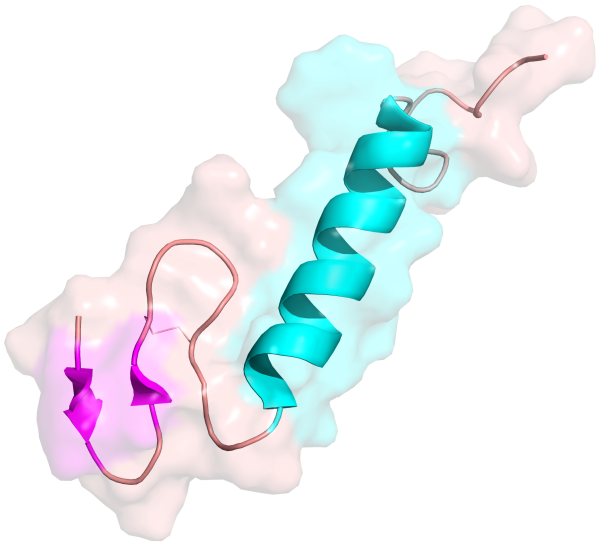


FIGURE 1.2: The three dimensional structure of the membrane bound class IIa bacteriocin Sakacin P. Moving left to right, the N-terminal β-sheet domain is marked in pink ahead of the Cystein disulfide bridge. Between the two Cys residues is the hinge-region connecting the two separate domains. The C-terminal domain α-helix is marked in *cyan*. (PDB ID: 1OG7)

hairpin structure is stabilised by the presence of yet another Cys-Cys disulfide bridge. The three-dimensional structure of a class IIa bacteriocin, Sakacin P, is shown in Figure 1.2. Note that Sakacin P does not contain the latter disulfide bridge. One region that is not shown in the figure is the leader-sequence of the pre-mature bacteriocin. This is a short peptide commonly 18-27 amino acids long (R., 1993), which is recognized and cleaved off by the ABC transporter during secretion from the cell (Drider et al., 2006).

1.1.2 Interraction with the Mannose Phosphotransferase System

Class IIa bacteriocins kill bacteria by creating pores in the cytoplasmic membrane through an interaction with the sugar transporter Mannose Phosphotransferase System (Man-PTS). These transporters are a part of the phosphoenolpyruvate transport systems (PTS), which are characterised by simultaneously phosphorylating sugars while importing them (Postma et al., 1993). The Man-PTS consists of four subunits: IIA, IIB, IIC, and IID. IIA and IIB act together, often in the same protein, in the cytoplasm, while the IIC and IID subunits are individual transmembrane proteins, responsible for the transportation of the sugars. Kjos et al. (2011) shows that it is these subunits that are targeted by the class IIa bacteriocin. Specifically,

there is an extracellular loop on the IIC subunit that is responsible for targeting by the class IIa bacteriocin (Kjos et al., 2010). After the N-terminal region binds to the transporter, the hydrophobic hairpin structure at the C-terminus penetrates the IIC component, forcing it open resulting in a pore for cell leakage. Thus, if a targeted bacteria is to survive in an environment filled with bacteriocins, it needs to rid itself of the receptor proteins. The cell does this through endocytosis, by retracting the Man-PTS into the cell for destruction in the lysosomes. It will also need to down-regulate the expression of the Man-PTS operon, thus developing a resistance to the bacteriocin. However, by doing this, the bacteria is now weakened compared to their competitors. Therefore they will be more likely to be outcompeted. As bacteriocins target bacteria that are genetically close to the producer, the producer has to have some form of separate defence mechanism, else it falls into the same trap as the target cells.

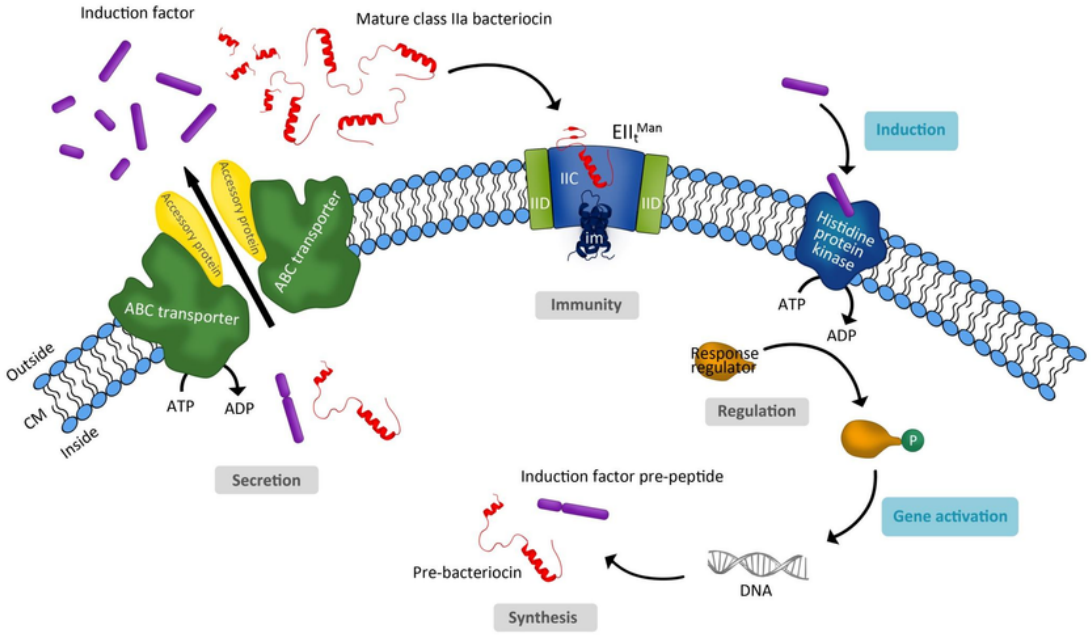


FIGURE 1.3: The complete pipeline of a class IIa bacteriocins life-cycle, from synthesis to attachment to the Man-PTS. (Ennahar et al., 2000)

1.1.3 Immunity Proteins

This defence comes in the form of a protein that renders the producer immune to the bacteriocins antimicrobial activity. This immunity protein is synthesised by a gene directly downstream from the bacteriocin 1.4. These genes are expressed concomitantly, meaning that they are produced together, and are regulated through quorum sensing (Drider et al., 2006). This implies that these bacteria become sensitive to their own bacteriocins should production seize. This exact occurrence has been observed with the Sakacin P producing *Lactobacillus sakei* LTH673 strain,

as it becomes more sensitive to its own bacteriocin when put in an environment where bacteriocin expression is down-regulated (Fimland et al., 2002). The proteins themselves range in sizes from 81 to 115 amino acids, although in contrast with the bacteriocins, which have distinctly defined motifs, these proteins do not share any clear consensus sequences (Kjos, 2012).

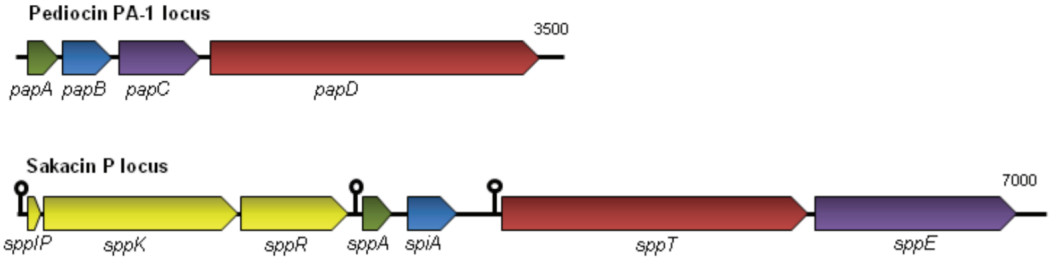


FIGURE 1.4: Layout of bacteriocin operon. The structural bacteriocin is marked *Green*, directly following this is the immunity protein (*Blue*). Following the immunity protein we find the accessory protein (*purple*) and transport proteins (*red*). The order of these last two tend to vary. In some bacteriocin operons there are regulatory genes incorporated into the operon. (*yellow*) (Kjos, 2012)

The proposed mode of action for these proteins is presented in Figure 1.5. The protein is thought to bind to the intercellular part of the Man-PTS and in some way hinder the pore-formation by the bacteriocin. It is unknown whether the immunity protein interacts directly with the bacteriocin, however, Johnsen et al. (2005) have found that the C-terminal region of the bacteriocin is involved in specific recognition of the C-terminal half of the immunity protein.

One last thing that should be mentioned is that the *L. sakei* LTH673 strain discussed above carried another immunity gene, *OrfY*, which inferred resistance to multiple class IIa bacteriocins, not except its own Sakacin P (Brurberg et al., 1997). This immunity protein is situated without any corresponding bacteriocin but was co-regulated along the regular bacteriocin operon. Since then several other similar immunity proteins have been found in different species (Quadri et al., 1994; Métivier et al., 1998). The presence of these immunity proteins could indicate that the same development that has been seen with multi-drug-resistant bacteria could happen with bacteriocin sensitive bacteria should a functional application be put forward (Drider et al., 2006).

1.2 coevolution

Apart from bacteriocin, it is necessary for the reader to have a good grasp on what coevolution is, and what this means on the molecular level. Evolutionary theory is the study of how

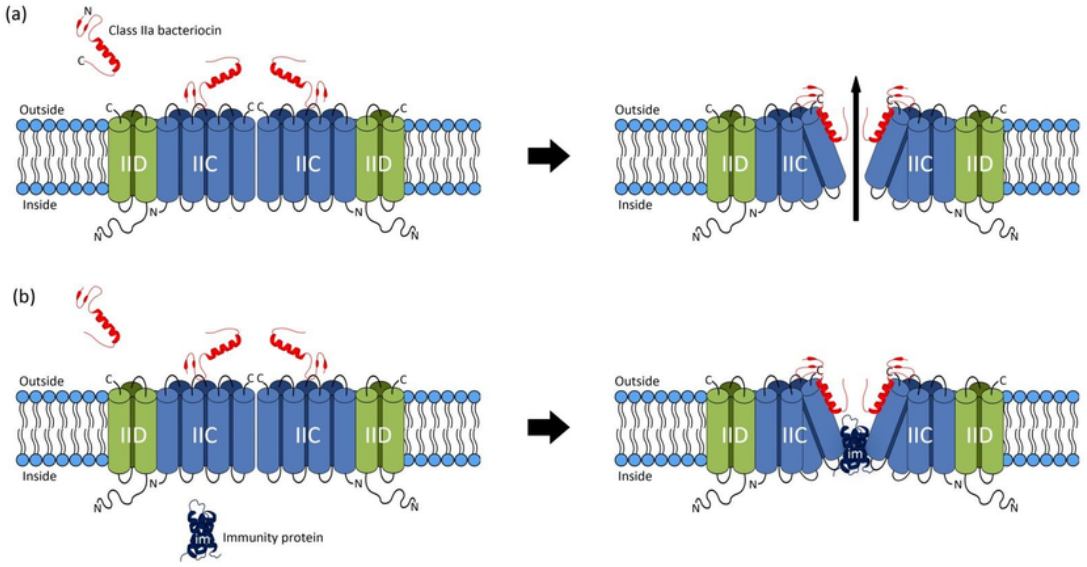


FIGURE 1.5: "Proposed model of the mode of action (A) and immunity (B) in class IIa bacteriocin systems. The class IIa bacteriocin is depicted in red. (A) The N-terminal -sheet containing part of the bacteriocin initially interacts with an extracellular loop (highlighted) of the man-PTS IIC protein. The helix-containing C-terminal part then engages in specific interactions with transmembrane helices of the IIC and/or IID proteins to cause conformational changes which, in turn, lead to pore formation and eventually cell death. (B) In immune cells, the bacteriocin mediates the same conformational changes, but the pore is blocked by a specific immunity protein (blue) which binds tightly to man-PTS." - Kjos et al. (2011)

living organisms change and adapt over time, based on this John N. Thompson (1994) defines *co*evolution as:

> The reciprocal evolutionary change in interacting species

What this means is that changes in one organism affect and induce changes in another. A simple example is the coevolution of prey and predators. Whenever a new mutation arises in a predator - one that increases fitness - the selection pressure on the prey increases. This will, in turn, cause the prey to evolve or die out. Should it develop to be more fit, the selection pressure will shift back unto the predators, and a cycle of coevolution is formed. In this sense coevolution is natures arms race. From this example, we see that any community of interacting entities can build coevolutionary dependencies, therefore, this applies to proteins as well.

Coevolution between proteins happens as a result of the interconnected systems that they are a part of. As these interactions are based on the sequence and structure of the proteins, changes to the structure can potentially alter the interactive factor between them. As with the example above, interacting proteins will affect each other as they evolve. As an effect of this coevolution, we would generally think that the evolutionary history of interacting proteins is related, which proved to be the case. It has been found that interacting proteins often share similar phylogenetic

trees (Fryxell, 1996). Coupled with the rise of next-generation Sequencing, this knowledge has led to a variety of different methods to measure coevolution (de Juan et al., 2013), discussed further in the Methods.

1.3 The goal of this study

In this study we intend to study the evolutionary dependency between the class IIa bacteriocin and their immunity proteins. Our hypothesis is that the two proteins are coevolved as the proteins holds a defense-attack interaction. This means that should one change too much, this interaction may be halted, thereby killing the bacteriocin producer. We also hypothesize that the non-conserved C-terminal ends hold the most influence on this relationship, as earlier tests have shown that the C-terminals of either protein are involved in recognition of each other. Should we be able to discard the null-hypothesis, this would be an indication that the two proteins interact directly, which may help in the ongoing effort to fully understand the mode of operations of class IIa bacteriocins and their immunity proteins.

Alongside this goal, we also hope to further annotate bacteriocins through genome-mining.

Chapter 2

Methods

The programming language R has been used to carry out every analysis and procedures in this study. R is a freely available programming language specialising in statistical studies and can be downloaded from *http://www.r-project.org/*.

2.1 Sequence Alignments

In bioinformatics, there are few algorithms used as commonly as sequence alignment, and this thesis is no outlier. Most tools used in this thesis apply some form of sequence alignment, multiple or pairwise. Thus it is necessary to understand what sequence alignments are.

We align sequences when we are looking for similarities between them. DNA and protein sequences are made up of either four different nucleotides [A,C,G,T] or 21 different amino acids [A,B,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y]. And the further back in time two species split off from one another, the more these sequences will differ. This is a prime principle in evolutionary theory, which allow us to estimate the relation between sequences based on how different they are, up to a point. A multitude of techniques use this principle, whether it is to detect homologs or construct phylogenetic trees, sequence alignment is at the core. First off, the pairwise sequence alignment.

2.1.1 Pairwise Sequence Alignments

When talking about pairwise sequence alignments, there are two primary algorithms that are at the core of the bioinformatic landscape, the Needleman-Wunsch and Smith-Waterman algorithms. Both represent a method for aligning a query sequence to a reference, but where they

differ is that the prior returns an alignment of the query mapped to the entirety of the genome, a global alignment, while the latter returns the query aligned only to the best matching region of the reference, a local alignment. Global alignments performed today build on the foundation laid by the Needleman-Wunsch algorithm, named after it's creators Saul B. Needleman and Christian D. Wunsch in 1970 (Needleman and Wunsch, 1970). The algorithm is intended for use when the query and reference are of approximately the same size, and it works by building a matrix from the two sequences represented as the rows and columns. From there it uses a scoring system to determine the score of each cell derived from the previous diagonally aligned cell, and whether the letters in the current column and row match. Gaps can be introduced to the alignment if a cells score is derived either from the previous horizontal or vertical cell. Introducing gaps usually harms the score more than a mismatch as insertions/deletions are biologically more taxing to an organism. After that, the optimal global alignment is inferred by tracing the steps from the last cell back to the starting cell, which is a cell that matches an empty start row and column in position [1,1]. A global alignment of a query sequence of GCA to the reference ATGCAGG would end up like this.

$$\begin{array}{cccc} 1' & ATGCAGG & 7' & Reference \\ & --GCA-- & & Query \end{array}$$

However global alignments are intended to be used when the query and the reference are of similar length. If this is not upheld and we align a short query to a long reference, the global alignment would introduce a high amount of gaps and the query might be stretched between several matching regions. Such errors are unacceptable as we will compare short sequences to large genomes to find potential homologs. For such a task the local alignment is a more surefire approach. Local alignments build on the work of Temple M. Smith and Michael S. Waterman in 1981 (Smith and Waterman, 1981), named the Smith-Waterman algorithm, which proposes never to allow any cells value to go below zero while allowing every cell with the value zero to be a potential start cell. Thus the matching regions between to sequences becomes more clear, and the optimal local alignment is easily found by backtracking from the cell with the highest score to its starting cell. If we align the same sequences as before we would now only get the matching local region.

$$\begin{array}{llll} 3' & GCA & 5' & Reference \\ & GCA & & Query \end{array}$$

Both of these algorithms perform with a time complexity of O(nm) where n is the length of the query and m is the length of the reference. It should be noted that both of these algorithms have been improved upon and are only the foundation of what is in use today, and most alterations are related to reducing running time and space complexity.

As a final note, there are also methods that combine the two outlined methods, called semiglobal methods, where a global alignmnet is performed on a local segment of either sequence.

2.1.2 BLAST

At a later point, we will align a plethora of protein sequences together in an attempt to detect homologs. The tool we will be using for this is BLAST, and so an explanation is in order.

Basic Local Alignment Search Tool (BLAST) is a heuristic tool designed to quickly identify local alignments between a query sequence and one or more reference sequences (Altschul et al., 1990) by the use of pairwise sequence alignment. As this is a heuristic approach to aligning sequences, it does not look for the optimal alignment. Instead, it scans for the approximate optimal alignment. How it achieves this is by first breaking the query sequence into k-letter words (*k-meres*) of length 3 for proteins and length 11 for nucleotides by default. This means that for the amino acid sequence MDPQVTR the *3-meres* would be [MDP, DPQ, PQV, QVT, VTR]. The next step is to score every possible *3-mere* by how well they align with these *3-meres* using a standard substitution matrix for amino acids (commonly BLOSUM62), and a mismatch scoring system for nucleotides. In all, there are 20^3 amino acid *3-meres* that need to be scored against each word, and 4^{11} nucleotide *11-meres*.

3-meres that score over whichever threshold has been set are taken to the next step, wherein the *3-meres* are used to search through a database in the blast format for exact matches. Next, the short alignments are extended in both directions in an attempt to increase the alignment score. This continues until the score dips below the previously set threshold and the alignment up until that point is returned with the other potential alignments.

Other forms of BLAST exist that could be useful to the analysis, such as PSI-BLAST which is designed to find distantly related homologs of a sequence by an iterative process where it

integrates the results from a blast into the query before blasting again. We have decided to incorporate a similar approach using HMMER instead of BLAST, a choice that is discussed in section 4.1.2.

2.1.3 Multiple sequence alignments and MUSCLE

While aligning two sequences together is an essential and ubiquitous task in bioinformatics, we often also want to look at more than two homolog sequences at the same time. This is important as it is the most detailed representation of evolutionary relationships between sequences we are currently able to create. This information is applied in a multitude of methods such as in the creation of phylogenetic trees to display the evolutionary distance within the alignment, and for example to produce profile Hidden Markow Models as described in section 2.1.4.

However, attaining an optimal multiple sequence alignment is less straightforward when compared to pairwise alignment. While finding the optimal alignment between two sequences is performed in a two-dimensional space, aligning n sequences requires an n-dimensional space. This increases the workload exponentially, and thus a "naive" approach to finding the optimal alignment of multiple sequences quickly becomes impossible. As such there have been many algorithms dedicated to increasing speed at the cost of accuracy.

There are four different branches of these methods, of which we employ two: The iterative methods, and Hidden Markow models discussed in section 2.2. Iterative methods build on the progressive approach of making a guide tree through clustering and combining the pairwise alignments in order of the most similar pairs. The main difference is that while progressive methods never revisit aligned sequences, the iterative method allows for realignment of previous partially aligned sequences.

The iterative MSA tool used in this study has been Muscle v.3.8.31 (Edgar, 2004a) as it produces reliable alignments at little cost to accuracy. The process works in these stages (Edgar, 2004b):

1. *k-mere* clustering is performed to create the guidance tree. This is a quick way to produce a guidance tree, at the cost of accuracy when compared to the pairwise identity trees used in for example ClustalW. It works by calculating the distance from counting and comparing *k-meres* of all sequences.

2. A progressive alignment method is used to align pairwise sequences to each other, sequences to profiles or profiles to profiles as the alignment process moves closer to the root of the guide tree. This process results in a rough MSA.

3. A new guide tree is produced based on the pairwise sequence identities of the alignments. This is done as the robustness of the original guide tree is unreliable, due to the *k-mere* clustering. The progressive method is then re-applied to build the MSA until the alignment is stabilised.

4. The last step is known as "Tree dependant refinement". In this step, the tree is split at random into to two sub-trees, whose profiles are aligned in the same pairwise alignment fashion as before. The resulting MSA is kept should the alignment score be higher than the previous one.This step is repeated a set number of times (default 16) or until convergence is reached.

2.1.4 HMMER and Hidden Markow Models

A relevant tool that uses the MSAs produced by, for example, MUSCLE is HMMER. This tool uses the inherent probabilistic properties of Hidden Markow Models (**HMM**) to align and score sequences based on similarity to the original MSA.

To fully understand how this works, it is crucial to understand Markow chains, and their sub-category Hidden Markow Models. The defining feature of a Markow Chain is that the probability of moving from one state to another is solely based on what position one is currently, disregarding past movements. A sound example of this would be a game of dice. For simplicity's sake, this game of dice will be played using only one die. This die is fair, and thus each result has the same probability of occurring, 1/6. This probability stays the same no matter what the results have been before. Should you roll any given sequence of numbers in a set of throws, the likelihood of the next toss would stay the same; 1/6th for all sides. Thus a game of dice can be viewed as a Markow-chain.

A Hidden Markow Model, however, does not revolve solely around endlessly repeating the same task. The difference is made by the addition of hidden states where, between each iteration of the process, there is a chance of to a different "state" with its own set of probabilities. In the book, Biological Sequence Analysis (Durbin et al., 1998), a similar casino as described above is given. However, this time we visit an "occasionally dishonest" casino. In this casino, the game of dice is mostly played as before, but there's a twist. Every time someone throws a die in this casino, there is a slight chance that the casino will switch out the fair die for a loaded die. By doing so the the gambler has entered a "hidden" state with a new set of probabilities (Figure 2.1). And with each throw the gambler makes there is a new chance that the casino, in fear of being ousted, switches out the loaded die for the fair die, thus moving the gambler back into

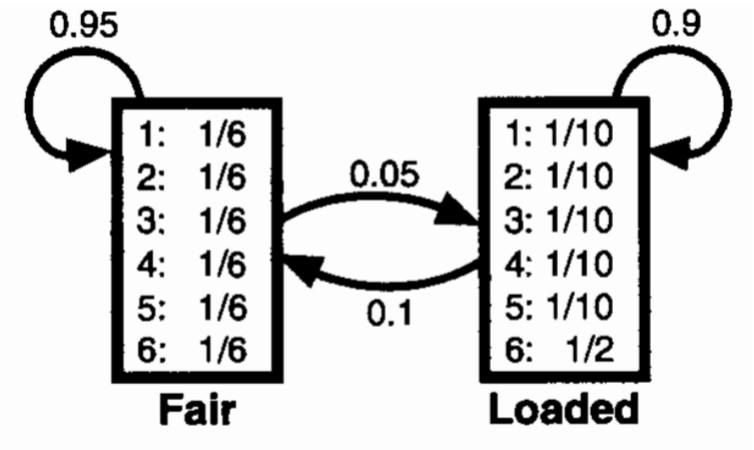


FIGURE 2.1: The Hidden Markow Model of the occasionally dishonest casino.

the "fair" state. And should the casino ever add more dice, or "states", these would then have probabilities for the gambler to jump between them.

If we are to put these examples in a bioinformatical context, the first example would be the same as if we represent a protein sequence as a set of probabilities of each amino acid occurring, with no thought for which "state" we are in. The second example is more intriguing, as for each amino acid expressed by the model, we would have a chance to move to a different set of probabilities. This hidden state resembles insertions, which are random additions to the DNA that neither weakened nor killed the host. This is however not a good enough simulation of a given sequence, as all states have the potential of acting as insertions.

Thus another model is proposed, a profile Hidden Markow Model. In this model, each of the positions of an MSA that describe more amino acids than gaps is represented as "main" states each containing the probabilities of the amino acids found at that position. These states express only one element before moving to the next state or one of the hidden states, either the insertion-state or a deletion-state. The probability of migrating to any of these states is inferred from the MSA, where a position with more gaps than amino acids are expressed as an insertion state, while the probability of expressing the deletion state is found by the rate of gaps to amino acids in a position with more amino acids than gaps. The likelihood of moving to the next main state is described by how many of the sequences that forego insertions or deletions at this point. Lastly, pseudo counts are added to all amino acids so that the profile can describe sequences outside the ones it was built from. This profile can then be used to compare a sequence to the MSA based on the probabilities of the profile.

Concretely, the HMMER3 software can utilise this latter method by first building a profile from an MSA and then search through either nucleotide or amino acid references to look for matching sequences. These sequences are areas of the reference that have a high probability

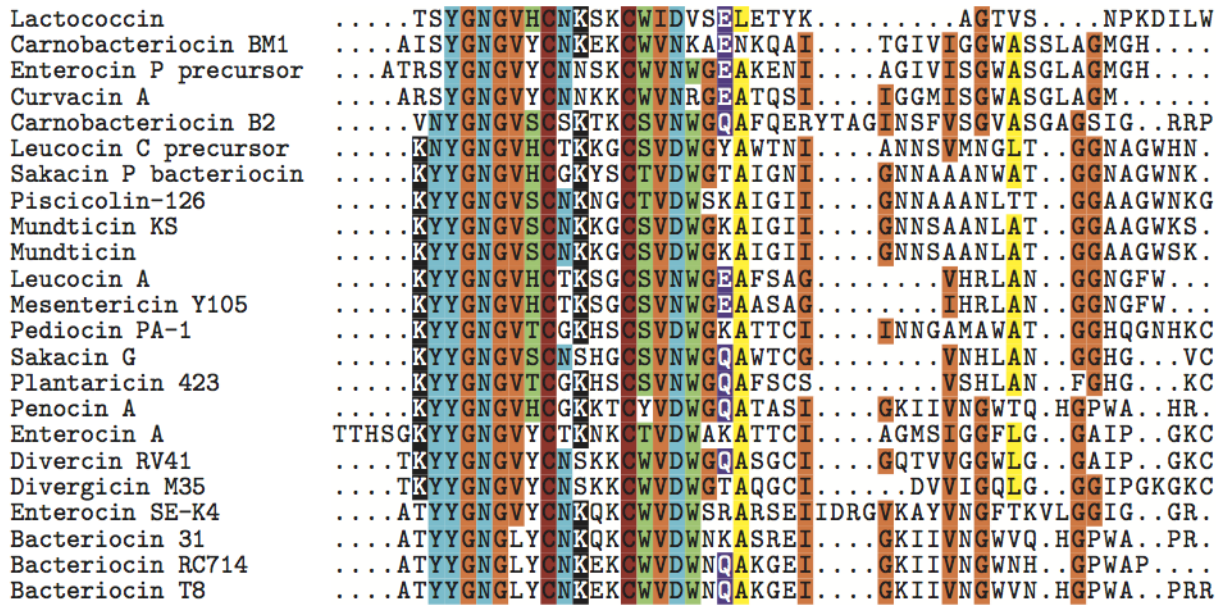


FIGURE 2.2: The MSA of the original 23 sequences used. The sequences are a subset of the bacteriocins listed in the review article by J. Nissen-Meyer et al. Nissen-Meyer et al. (2009).

of being expressed from the profile, and thus a high likelihood of being a homolog to the seed sequences of the profile.

2.2 Locating bacteriocins with HMMER3

HMMER v.3.1b2 is used throughout this study to find potential homologs of the known class IIa bacteriocins. As stated in the user guide (Sean R. Eddy, 2015):

> HMMER is used to search sequence databases for homologs of protein or DNA sequences, and to make sequence alignments.

HMMER does this by utilizing the probabilistic properties of a Hidden Markow Model (HMM).

2.2.1 Choice of sequences

For us to even be able to craft a profile HMM, we first need a set of known bacteriocin sequences. As such a set of sequences listed in a review article by Nissen-Meyer et al. (2009), detailing the class IIa bacteriocins were chosen. These sequences were then aligned with Muscle, and the resulting MSA (Figure 2.2) was used to build an amino acid profile HMM with HMMER3s built-in function hmmbuild.

2.2.2 The NCBI database

HMMER3 needs sequences to test against, and for that, we need raw data. For us, this entailed downloading all publicly available prokaryotic genomes at the American National Center for Biotechnology Information (NCBI) (Coordinators, 2016) available on the 1st of January 2018. NCBI hosts the Genbank database which is "a comprehensive database that contains publicly available nucleotide sequences for more than 260 000 named organisms" (Benson et al., 2008), as well as the Reference Sequence (RefSeq) database (O'Leary et al., 2016) containing hand-picked and annotated DNA, RNA, and protein sequences. Individual laboratories submit these sequences, and through a collaboration known as the International Nucleotide Sequence Database Collaboration (INSDC) (Cochrane et al., 2016) they share data with other datacentres around the world. The major contributors are the European Bioinformatics Institute (EBI) (Emmert et al., 1994), which hosts the EMBL-database who collects European sequences, and the DNA Data Bank of Japan (DDBJA) (Kodama et al., 2018), which is the only genetic data bank in Asia.

2.2.3 Iterative Hmmsearch

As our profile was built up from protein sequences, we had to rely on HMMERs built-in hmmsearch function to find homologs. Hmmsearch takes input amino acid sequences and aligns them to an input pHMM, and stores the sequences with a hit above whatever E-value threshold we have set. For our search, this threshold was set to the default value of 0.05, as not to miss any potential bacteriocin sequences hiding within the lower values. To increase the speed of the algorithm as well as reducing the necessary memory usage we used the micropan-package (beta-v.1.1.2) in R to locate all leading Open Reading Frames (ORF) sequences in the genome with a nucleotide sequence length between 70 and 1000. The algorithm then translated and aligned these sequences to the profile using *hmmsearch* as described above. The resulting sequences were then filtered at three separate E-values for the three sets of genomes, for each of the three separate iterations that we ran (table 2.1). For each iteration, a new profile was built using a subset of the resulting leader-less bacteriocins (leader peptides were removed manually at the GG-motif) from the previous iteration. Seeing as certain species are over-represented in the databases, we decided to only use the most common sequence for each species. The profiles are shown in Appendix figure A.2. For the third iteration, we added organisms that were missing from the original alignment back into the MSA. This decision was made to circumvent what we now know was a bug in the micropan package, further discussed in section 4.1.4. The workflow

is shown in figure 2.3. Similar approaches can be found in other studies where HMMER has been used for genome mining (Morton et al., 2015).

Iteration	Complete	Scaffolds	Contigs
1	$1e-5$	$1e-7$	$1e-6$
2	$1e-11$	$1e-12$	$1e-12$
3	$1e-13$	$1e-10$	$1e-10$

TABLE 2.1: E-value thresholds used to filter the results from the hmmsearch in each iteration for each for the three genome databases. The remaining sequences are then aligned in order to create the next pHMM. The third iteration marks the end of the search and are the thresholds set to filter the sequences for the result in section 3.2

2.3 Locating immunity proteins

With the knowledge of a bacteriocins location in a genome, and the knowledge that the immunity gene is usually the first ORF either upstream or downstream from the bacteriocin gene (Figure 1.4), a short workflow was developed:

1. Extract the downstream sequence of the 1000 next nucleotides from the genome fasta file.
2. Use the micropan package to create a gff-table of all LORFs in this sequence
3. Look for the first sequence with a length of 50 amino acids.
4. If none are found repeat the procedure for the upstream sequence of equal length.
5. If nothing is found return null

2.4 Measurement of co-evolution

There are multiple methods for measuring co-evolution between proteins(de Juan et al., 2013), and a traditional set of techniques are known as mirror-tree methods. At their core, they all revolve around computing a correlation statistic between two phylogenetic trees built from separate protein distances. The idea is that if two proteins are co-evolved a mutation in one of the proteins will incur a change in the other, lest they stop functioning in tandem. From this, we can infer that the phylogenetic trees of coevolved proteins will look similar to each other, as a change, or branch, in one tree will incur a branch in the next tree.

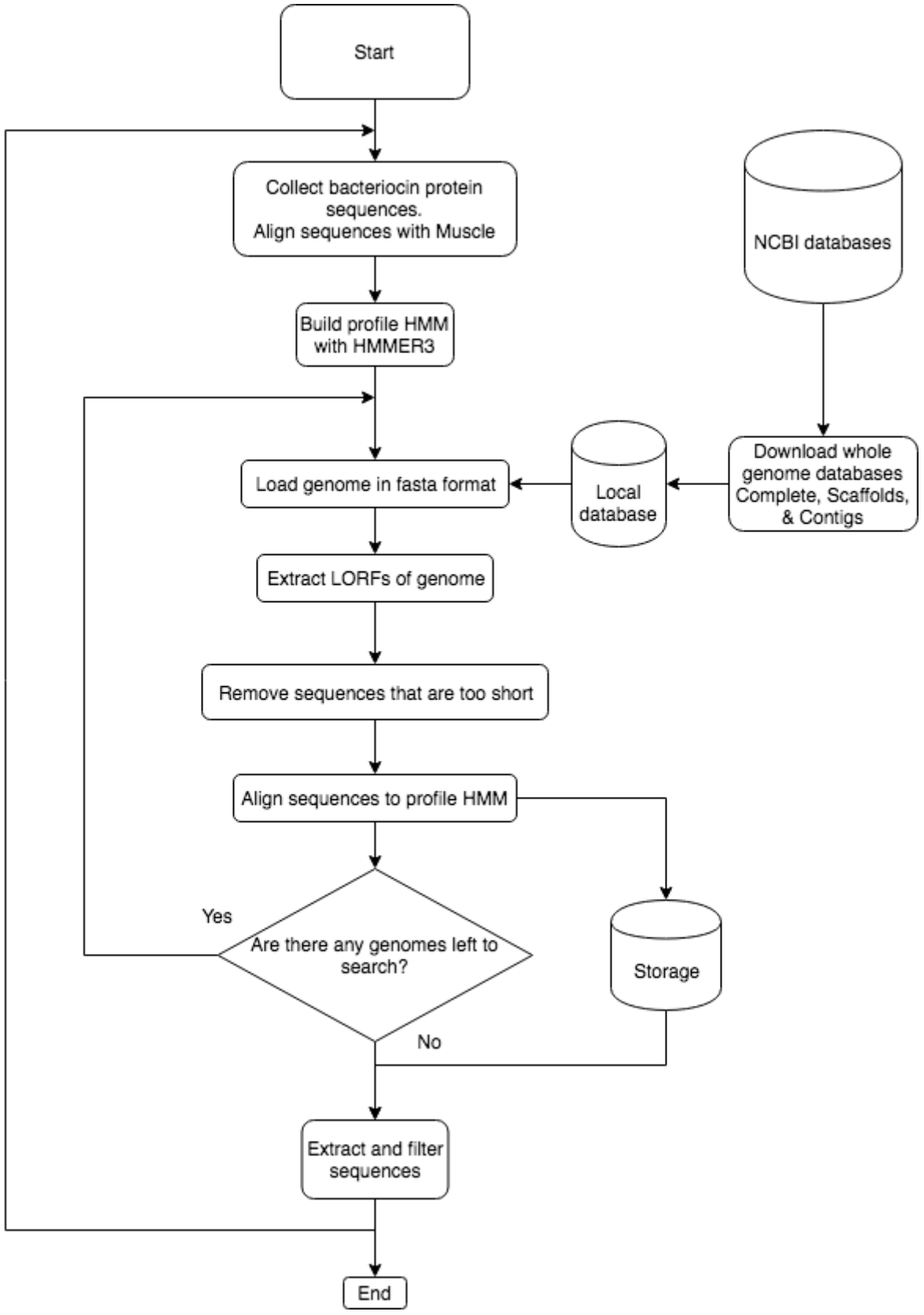


FIGURE 2.3: Implementation of itereative search process

We apply two forms of mirrortree methods, one using the MirrorTree web server(Ochoa and Pazos, 2010), and the other is a variation of the pMirrorTree described in Detection of significant protein co-evolution (Ochoa et al., 2015).

To perform these methods, we craft two data-sets containing both bacteriocins and immunity proteins. The first data-set contains every unique bacteriocin sequence found in *E. faecium* and *E. faecalis*, provided that the immunity proteins are also found for said sequence. The second set contains every unique bacteriocin and immunity protein found in the *Streptococcus* genus.

2.4.1 Mirrortree

The simple mirrortree application infers distances between organisms from the phylogenetic trees and calculates the linear Pearson correlation between the two emerging distance vectors. The Pearson correlation is a measure of linear correlation between two variables and ranges from 1 to -1. The correlation is calculated by equation 2.1 (Pazos and Valencia, 2001).

$$Cor_{[R,S]} = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i - \bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}} \tag{2.1}$$

R_i is the branch length of comparison i in tree R, while S_i is the branch length of the same comparison in tree S. $\bar{R}$ and $\bar{S}$ are the average branch length of all comparisons in tree R and S respectively. Note that the Pearson correlation looks at the linear correlation between two sets of data, so even though the branches in one tree are longer than the other, the correlation might still fall to 1 or -1.

The web server also returns a P-value score for each correlation run based on tabulated P-values from earlier runs where the interactions were known (Pazos and Valencia, 2001).

We supplied MSAs of the *Enterococcus* data sets for the first run, and another run with the *Streptococcus* data set. Lastly, a series was performed with an MSA of each species that resulted from the hmmsearch.

2.4.2 P-MirrorTree

Next an interpretation of the P-MirrorTree (pMT) approach, described in the paper *Detection of Significant Protein Coevolution* (Ochoa et al., 2015), was applied to the data. The paper

posits that the significance assessments of the correlations found by previous approaches fail to properly adjust for the inherent co-dependencies within the trees, as they state:

> On the methodological side, it is well known that the internal codependencies between the values of distances matrices burden the significance assessment of a given correlation coefficient. The tabulated P values, regularly used to assign correlation significances, assume the independence of the vectors components. Since the distances in the phylogenetic trees cannot freely change to adopt any possible value, strictly speaking, these P values are not adequate, in spite of having shown an improvement on the interaction prediction (Juan et al., 2008).

Secondly, the branch lengths are potentially constrained to have similar values due to the low evolutionary distance between the organisms that are measured. This could cause two non-coevolving sequences to appear coevolved merely because the organisms are evolutionary close. Another issue is that mirrortree methods have been shown to increase performance when trees made from data with a non-redundant taxa-pool is used (Herman et al., 2011)(Muley and Ranjan, 2012). As a result, most studies of co-evolution have discarded the use of significance levels when looking for protein co-evolution.

To alleviate these problems Ochoa *et al.* devised the P-MirrorTree method, wherein the correlation between the target sequences are measured against a set of trees whose branches have been swapped with each other in an iterative process. The shuffling procedure should produce trees with correlations centred around values expected of non-coevolving proteins given the evolutionary distance between the organisms. Thus a null distribution is formed and will allow us to infer an empirical P-value for the correlation found between the target sequences.

We apply a similar method however we will be using the evolutionary distances directly instead of building a tree and then inferring the distances between proteins from the branch lengths. This is done to eliminate the information loss incurred when constructing a phylogenetic tree, whether it be done by Neighbour-Joining or Maximum Likelihood methods. This is achieved by vectorising every distance matrix for every protein, and adding these vectors to a new data frame, a "super" distance matrix if you will.

To achieve this, we will first have to find all the proteins the organisms have in common. The following steps are heavily inspired by the case study contained within the micropan package documentation (Snipen and Liland, 2015).

First, we use Prokka v.1.12 (Seemann, 2014) which is a program designed to find and annotate all potential proteins within a genome. The resulting files containing every protein found for all

genomes will then be blasted against each other by the use of the BlastAllAll function found in the micropan package. This should then result in a matrix containing all hits, with associated bit-scores. This bit-score will allow us to calculate the blast distances between all sequences according to equation 2.2.

$$bDist_{i,j} = \frac{1}{2}\left(2 - \frac{S_{i;j}}{S_{i;i}} + \frac{S_{j;i}}{S_{j;j}}\right) \tag{2.2}$$

Where $S_{i;j}$ is the bit score when sequence i is blasted against sequence j as the reference.

These distances can then be used to cluster the sequences using complete linkage which should leave us with an adequate number of clusters that contain precisely one sequence for every genome. From this step onward we deviate from the case study to implement the pMT-method.

Having extracted the clusters, we now calculate the evolutionary distances between the proteins within each group using the seqinr-package (v.3.4-5) in R, which returns a distance matrix. Said matrix is then added to the super distance matrix as mentioned earlier.

Having filled the super distance matrix, we will commence the swapping procedure. We take a similar approach to the authors of the original paper, however where they set a probability for each branch to swap, we swap a set of columns for each iteration. The procedure follows as such:

1. Randomly select two protein vectors.
2. Randomly select a genome.
3. For the two protein vectors, swap every column containing distances involving this genome.

This procedure is repeated 100000 times before correlations between all pairs of proteins are calculated. This should result in a null distribution similar to that achieved in the original method.

Lastly, the distance matrices for the bacteriocins and immunity proteins are formed, and the Pearson correlation is calculated. This correlation can then be compared to the distribution to find the empirical P-value according to equation 2.3.

$$P_{cor(bact,imm)} = \frac{N_{Above}}{N_{Total}} \tag{2.3}$$

A workflow of this procedure is available in figure 2.4.

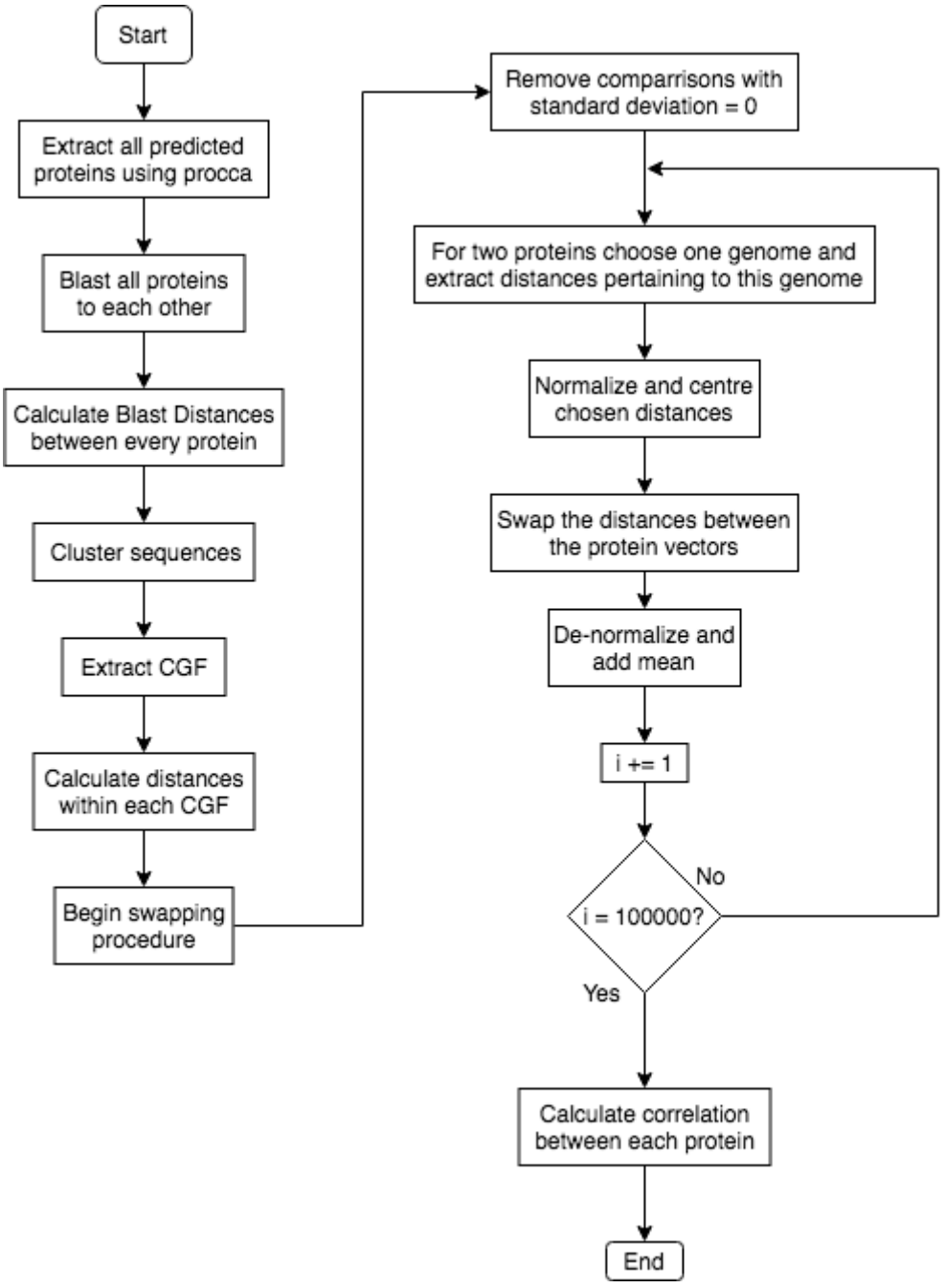


FIGURE 2.4: The altered implementation of the pMirrorTree method

2.4.3 Measuring correlation while cutting off amino-acids from either terminal

The proposed mode of action for the class IIa bacteriocin is that the conserved N-terminal sequence binds to an outer loop on the Man-PTS IIC as described in section 1.1.2. It follows that only the C-terminal half is allowed to interact with the immunity protein left in the cytosol (Figure 1.5. Based on this an iterative process was conceived where we remove amino acids from

the N-terminal of the bacteriocins one at a time while calculating the correlation in between each removal. The idea is that since only the C-terminal half is allowed to interact with the immunity, the correlation should be highest between the C-terminal half of the bacteriocin and the immunity protein. We then apply the same strategy to the C-terminal half using the MSA with the most significant correlation from the N-terminal run to look for another optimum, and as a form of redundancy.

These new MSAs are then tested against the swapped distance frames as described in section 2.4.2, and P-values are computed according to equation 2.3.

2.4.4 Principal Component Analysis of the distances

To further explore the two data sets, we perform two Principal Component Analyses (PCA). As each column of the super distance matrix represents a dimension for which every protein holds a value, our data sets will contain more than a hundred thousand comparisons each. This dimensionality makes it nigh impossible for our human mind to find patterns that describe the data in any meaningful way. To combat this "Curse of Dimensionality", the famous statistician K. Pearson devised the PCA Pearson (1901). This statistical method reduces dimensionality by calculating new vectors in the n-dimensional data that explain the linear relations between the values as best as possible, the Principle Components (PC). Once these components are calculated, the idea is that we are now able to more clearly visualize the linear patterns within the data sets in a 2- or 3-dimensional space.

Inspired by the article *Analysis of evolutionary patterns of genes in Campylobacter jejuni and C. coli* Snipen et al. (2012), we will normalize the super distance matrices by dividing the distances for every comparison by their mean so that we only consider their relative difference. After this, we carry out the PCAs, which should leave us with a decomposed data set we can visualize.

Chapter 3

Results

3.1 Seed sequences

HMMER3s *hmmbuild* function was used to construct the first pHMM from the initial sequences in Figure 2.2. The profile is represented as a logo in Figure A.1.

3.2 HMMsearch

Table 3.1 show the number of bacteriocin sequences and species found for each run of the hmmsearch. Each iterations results have been filtered by the E-value threshold given in Table 2.1

Iteration	Sequences	Species
1	928	46
2	918	53
3	982	65

TABLE 3.1: A tabular representation of how many bacteriocin-like sequences were found in all genomes across the three databases. Note that at least 600 sequences are from *Enterococcus Faecium* in every iteration.

The results from the third iteration were then handed over to Morten Kjos and Dzung B. Diep (Diep, 2018), who identified a number of interesting hits should they prove to be real bacteriocins. Knowing that a large portion of our hits originated from the Contig and Scaffold databases, which are less reliable than the Complete database, the parts containing these sequences were blasted against the NCBI nucleotide database to identify possible contaminations (table 3.2).

Species	Contamination	Contaminant species
Bavariicoccus seileri	+	Mixed
Chlamydia trachomatis	+	S. dysgalactiae
Clostridioides difficile	+	E. faecium
Clostridium beijerinckii	-	
Clostridium saccharobutylicum	-	
Escherichia coli	+	E. faecium
Klebsiella oxytoca	+	E. faecium
Listeria aquatica	-	
Marinilactibacillus psychrotolerans	+	C. Maltoromorticum
Oceanobacillus oncorhynchi	-	
Multiple hits in the *Streptococcus* genus	-	
Terribacillus saccharophilus	-	

TABLE 3.2: List of organisms that were unexpected according to personal communication with Dzung B. Diep (Diep, 2018). The table also lists if the sequences seem to be contaminations, and if so which species they are contaminated by.

The contaminant hits were then dismissed, along with the *Oceanobacillus* as it showed little resemblance to other Class IIa bacteriocins.

Genus	Percentage	Hits
Terribacillus	25.00	1
Listeria	5.26	2
Leuconostoc	16.67	3
Clostridium	1.54	9
Pediococcus	30.00	9
Carnobacterium	41.67	11
Streptococcus	22.67	61
Lactobacillus	10.28	80
Enterococcus	19.56	797

TABLE 3.3: The rate (%) of bacteriocin-carrying species compared to all species within the genera

Table 3.3 shows the percentage of species carrying bacteriocins within each genus.

The complete tables of hits and sequences are available via contact.

3.3 MirrorTree

Having formed the two datasets, *Enterococcus* and *Streptococcus*, the MSAs of each group were submitted to the MirrorTree server for analysis. The results can be seen in figure 3.1 and 3.2 respectively.

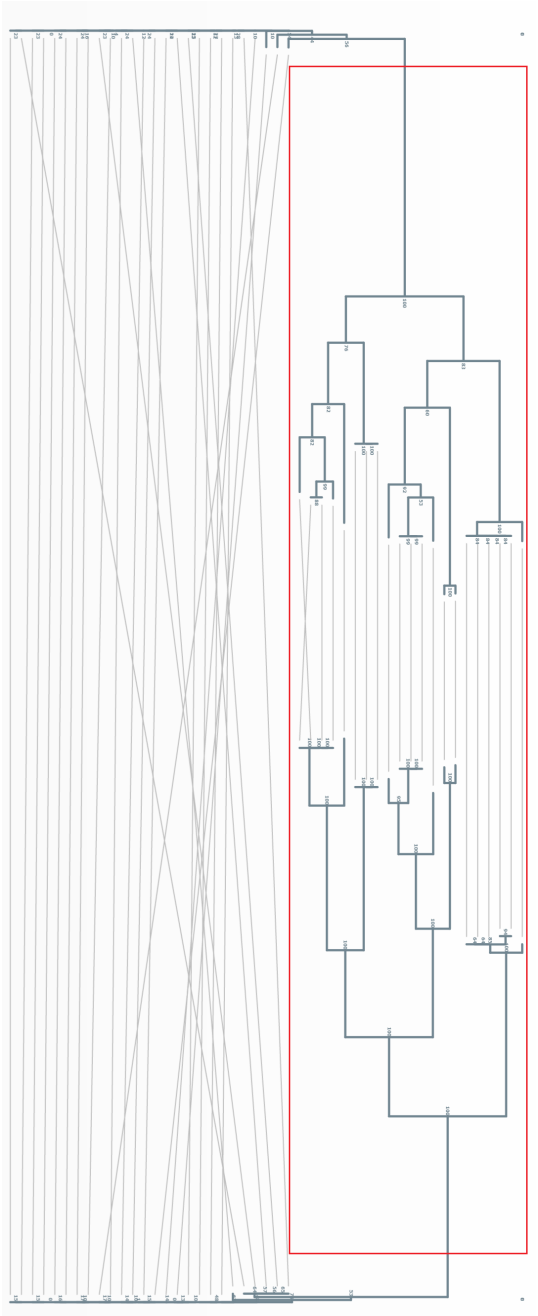


FIGURE 3.1: The resulting mirrortree from the MirrorTree server (Ochoa and Pazos, 2010) using the *Enterococcus* data set. *E. faecium* and *E. faecalis* (*red* box) are visually distinct from one another forming two clear groups in the tree, both in the bacteriocin tree (top) and the immunity protein tree (bottom).

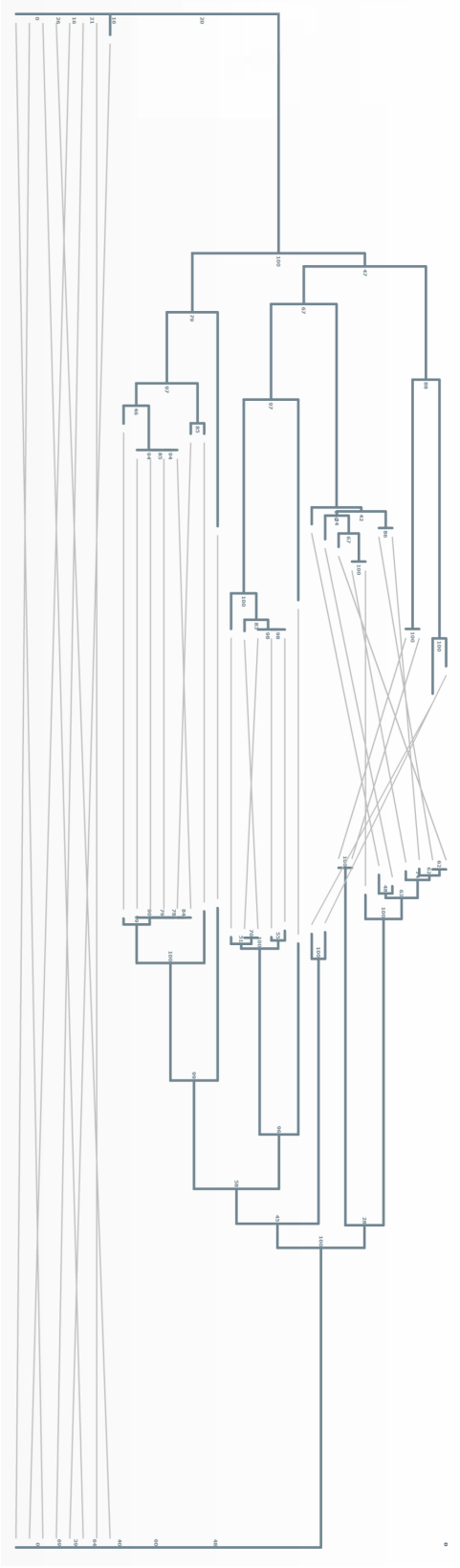


FIGURE 3.2: The resulting mirrortree from the MirrorTree server (Ochoa and Pazos, 2010) using the *Streptococcus* data set. The top tree is built from the bacteriocins, while the bottom tree is built from the Immunity proteins

3.4 pmirrortree

As with the regular mirrortree method, the altered pMT correlation was calculated as described in section 2.4.2 for both sets. In the distribution plots we have marked two correlations. One is for the full length of the bacteriocin MSAs when compared to the immunity protein, this is always marked as a green line. The second is the same correlation, but only parts of the bacteriocin MSA is used, marked as *red*. The latter approach is discussed in section 3.5.

3.4.1 *Enterococcus*

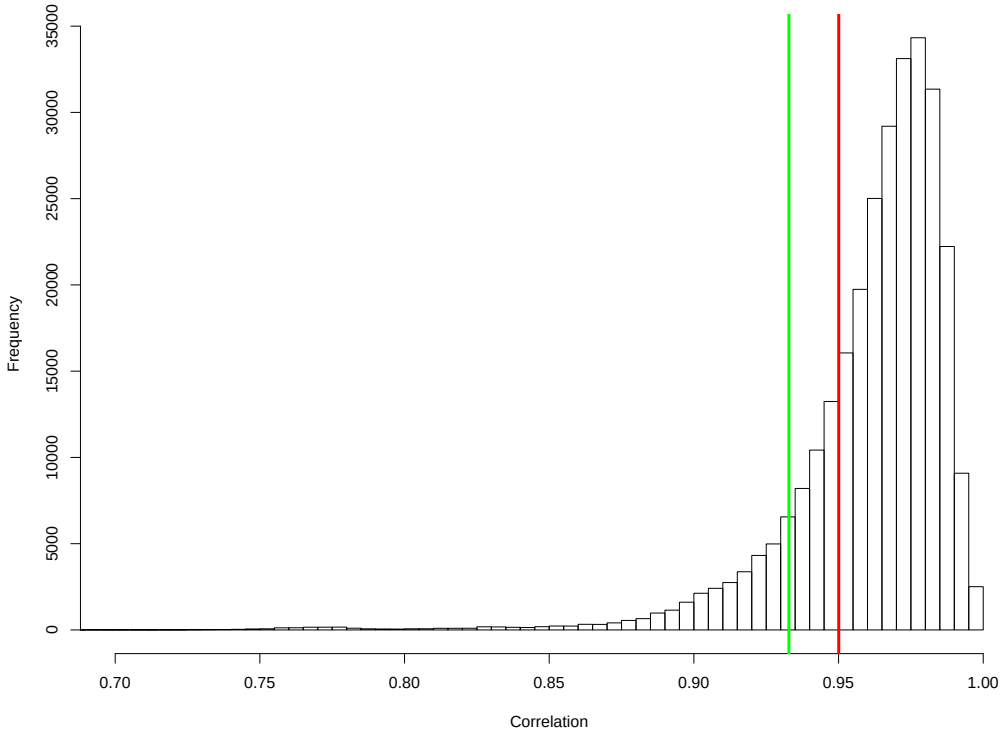


FIGURE 3.3: The distribution of correlations in the *Enterococcus* genomes analyzed. The two marked lines indicate the two correlations calculated for the bacteriocin and immunity protein. The first using the whole bacteriocin (*green*), the second using the cropped sequence found in section 3.5 (*red*)

The native correlations were plotted

First the native correlations of the distances were plotted to observe where the target correlations fell in this distribution. We see immediately that almost all proteins that were found share an extraordinarily high correlation. Had this been a complete set of shared proteins we would expected to find proteins with low correlations (Pazos and Valencia, 2001). And although the

correlation for the bacteriocin and immunity protein was markedly high, coming out at 0.9328, we see that it is grouped before the peak of correlations have been reached in Figure 3.3.

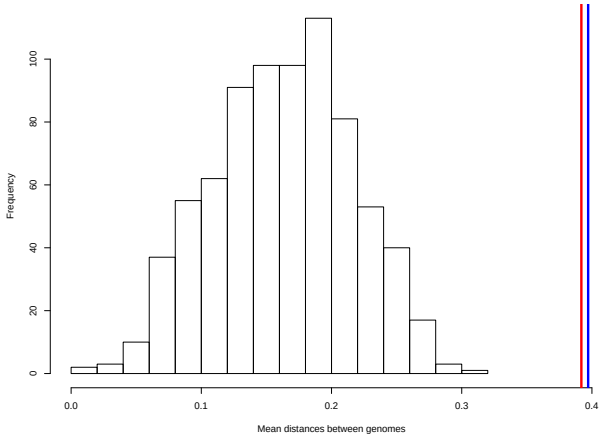


FIGURE 3.4: A histogram of the mean distance for each protein in the *Enterococcus* set. The mean distance for the bacteriocin (*red*) and immunity protein (*blue*) are marked as vertical lines.

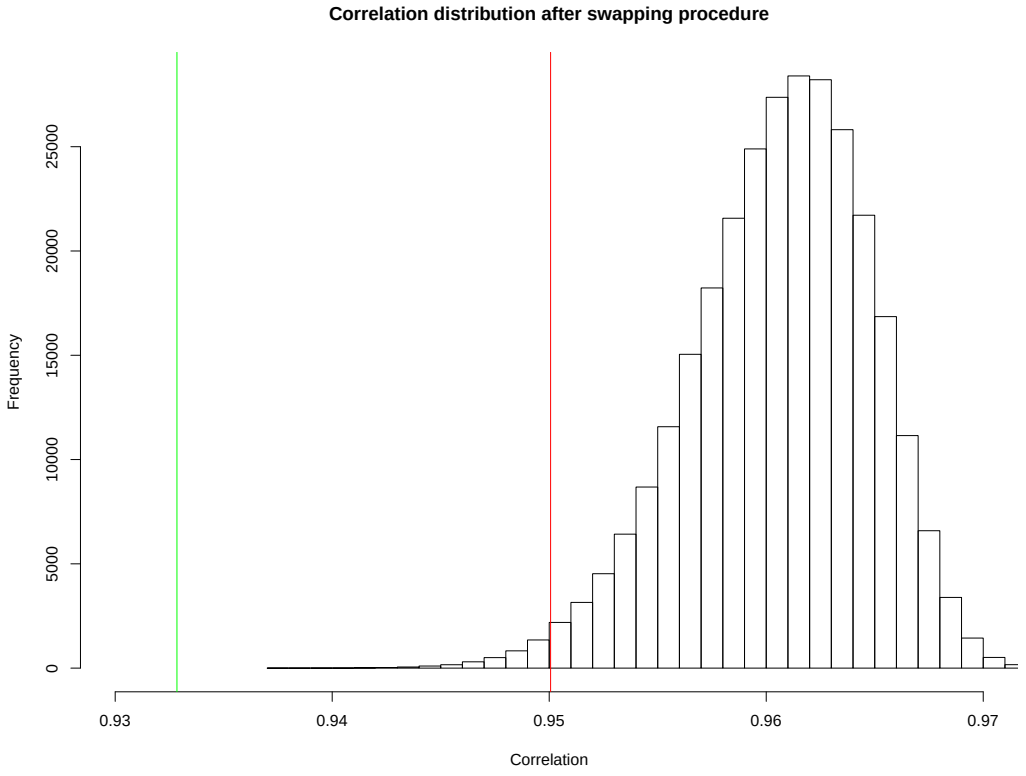


FIGURE 3.5: The correlation distribution of the *Enterococcus* proteins after the shuffling procedure has been carried out. The two target correlations have been marked as in figure 3.3.

The swapping procedure was caaried out, and all correlations were calculated. Figure 3.5 shows these correlations as a histogram, where the correlation between bacteriocin and immunity is

marked in *green*. The empirical P-value for this correlation is 0.9997

3.4.2 *Streptococcus*

Next up, the *Streptococcus* protein distance correlations were calculated and plotted as before.

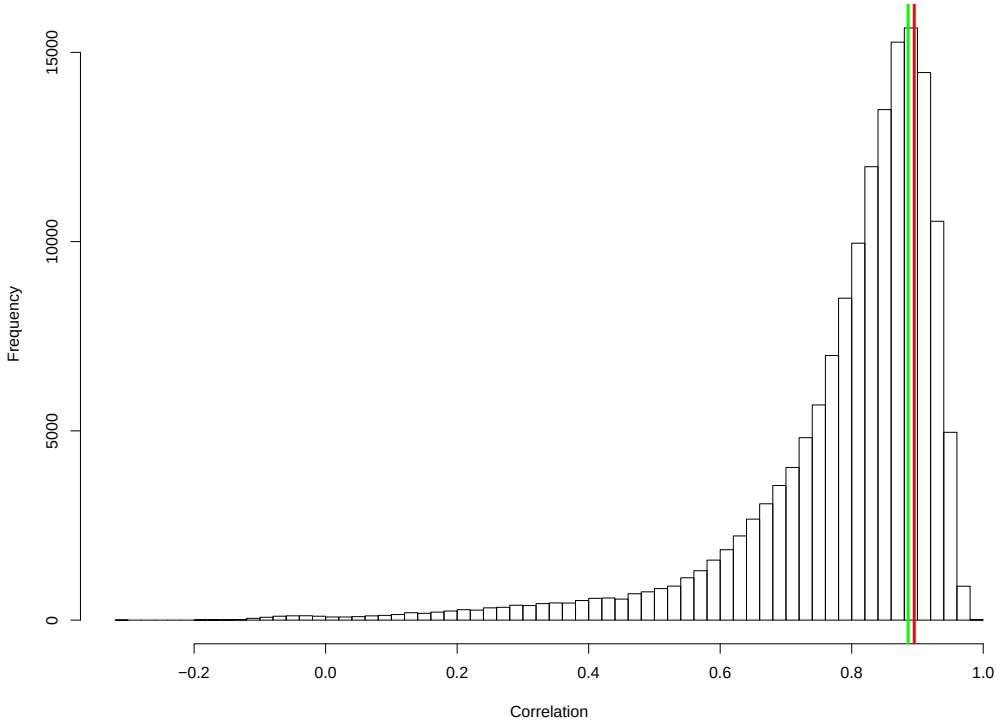


FIGURE 3.6: The distribution of correlation scores derived from the distance matrix before the swapping procedure had been applied to the data. The correlation between the bacteriocins and immunity proteins are marked at two locations, one for the score of the entire bacteriocin is used for the calculation (*green*), and one where only the optimal cut-off from the N-terminal was used (*red*)

The results are quite different this time compared to that of the *Enterococcus* set. From Figure 3.6 we see that the spectrum of correlations is more varied than in the previous data set, in spite of a lower number of proteins collected. The correlation between the bacteriocin and immunity protein was calculated to be 0.8856.

Performing the shuffling procedure on the distance matrix, we generate the null-distribution shown in Figure 3.8. Comparing the correlation between bacteriocin and immunity protein to the distribution, we find that the correlation is highly unlikely given that they are not coevolved. The empirical P-value comes out at 0.0007 which indicates a strong coevolutionary force between the proteins (Ochoa et al., 2015).

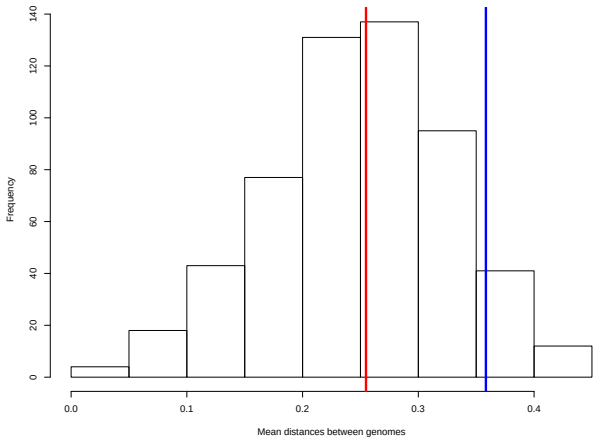


FIGURE 3.7: A histogram of the mean distance for each protein in the *Streptococcus* set. The mean distance for the bacteriocin (*red*) and immunity protein (*blue*) are marked as vertical lines.

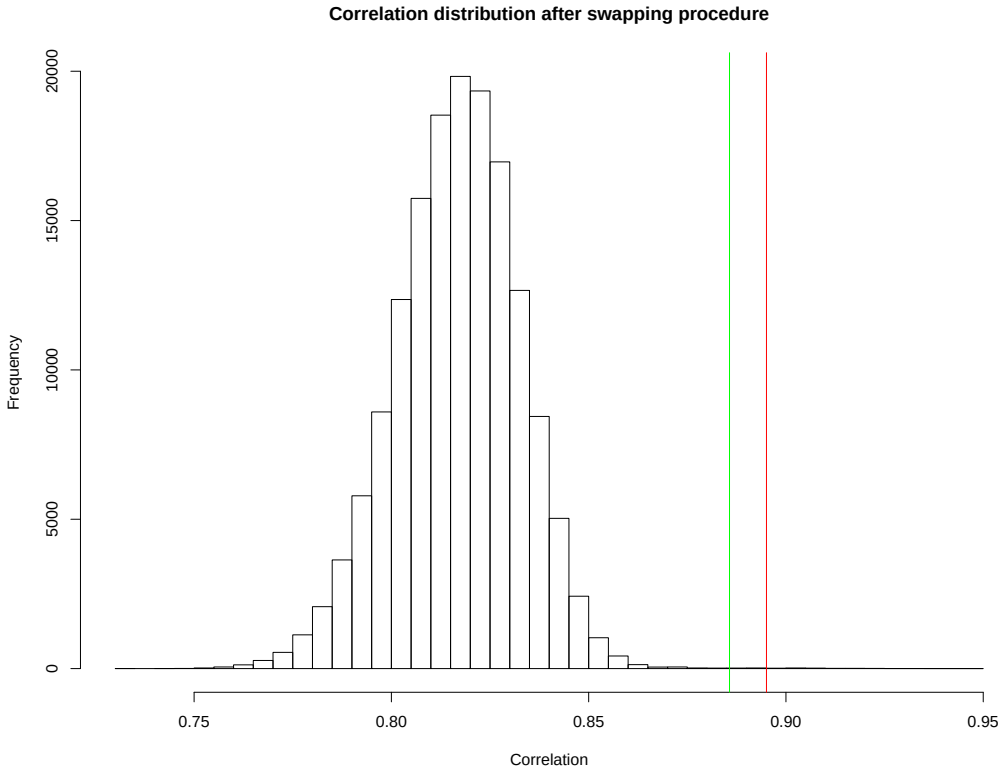


FIGURE 3.8: The distribution of correlation scores derived from the shuffled distance frame. The correlation between the bacteriocins and immunity proteins are marked at two locations, one for the score of the entire bacteriocin is used for the calculation (*green*), and one where only the optimal cut-off from the N-terminal was used (*red*)

3.5 Exploratory analysis to identify correlating regions

Below we present the findings form the exploratory analysis where we cut off amino acids from either terminal of the bacteriocins as described in the Methods section. Interestingly the two

bacteriocin-sets react differently to the method.

3.5.1 *Enterococcus*

Before presenting the findings in this section we have to mention that the empirical P-value found using the pMT is 0.9288. This means that the correlation is not significant when compared to the correlations of what should be un-correlated proteins. This is discussed further in section 4.3. Never the less, the findings are still of interest.

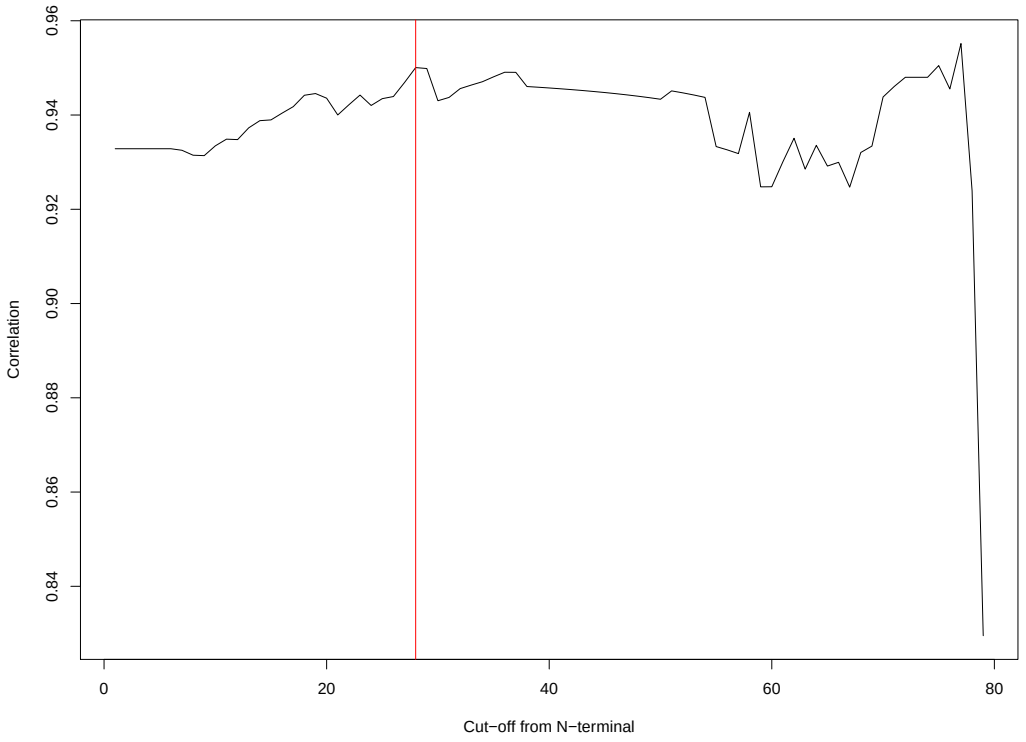


FIGURE 3.9: A plot of the resulting correlation scores for the bacteriocins in the *Enterococcus* set using the cropping method from the N-terminal. There is an optimal cut off point at position 28, which corresponds to right after the GG-motif in the MSA.

Cropping the sequences from the N-terminal we see that the highest correlation is found in position 28, which translates to the start of the bacteriocin after the leader peptide cut-off point by the double Gly-motif. This could indicate that, at least for the *Eneterococcus* genus, that the leader peptide is not particularly co-evolved to the immunity protein. This it could also be caused by the narrowness of the data set, as the sequences could simply be to conserved, so that any mutation will affect the correlation negatively.

From the previous results we found the optimal cut-off point from the N-terminal, and continuing on from there we crop the now shortened MSA from the C-terminal. Figure 3.10 clearly shows

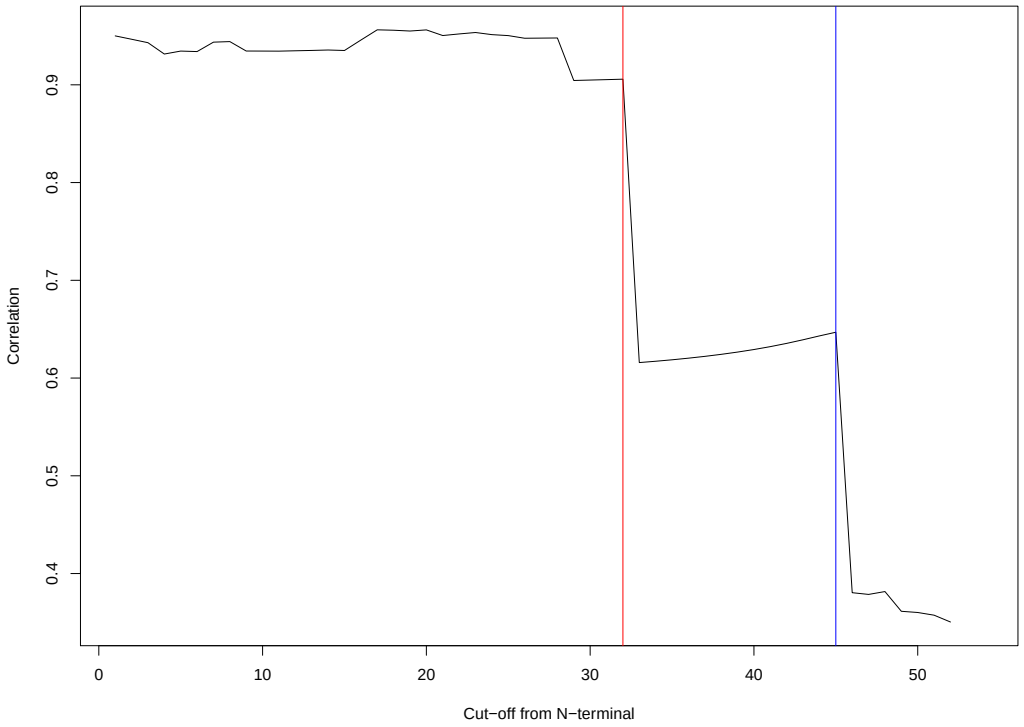


FIGURE 3.10: Results from the window-method cutting of from the C-terminal. The two major dips in correlation correspond to the first cystein (*red*) downstream from YGNG-motif, and the second corresponds to the tyrosine (*blue*) of the YGNG-motif.

two points that are crucial for correlation in the N-terminal half of the protein, namely the tyrosine of the YGNG-motif, shown in blue, and the first cystein downstream to said motif, marked in *red*. interestingly this sudden dip in correlation is not seen when croping them from the N-terminal half, neither do we see any notable increase in correlation by cropping from the C-terminal. Both of these findings indicate that the C-terminal half contains the most coevolved regions, which fits the going theory that a region of the C-terminal that interact with the immunity protein (Johnsen et al., 2004).

As stated, this new correlation, where the N-terminal is cropped at position 28, to the correlation distribution in section 3.4.1, results in a higher empirical P-value of 0.9288. This is marked as *red* line in figure 3.5.

3.5.2 *Streptococcus*

The same cropping procedure was performed on the *Streptococcus* bacteriocin MSA. This cropping method shows a rise in correlation after removing the 27 first positions from the N-terminal. Cropping to this position leaves the consensus leader-peptide region intact, marked *red* in Figure

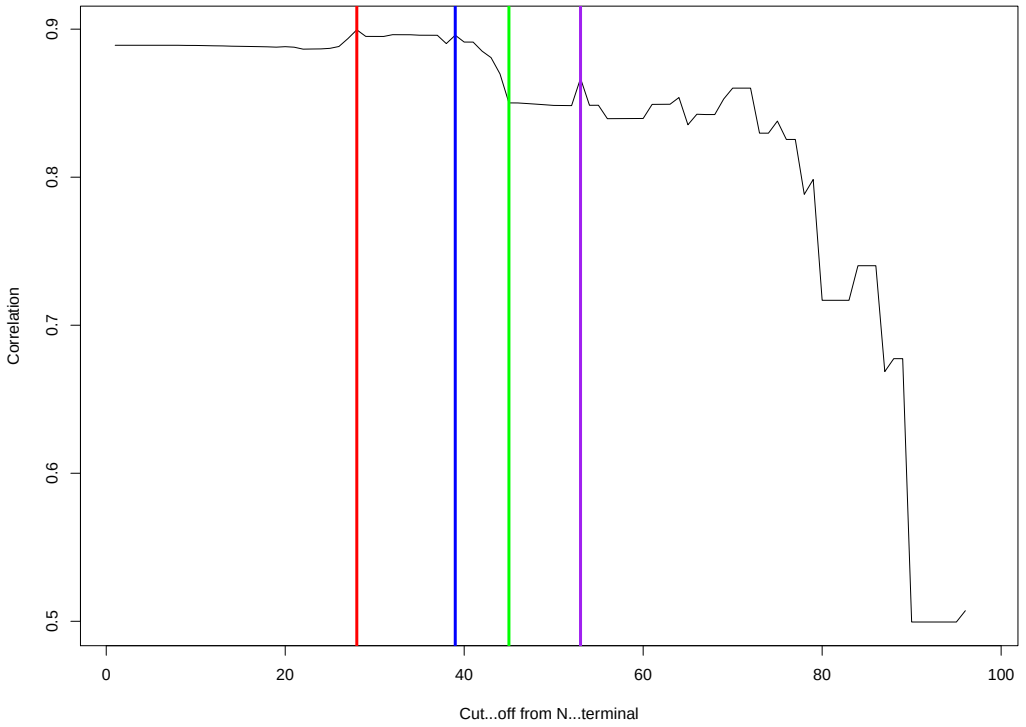


FIGURE 3.11: The correlations of the window-method as it is applied to the N-terminal of the bacteriocins found in the *Streptococcus* data-sets. The highest correlation is found at position 28 (*red*), which is the start of the common leader peptide. Another peak is found when calculating correlation from position 39 in the MSA, which corresponds to a variable amino acid just upstream from the Gly-Gly processing cite (*blue*). After this point the correlation drops until position 45 (*green*), corresposinding to the first glysine in the Gly-Gly processing cite. Lastly a new peak is reached at position 53 (*purple*), corresponding to a position just ahead of the pediocin -box motif.

3.11 and 3.12. The correlation drops once we remove the last amino acids before the GG motif, this span is marked from position 39 to position 45 (*blue* to *green*). Lastly we see a spike in correlation just upstream from the pediocin-box motif, at position 53 marked in purple. using the correlation from the optimal cut-off point, position 28, during the pMirrorTree test, we get a new empirical P-value of 0.0005, marked *red* in Figure 3.8.

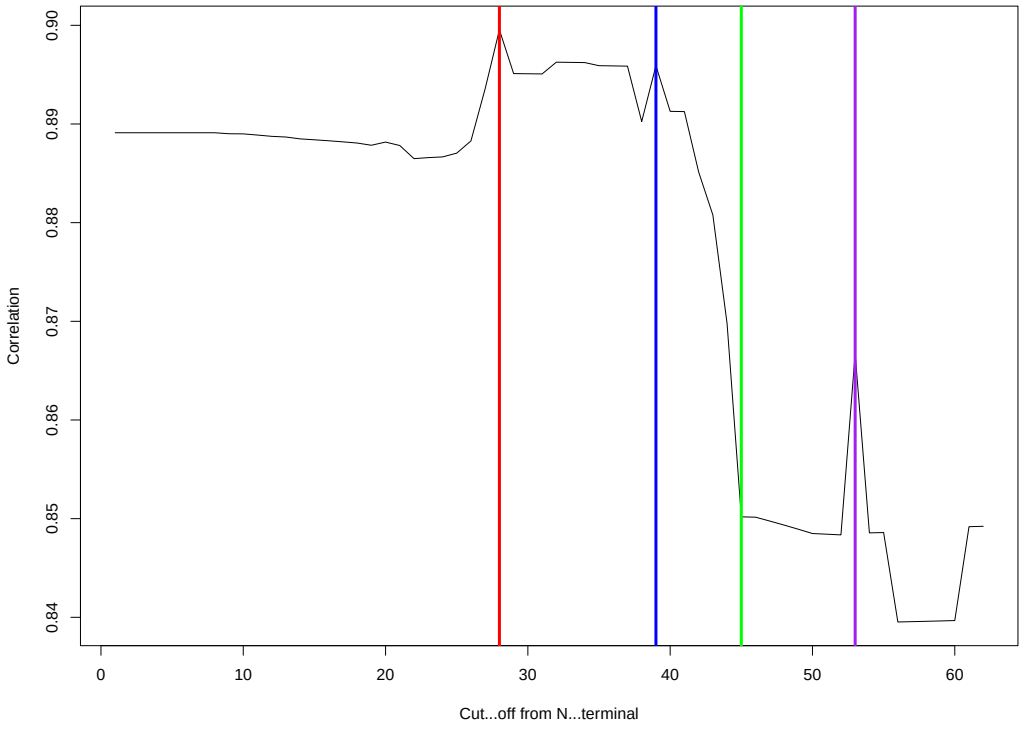


FIGURE 3.12: Subset of figure 3.11, showing the fluctuations between position 1 to 60. The three marked sections are the start of the leader peptide (*red*), a variable amino acid just upstream from the double Glysine-motif (*blue*), the first glysine of the GG-motif (*green*), and the beginning of the pediocin-box motif (*purple*)

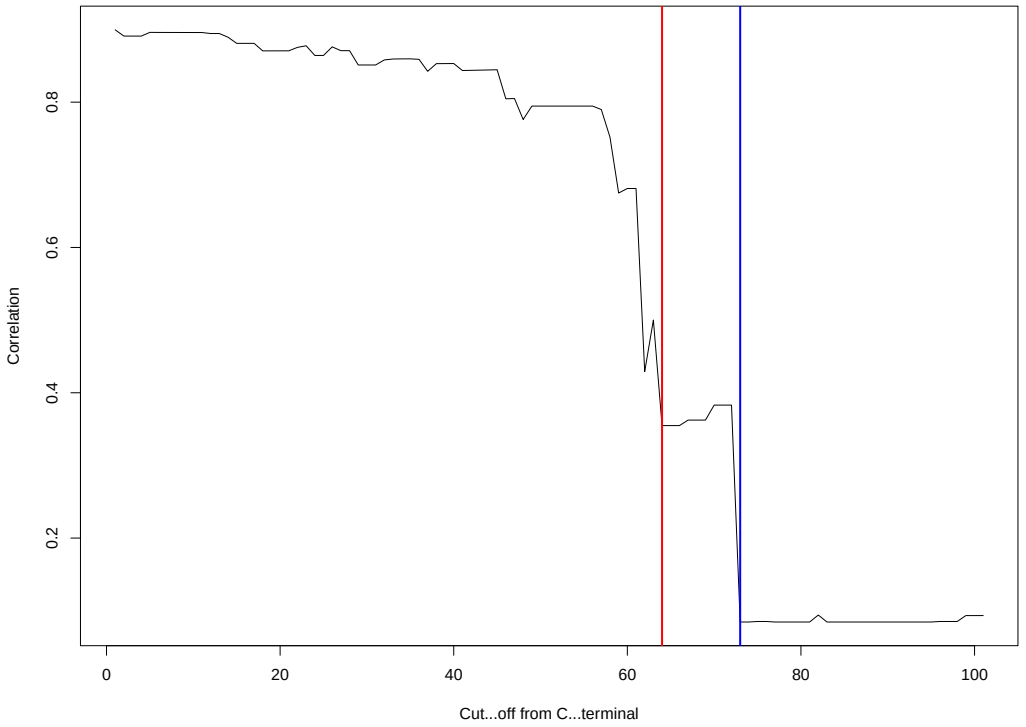


FIGURE 3.13: The window method applied to the C-terminal of the bacteriocin MSA, using the optimal cut-off from the N-terminal (position 28) found in figure 3.12. Two points are marked where a major drop in correlation has occurred, at position 64 (*red*) and 73 (*blue*) from the C-terminal. These correspond to one of the Cysteine needed to form the disulfide bridge, and the pediocin box motif.

Figure 3.13 shows a steady decrease in correlation while cropping from the C-terminal. Two sharp declines are marked *red* and *blue*, corresponding to one of the two cysteines needed to form the disulfide bridge, and the pedicin-box motif respectively.

3.6 PCA

A PCA was performed for both data-sets.

3.6.1 *Enterococcus*

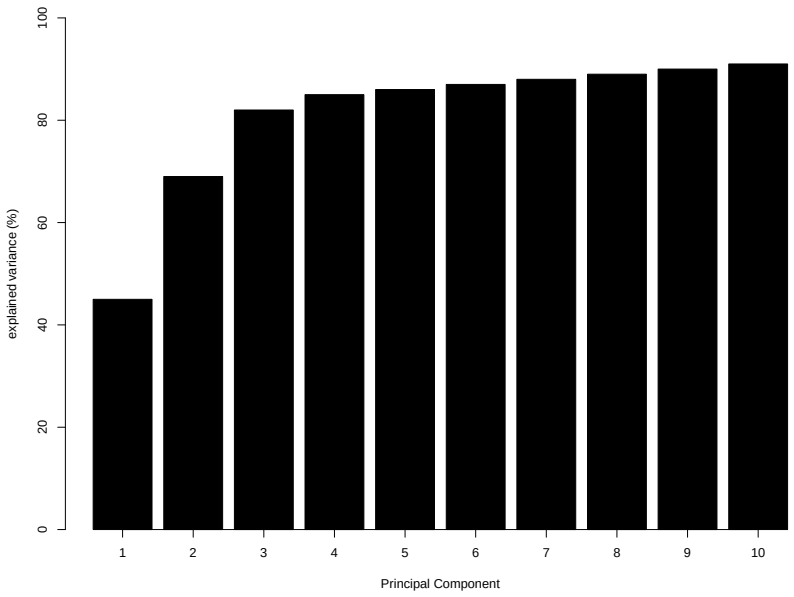


FIGURE 3.14: The cumulative explained variance in the principal component analysis of the evolutionary distance matrix for the *Enterococcus* data-set. Only the first 10 of the 766 PCs are shown. Using three PCs to plot the variance, over 80% of the variance is explained.

Figure 3.14 shows cumulative sum of explained variance over the first 10 components. The first direction accounts for over 40% of the variance in normalized evolutionary distances, and including the three first components we capture 80% of the variance. The succeeding components add gradually less variance to the explained sum, and we therefore assume this variation to be unimportant, and move forward using only the three first components.

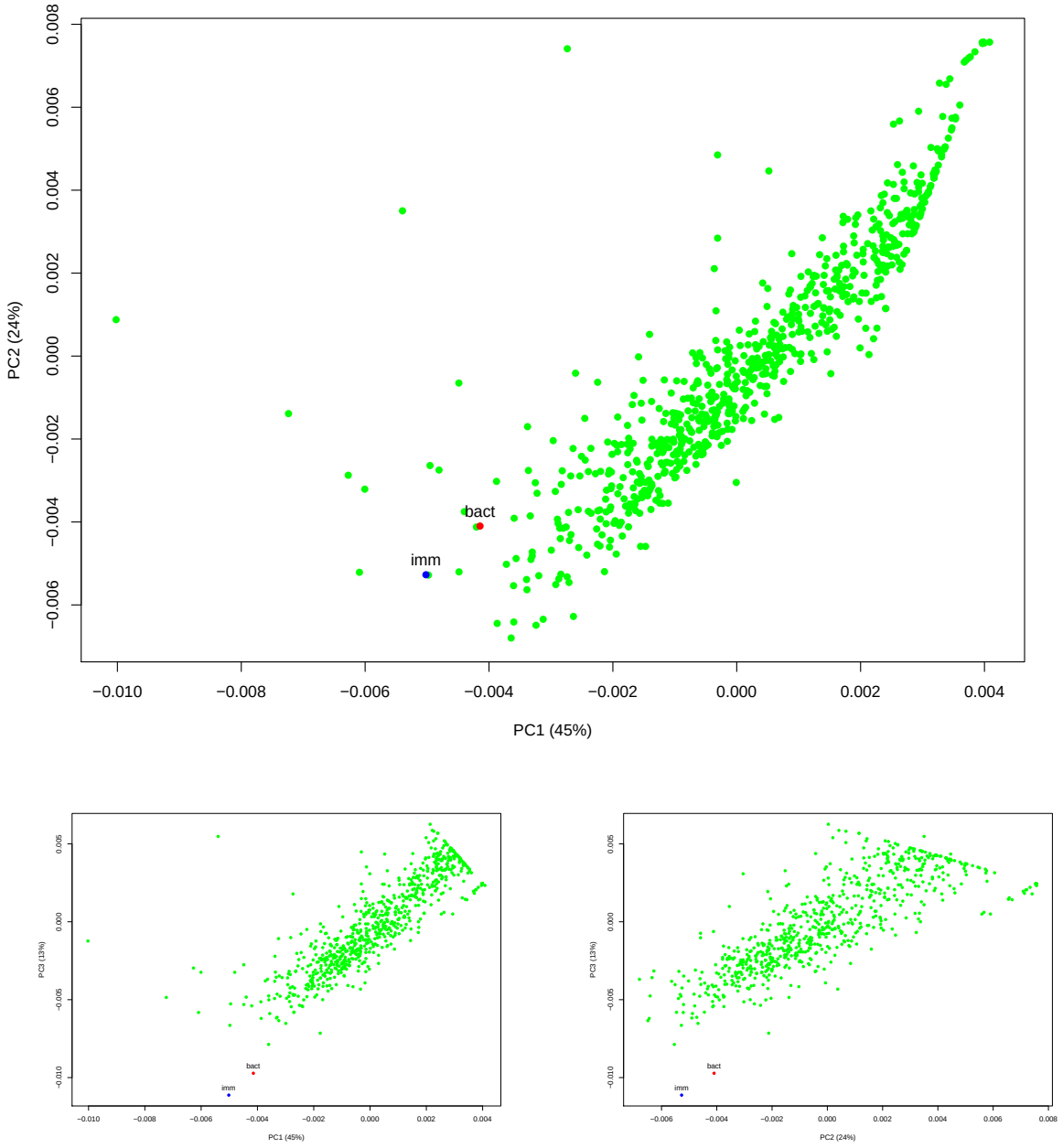


FIGURE 3.15: Plot of the top three principal components. The bacteriocin (*blue*) and immunity protein (*green*) are visually distinct from the CGFs (*red*), however always in close proximity to each other.

Figure 3.15 shows each protein family in the space spanned by the first three principal components. Each green point corresponds to the Central genes found through clustering, while the *red* and *blue* points corresponds to the bacteriocin and immunity proteins accordingly.

3.6.2 *Streptococcus*

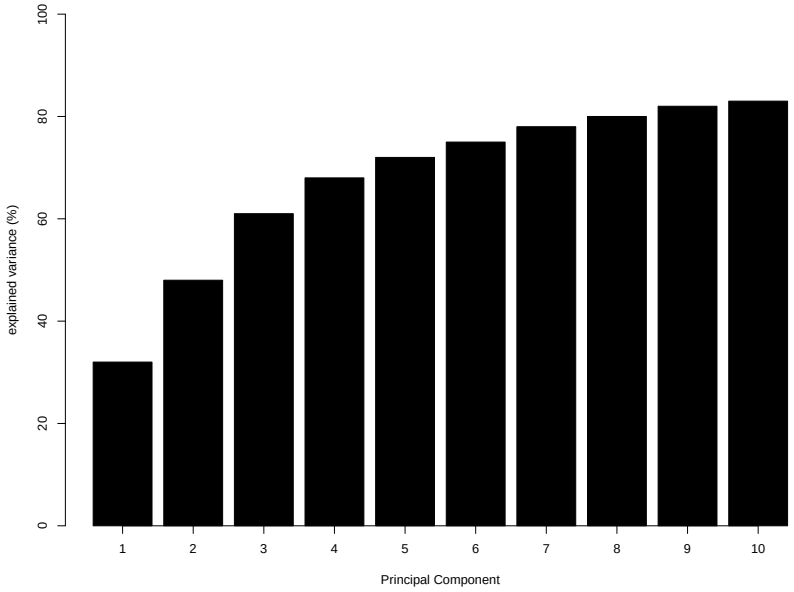


FIGURE 3.16: The cumulative explained variance in the principal component analysis of the evolutionary distance matrix for the *Streptococcus* data-set. Only the first 10 of the 528 PCs are shown. By using the first three Pricipal Components to plot the variance in figur 3.17, over 60% of the variance is explained.

Figure 3.16 shows the cumulative sum of explained variance over the first 10 components in the *Strepotococcus* set. The first component accounts for 30% of the variance in normalized evolutionary distances, and including the three first components we capture 60% of the variance. The succeeding components add gradually less variance to the explained sum, and are assumed to explain unimportant variance this variation to be unimportant, and move forward using only the three first components.

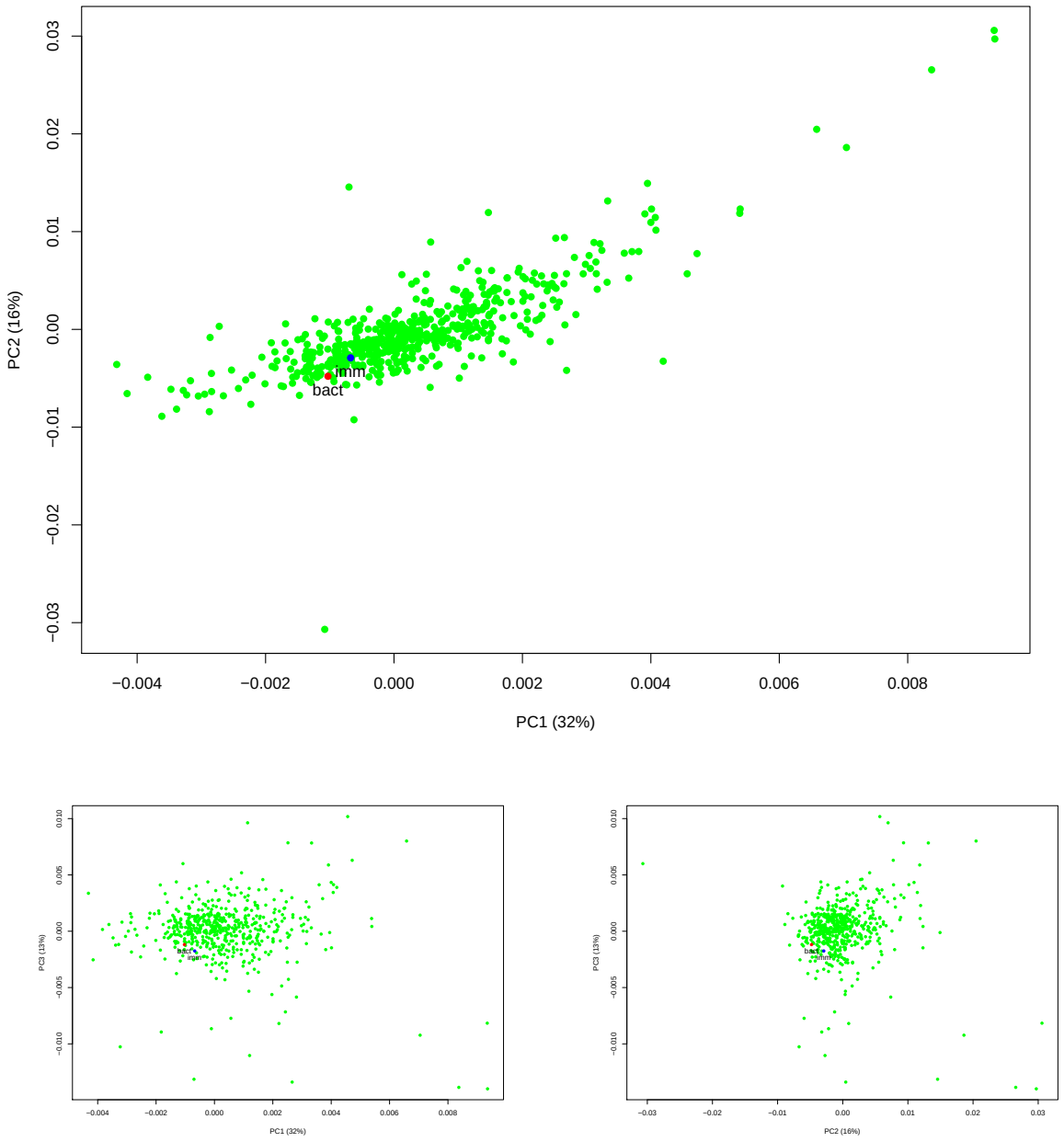


FIGURE 3.17: Plot of the top three principal components. The bacteriocin (*red*) and immunity protein (*blue*) are in close proximity to each other in every PC, while remaining close to the center of the CGFs (*green*).

A PCA was performed on the normalized distances The bacteriocin and immunity proteins are again shown to be situated close in the PCA-plot in Figure 3.17. The bacteriocinc (*red*) and immunity protein (*blue*) are found to be situated close together.

Chapter 4

Discussion

4.1 HMMER search of all genomes

We carried out the iterative HMMsearch as described in the Methods, and for each run, we managed to catch an increasing number of sequences, as intended. We see a minor decrease in detected sequences after the second iteration, while the number of individual species is increased. This decrease in sequences can most likely be attributed to the sharp increase in the E-value threshold chosen to filter the sequences. On that note, the E-values used to filter the results (Table 2.1) are chosen just above regions were the E-values sharply drop, or where non-bacteriocin-like sequences are found.

The decision to use a subset of the bacteriocins listed in the review article by Nissen-Meyer et al. (2009) was made because these were the only sequences that were publically available in the NCBI database in September of 2017.

4.1.1 Comparing hits to the BAGEL4 database

After reviewing the results of the bacteriocin mining procedure, we can see that we have found several new class IIa bacteriocin-carrying species. Comparing our sequences to other bacteriocin databases such as BactiBase (Hammami et al., 2016) and BAGEL4 (de Jong and van Heel, 2017) (Table A.1), we observe that several of our sequences are not found in either databases. Before we can definitively state that these are in fact new class IIa bacteriocins, these sequences need to be tested in a lab of course, but the data gathered should help with finding new candidate species.

4.1.2 HMMER over BLAST+

The decision to create a profile HMM of the sequences was made at the start of the analysis, as we feared that blasting the sequences would not be specific enough. Early tests using the blast engine at blast.ncbi.nlm.nih.gov seemed to confirm this, complemented by the fact that other bacteriocin mining tools also incorporates Hmmer in a similar manner (Morton et al., 2015).

4.1.3 The hmmsearch pipeline

A choice was made only to use the mature bacteriocins to build the pHMMs for the searches after each iteration. This is done as the original seed sequences, shown in Figure A.1, were mature sequences. This might not have been the right choice, as this omits the conserved GG-processing cite out of the model, which means there is one less conserved structure for the profile to latch on to. However, as our pipeline managed to locate several potentially new bacteriocins, this does not seem to have hampered us.

4.1.4 The missing YGNGV-motiv

After just the first run of the hmmsearch pipeline, we found a glaring issue. None of the bacteriocins that were located contained a YGNGV-motif. The YGNGL-motif was found as expected, but unexpectedly a new motif had appeared, YGNGT. As a never before seen motif this was suspicious, but after combing the code for errors, finding none, these sequences were included in the iterations to follow. In the last iteration of the search an attempt to rediscover the missing YGNGV motif was devised. By reintroducing known sequences containing the motif to the profile, the chances of finding such sequences should increase. This too failed. It was only after we finished the searching process that we located the error. At the time we used an unreleased version of the micropan package which contained a bug in the translate function, which translated codons coding for Valine [V] to Threonine [T]. After updating the package to version 1.2 all sequences with the YGNGT motif were correctly translated as YGNGV.

This bug did most likely very little to the specificity of the method, as the sequences that were aligned to the profile also had the translational error involved. What might have caused a problem, however, was the inclusion of YGNGV-motif carrying sequences into the third iteration, as this would lower the relative probability of finding either Threonine (as it was expected with the bug) or Leucine in that position.

4.2 Mirrortree server

After getting the results of the hmmsearch, we decided to run two different mirrortree approaches. First, we would run the sequences through the MirrorTree Server (Ochoa, 2010), then we would try and implement the pMT-method into R. To do this two sets of data were constructed, one with a narrow spectrum of species, and one with a broad spectrum. We tried to build a third data set of the *Lactobacillus* genus, as we had many hits within this genus as well. However, due finding a low number of shared genes during clustering, this data set was abandoned.

When applying the standard mirrortree method to the bacteriocins and immunity proteins of both data sets, the comparisons (Figure 3.1 & 3.2) show a very high correlation between the trees. The *Enteroccus* comparison had a correlation of 0.937, and the broader *Streptococcus* set had a correlation of 0.889. In line with the findings of Pazos and Valencia (2008), this high correlation is indicative of co-evolution.

As stated, the *Enterococcus* set retained a higher correlation than the *Streptococcus* set. This difference was expected as the *Enterococcus* set was comprised of only two species, compared to the *Streptococcus* set that covered 16 different species. The narrow species pool of the *Enterococcus* set limits the mutation-rate of both proteins, as we expect protein variance to be low when they are compared within a species, in contrast to when we compare them between species.

Moreover, we also see that the branch lengths between the *E. faecium* sequences are less varied than between the *E. faecalis* sequences, both in the bacteriocins and immunity proteins. This indicates that the bacteriocin-producing strains of *E. faecalis* are more varied within the than the bacteriocin-producing *E. faecium* strains.

Seeing as the given p-values for both trees are below 0.000001, the correlation is seen as significant. This in combination with the high correlations indicates, according to Pazos and Valencia (2001), that these proteins are correlated. However, as noted in 2.4.2, these p-values are not robust, and often disregarded entirely (Ochoa et al., 2015). The score is especially unreliable when the pool of compared organisms is small. We are therefore not able to say that there is a **significant** co-evolution between the bacteriocins and immunity protein based on this method.

4.3 pMT

To remedy this, we carried out the pMT-method as described in Section 2.4.2. First, the native correlations of the distances were plotted to observe where the target correlations fell in this distribution.

Looking at the results from the *Streptococcus* set, we see that the correlations between the clustered proteins are leaning heavily on the correlated side. In spite of this the correlation measured between the bacteriocin and the immunity proteins outclass most correlations, even before the swapping procedure. Comparing our target correlation to the "null-distribution" in figure 3.8, we clearly see that the bacteriocin and immunity protein are highly coevolved, according to the interpretation described in *Detection of significant proteincoevolution* (Ochoa et al., 2015), with a P-value of 0.0007.

Figure 3.6 and 3.7 show that the shared proteins are more diverse, as is to be expected when we look at 16 species compared to 2 in the previous run. However, we are only looking at the proteins that the same clustering algorithm as before managed to find. Therefore we are still probably looking at a distribution favouring coevolving proteins. Interestingly we only see a slight drop in the correlation of the bacteriocin and immunity protein. This is likely one of the most reliable indications that these proteins are coevolved, as they are similarly correlated in two different genera. And when we run the swapping procedure to produce the "null"-distribution, we find that the correlation is highly significant, in spite of the shortcomings in the method. This would mean that if we were to apply the pMT method to the data set correctly, we would only increase the significance of our findings.

Taking a look at the *Enterococcus* set, however, and we see that almost all proteins that are found share an extraordinarily high correlation. Had this been a complete set of all shared proteins we would expect to find proteins with low correlations (Pazos and Valencia, 2001) as well, as it is highly unlikely that all proteins in a genome would coevolve. And although the correlation for the bacteriocin and immunity protein was markedly high, coming out at 0.9328, we see that it is grouped before the peak in Figure 3.3. A possible explanation lies in Figure 3.4. Here we see that the average distances for each protein-vector are fairly small, at least when compared with the distances in the bacteriocin- and immunity protein-vectors. It would seem that the clustering process was only able to extract the highly conserved and/or coevolved genes from the genomes. This poses a problem for our implementation of the pMT-method, as no matter how many times we shuffle around the distances, they would never truly represent a null distribution for our null-hypothesis. And this is, presumably, what has taken place during

our implementation of the pMT-method. Thus, a correlation as 0.9328 is seen as insignificant when compared to the other protein sequences.

This can probably be explained by the way we implemented the pMT method. If we look at the main difference between our implementation of pMT to that of the original paper, we only compare "trees" where all branches are shared. In the original paper, they matched all trees, regardless of how many branches they had in common. By not following the same principle we have created a process that results in a null-distribution where we assume that all the proteins gathered are independently evolved. A quick way to affirm whether our proteins represent all proteins found in the genome or not is to look at the correlation between the bacteriocin-vector and all other protein-vectors (Figure A.3). Here we see a sharp divergence from the norm, according to Figure 3.3, where correlations as low as this are hardly even present.

As the *Streptococcus* set did not show this same problem, it would seem that increasing the number of species in the *Enterococcus* set could remedy the results. By adding more species we would degrade the evolutionary pattern present in the conserved/coevolved sequences, which should in turn lead to a more varied distribution of correlations.

As the pMT method seems to struggle when the species compared are similar, we created an alternative approach. In this method we simply calculate the correlation between the bacteriocin and all the proteins found by clustering, and then compare the target correlation to the resulting distribution. This method showed quite clearly that the two proteins were more correlated to each other than the rest, as seen in Figure A.3 and A.4 where the target correlation is marked *red*. Therein we see, much clearer than the pMT method, that the sequences are considerably more coevolved to each other than to any other protein, especially in the *Enterococcus* set, where the pMT method failed to detect this. And while this is a crude approach, it should be able to show, when dealing with similar distributions as in this thesis, whether the two proteins you are investigating are coevolved or just a part of the natural correlations found in the data set.

4.3.1 Cropping method

The cropping procedure was thought to be an exploratory analysis to see if we could identify highly correlating regions by cropping from either terminal.The hypothesis was that since the C-terminal of both proteins are involved in the recognition process, these regions would be more correlated than the rest of the proteins. In the *Enterococcus* set we found the highest correlation by cutting away the leader peptide, cropping to just after the Gly-Gly processing

site (Figure 3.10). In essence, this means that correlation was highest when we compared the mature bacteriocin to the immunity protein. Oddly, we saw the opposite in the *Streptococcus* set, where correlation peaked when the entire, common, leader-peptide was present (Figure 3.12). It dips down once we remove a variable region between position 39 and 45, only rising when we compare the mature bacteriocin from the start of the pediocin-box. Based on this it would seem like the leader-peptide plays a role in the coevolution of the proteins. However, a counter-argument forms when we look at the same procedure carried out from the C-terminal.

By cropping from the C-terminal (Figure 3.13) we see that there is virtually zero correlation between the leader-peptide (*blue*line and outward) and the immunity protein, indicating that the increase in correlation may merely be a result of noise.

Cropping from the C-terminal results in sudden drops in correlation when the conserved pediocin box and cysteines are removed. This happens for both data sets. Interestingly these same drops do not occur when we crop from the N-terminal. This implies that the C-terminal half retains enough evolutionary movement to stay correlated to the immunity protein, indicating that the C-terminal is, in fact, more correlated than the N-terminal.

4.4 Principle Component Analysis

The PCA analysis revealed two things about our data sets. One, the *Enterococcus* bacteriocin and immunity protein are correlated independently of the genus. Two, the *Streptococcus* data set is not suited for principal component analysis.

For the *Enterococcus* data set we see that the three first principle components describe more than 80% of the variance found in the initial set. This indicates that there are patterns available in the data that the PCA can latch on to, which we see clearly in Figure 3.15. In the plots we see that most proteins align to a linear pattern, moving positively on all three PCs. As these proteins are highly conserved, as shown in Figure 3.4, this could be thought of as a species specific evolutionary pattern. The higher something climbs on these components, the more clearly they become a *E. faecium* or *E. faecalis* proteins. However, the bacteriocin and immunity diverges quite heavily from the norm in all three components, indicating that they evolve independently to the species that produce them. This in combination with the fact that they are situated closesly indicates that they are part of the same independant evolutionary pattern, and therefore coevolved.

The *Streptococcus* set, on the other hand, runs into problems rather quickly. At first glance we see a linear pattern in Figure 3.17 when comparing PC1 and PC2. The problem becomes

apparent when we look at the explained variance, as well as when we look at the plots where PC3 is involved. The explained variance is very low in these first components, which indicates that there is no clear pattern for the PCA to attach itself to. And while the PCA seems to find a linear pattern when comparing PC1 to PC2, this comparison explains less than 50% of the variance contained in the data set. And when taking into account the circular shape of the data found when comparing PC3 to either of the two first, there is no pattern to be gleamed. Both of these findings tells us that the data set is unsuited for PCA. A reason for this could be that the addition of more species has messed up the clear evolutionary pattern seen in the *Enterococcus* set. This is suported by the distances found in Figure 3.7 which cover a wider spread than in the *Enterococcus* set.

The bacteriocin and immunity proteins are still situated closely, so they are at correlated, yet we are not capable of saying much more than that.

4.5 Future research

First of all, the iterative HMMER approach seems to have worked like a charm, so letting this improve over more iterations seems like a good use of time. The main trick seems to make sure that it doesn't become overfitted to one particular bacteriocin (For example, adding all *E. faecium* sequences after a run would likely overfit the profile immediately). It could also be interesting to see if different seed sequences would produce different hits.

We have also managed to annotate new potential bacteriocin sequences. There is particular one new hit that should be tested, the *Listeria aquatica* hit. *Listeria* is one of the genus that we currently try to combat using class IIa bacteriocins (Drider et al., 2006). Should this hit turn out to be real, this would mean that the *Listeria* genus already has an immunity protein available to infer immunity. It is not unthinkable that this immunity protein could undergo changes to evolve to an orphan immunity protein similar to the previously mentioned *OrfY* immunity protein. Along with this the bacteriocin could spread to other parts of the genus, giving the bacteria a way to fight back against the LAB bacteria.

Should one still wish to perform/improve the pMT method, one should this time expand the method to compare all trees to each other, not just the ones where all the branches match. Along with this the *Enterococcus* set should be expanded to incorporate more species within the genus, while at the same time limiting how many sequences are included for each species.

Another analysis that can, and should, be run is a comparison between the Man-PTS and the immunity protein as well as the bacteriocin. For even as we find such strong correlation

between the bacteriocins and immunity proteins, it could still be interesting to see whether a similar correlation exists when comparing them to either the entirety of the IIC subunit of the Man-PTS, or a subset of it. This was intended to be performed in this study, but due to time constraints it was not followed through.

Chapter 5

Conclusion

In this thesis we set out to investigate whether the class IIa bacteriocins were coevolved with their immunity proteins. To accomplish this we used two subsets of class IIa bacteriocin producing species; one set consisting of *E. faecium* and *E. faecalis*, and another consisting of every unique bacteriocin carrying strains in the *Streptococcus* genus.

The conventional mirrortree method, through the MirrorTree server, found a strong correlation between the two proteins in both sets, indicating that the proteins are coevolved. The results from the pMirrorTree showed *significant* coevolution between the bacteriocin and immunity protein of the *Streptococcus* set, while the *Enterococcus* set failed to show *significant* coevolution through this method, in spite of a higher correlation than the prior set.

This result is contrasted by the PCAs however, as the plots from both sets clearly show that the bacteriocin and immunity proteins are correlated. Interestingly we see that the evolution of the bacteriocins and immunity proteins from the *Enterococcus* deviate from the patterns found by the PCA. And while little to no evolutionary pattern was found in the *Streptococcus* set, we still see that they are correlated.

Considering these findings, we can clearly identify that the bacteriocin and immunity protein are coevolved, in spite of the *Enterococcus* set failing the test. As discussed the way we implemented an extremely strict version of the pMT method, especially when the compared strains are similar. And it would be extremely suspicious if the class IIa bacteriocin and immunity protein in one species act and evolve completely different to those of another species. This in combination with the PCA and appendix figures allow us to state, with high certainty, that the proteins are coevolved.

The results from the cropping procedure indicated that the C-terminal region of the bacteriocin holds more influence on the correlation than the N-terminal region, which lends credibility to the hypothesis that the C-terminal halves are the interacting regions between the two proteins.

Lastly, we have been able to annotate potentially new class IIa bacteriocins through our iterative HMMsearch method. Noteworthy additions are *Listeria aquatica*, *Clostridium beijerinckii* and *Clostridium accharobutylicum*, and a slew of new hits in the *Streptococcus genus*.

Appendix A

Species	**HMMsearch**	**BAGEL4**
Bacillus coagulans	-	+
Carnobacterium gallinarum	+	-
Carnobacterium inhibens	+	-
Carnobacterium maltaromaticum	+	-
Carnobacterium mobile	+	-
Carnobacterium viridans	+	-
Clostridium beijerinckii	+	-
Clostridium saccharobutylicum	+	-
Enterococcus canintestini	+	-
Enterococcus durans	+	-
Enterococcus hirae	+	-
Enterococcus pallens	+	-
Enterococcus rivorum	+	-
Enterococcus sp.	+	-
Lactobacillus acidipiscis	+	-
Lactobacillus agilis	+	-
Lactobacillus aquaticus	+	-
Lactobacillus casei	+	-
Lactobacillus crispatus	+	-
Lactobacillus delbrueckii	+	-
Lactobacillus fuchuensis	+	-
Lactobacillus futsaii	+	-
Lactobacillus hordei	+	-
Lactobacillus murinus	+	-
Lactobacillus rennini	+	-

Lactobacillus rhamnosus	+	-
Lactobacillus ruminis	+	-
Lactobacillus sp.	+	-
Lactobacillus taiwanensis	+	-
Listeria aquatica	+	-
Pediococcus ethanolidurans	+	-
Pediococcus pentosaceus	+	-
Streptococcus criceti	+	-
Streptococcus downei	+	-
Streptococcus dysgalactiae	+	-
Streptococcus equi	+	-
Streptococcus equinus	+	-
Streptococcus gallolyticus	+	-
Streptococcus henryi	+	-
Streptococcus infantarius	+	-
Streptococcus macacae	+	-
Streptococcus mitis	+	-
Streptococcus mutans	+	-
Streptococcus oralis	+	-
Streptococcus pantholopis	+	-
Streptococcus pasteurianus	+	-
Streptococcus pseudopneumoniae	+	-
Streptococcus sobrinus	+	-
Streptococcus sp.	+	-
Terribacillus saccharophilus	+	-
Carnobacterium divergens	-	+
Carnobacterium piscicola	-	+
Enterococcus faecalis	+	+
Enterococcus faecium	+	+
Enterococcus mundtii	+	+
Lactobacillus curvatus	+	+
Lactobacillus plantarum	+	+
Lactobacillus sakei	+	+
Lactococcus garvieae	-	+

Lactococcus lactis	-	+
Lactococcus sp.	-	+
Leuconostoc carnosum	-	+
Leuconostoc gelidum	+	+
Leuconostoc mesenteroides	+	+
Listeria innocua743	-	+
Pediococcus acidilactici	+	+
Streptococcus thermophilus	-	+
Streptococcus uberis	-	+

TABLE A.1: A comparison of the class IIa bacteriocin carrying species we found and the ones that are listed in the Bagel4 database (*http://bagel4.molgenrug.nl/* Accessed: 2018-05-12)

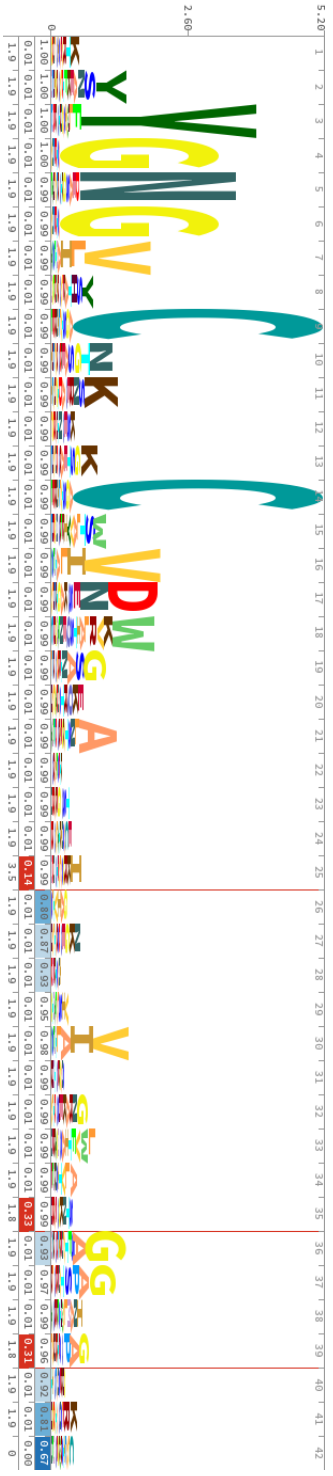


FIGURE A.1: Logo representation of original profile HMM. For each position the relative probability of having a set number. The logo was built using the Skylign tool at *skylign.org* Accessed: 03-04-2018 (J Wheeler et al., 2014).

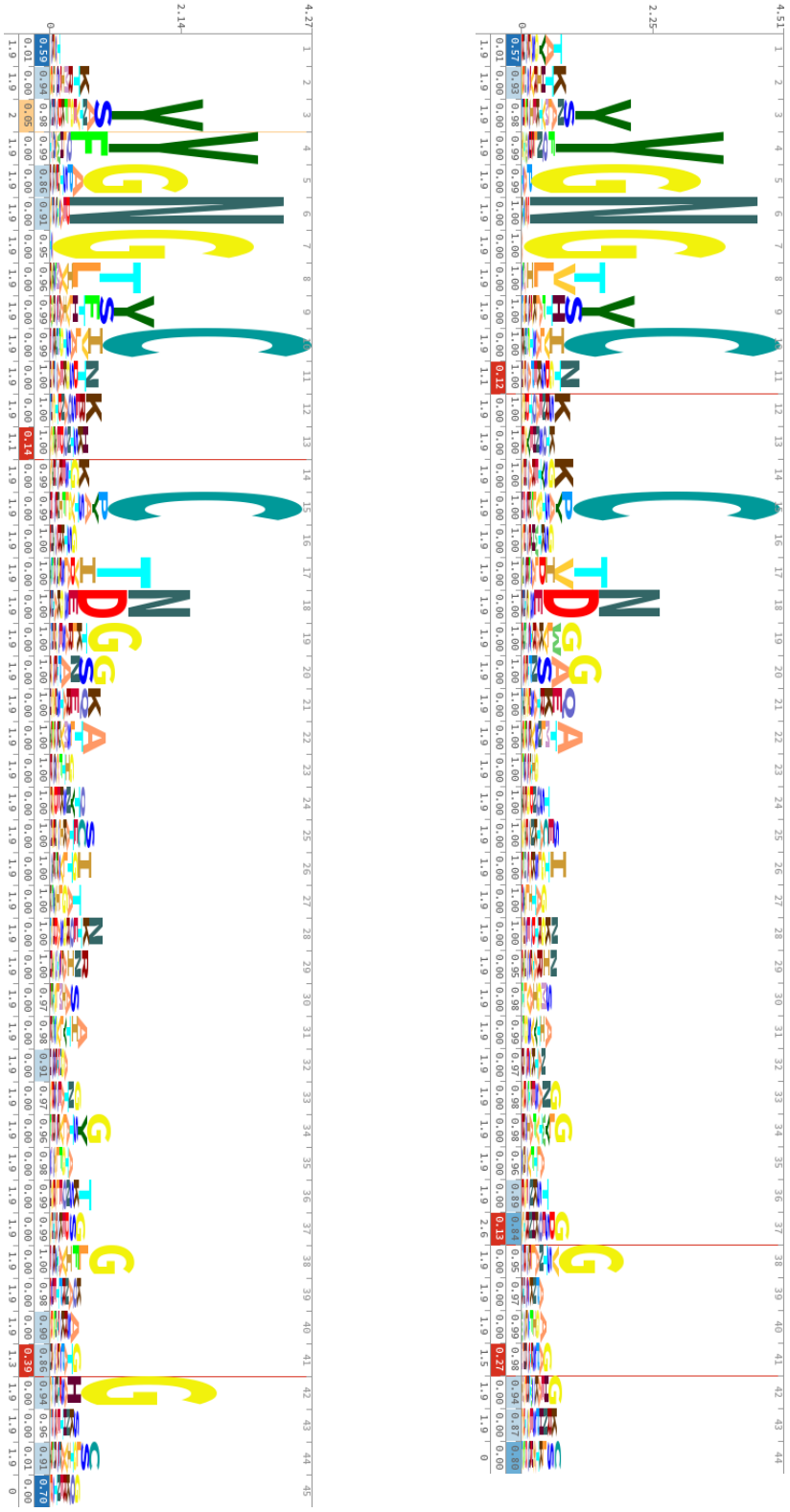


FIGURE A.2: Profile HMMs used to run the second (left) and third (right) iteration of the HMMsearch. The logos was built using the Skylign tool at *skylign.org* Accessed: 06-04-2018 (J Wheeler et al., 2014).

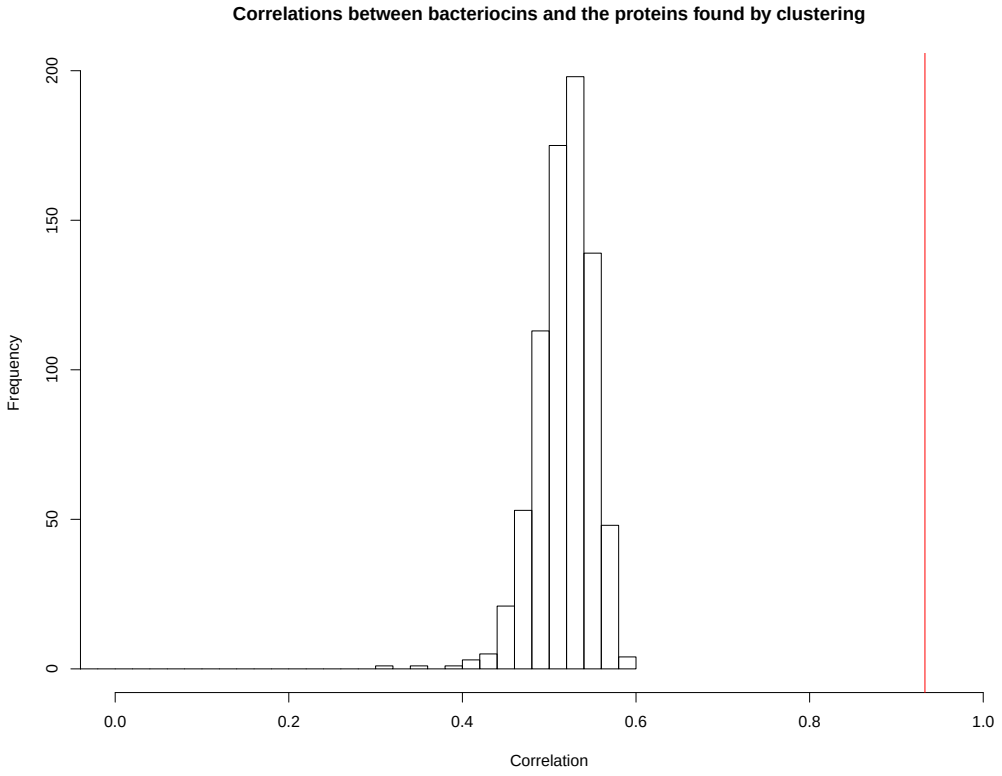


FIGURE A.3: A histogram showing the correlations found between the bacteriocin proteins and the proteins found by clustering in the *Enterococcus* data set.

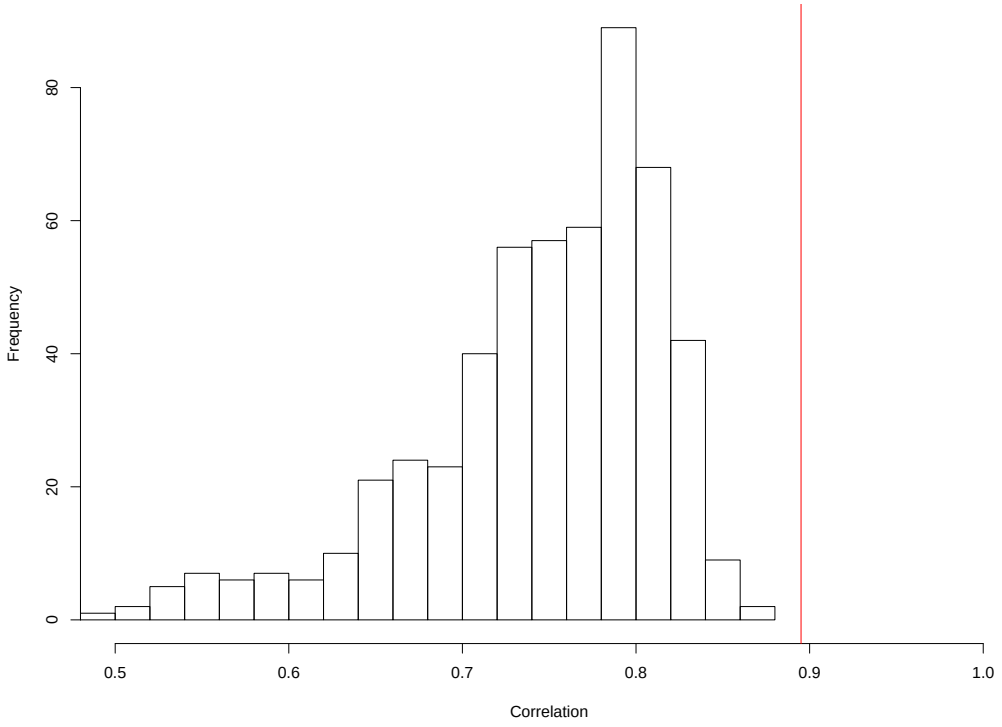


FIGURE A.4: Correlation distribution between bacteriocin and the clustered proteins in the *Streptococcus* data set

Bibliography

Abraham, E. P. and Chain, E. (1940). An Enzyme from Bacteria able to Destroy Penicillin. *Nature*, 146:837.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410.

Alvarez-Sieiro, P., Montalbán-López, M., Mu, D., and Kuipers, O. P. (2016). Bacteriocins of lactic acid bacteria: extending the family. *Applied Microbiology and Biotechnology*, 100:2939–2951.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2008). GenBank. *Nucleic Acids Research*, 36(Database issue):D25–D30.

Biswas, S. R., Ray, P., Johnson, M. C., and Ray, B. (1991). Influence of growth conditions on the production of a bacteriocin, pediocin AcH, by Pediococcus acidilactici H. *Applied and Environmental Microbiology*, 57(4):1265–1267.

Brurberg, M. B., Nes, I. F., and Eijsink, V. G. (1997). Pheromone-induced production of antimicrobial peptides in Lactobacillus. *Molecular microbiology*, 26:347–360.

Chopra, I. and Roberts, M. (2001). Tetracycline Antibiotics: Mode of Action, Applications, Molecular Biology, and Epidemiology of Bacterial Resistance. *Microbiology and Molecular Biology Reviews*, 65(2):232–260.

Cochrane, G., Karsch-Mizrachi, I., Takagi, T., and SequenceDatabaseCollaboration, I. N. (2016). The International Nucleotide Sequence Database Collaboration. *Nucleic Acids Research*, 44(Database issue):D48–D50.

Conly, J. M. and Johnston, B. L. (2005). Where are all the new antibiotics? The new antibiotic paradox. *The Canadian Journal of Infectious Diseases & Medical Microbiology*, 16(3):159–160.

Coordinators, N. R. (2016). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44(Database issue):D7–D19.

Cotter, P. D., Hill, C., and Ross, R. P. (2005). Bacteriocins: developing innate immunity for food. *Nature Reviews Microbiology*, 3:777.

Cotter, P. D., Ross, R. P., and Hill, C. (2012). Bacteriocins a viable alternative to antibiotics? *Nature Reviews Microbiology*, 11:95.

de Jong, A. and van Heel, A. (2017). BAGEL4.

de Juan, D., Pazos, F., and Valencia, A. (2013). Emerging methods in protein co-evolution. *Nature Reviews Genetics*, 14:249.

Diep, D. B. (2018). Personal Communication.

Diep, D. B., Skaugen, M., Salehian, Z., Holo, H., and Nes, I. F. (2007). Common mechanisms of target cell recognition and immunity for class II bacteriocins. *Proceedings of the National Academy of Sciences*, 104(7):2384–2389.

Drider, D., Fimland, G., Hechard, Y., McMullen, L. M., and Prevost, H. (2006). The Continuing Story of Class IIa Bacteriocins. *Microbiology and Molecular Biology Reviews*, 70(2):564–582.

Durbin, R., Eddy, S., Krogh, a., and Mitchison, G. (1998). *Biological sequence analysis*.

Edgar, R. C. (2004a). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Edgar, R. C. (2004b). MUSCLE User Guide. *Nucleic Acids Research*, 32(November):1–15.

Eijsink, V. G. H., Skeie, M., Middelhoven, P. H., Brurberg, M. B., and Nes, I. F. (1998). Comparative studies of class IIa bacteriocins of lactic acid bacteria. *Applied and Environmental Microbiology*, 64(9):3275–3281.

Emmert, D. B., Stoehr, P. J., Stoesser, G., and Cameron, G. N. (1994). The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Research*, 22(17):3445–3449.

Ennahar, S., Sashihara, T., Sonomoto, K., and Ishizaki, A. (2000). Class IIa bacteriocins: biosynthesis, structure and activity. *FEMS Microbiology Reviews*, 24(1):85–106.

Fimland, G., Eijsink, V. G., and Nissen-Meyer, J. (2002). Comparative studies of immunity proteins of pediocin-like bacteriocins.

Fleming, A. (1929). On the antibacterial action of cultures of a penicillium, with special reference to their use in the isolation of B.influenzae. *British journal of experimental pathology*, 10(1923):226–236.

Fryxell, K. J. (1996). The coevolution of gene family trees.

Hammami, R., Zouhir, A., Le Lay, C., Ben Hamida, J., and Fliss, I. (2010). BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiology*, 10(1):22.

Hammami, R., Zouhir, A., Le Lay, C., Ben Hamida, J., and Fliss, I. (2016). BactiBase.

Haugen, H. S., Fimland, G., Nissen-Meyer, J., and Kristiansen, P. E. (2005). Three-dimensional structure in lipid micelles of the pediocin-like antimicrobial peptide curvacin A. *Biochemistry*, 44(49):16149–16157.

Heng, N. C. K., Wescombe, P. A., Burton, J. P., Jack, R. W., and Tagg, J. R. (2007). *The diversity of bacteriocins in gram-positive bacteria*.

Herman, D., Ochoa, D., Juan, D., Lopez, D., Valencia, A., and Pazos, F. (2011). Selection of organisms for the co-evolution-based study of protein interactions. *BMC Bioinformatics*, 12.

J Wheeler, T., Clements, J., and Finn, R. (2014). *Skylign: A tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models*, volume 15.

Jack, R. W., Tagg, J. R., and Ray, B. (1995). Bacteriocins of gram-positive bacteria. *Microbiological reviews*, 59(2):171–200.

Joerger, M. C. and Klaenhammer, T. R. (1986). Characterization and purification of helveticin J and evidence for a chromosomally determined bacteriocin produced by Lactobacillus helveticus 481. *Journal of Bacteriology*, 167(2):439–446.

John N. Thompson (1994). *The Coevolutionary Process*. University of Chicago Press.

Johnsen, L., Fimland, G., Mantzilas, D., and Nissen-Meyer, J. (2004). Structure-function analysis of immunity proteins of pediocin-like bacteriocins: C-terminal parts of immunity proteins are involved in specific recognition of cognate bacteriocins. *Applied and Environmental Microbiology*, 70(5):2647–2652.

Johnsen, L., Fimland, G., and Nissen-Meyer, J. (2005). The C-terminal domain of pediocin-like antimicrobial peptides (class IIa bacteriocins) is involved in specific recognition of the C-terminal part of cognate immunity proteins and in determining the antimicrobial spectrum. *Journal of Biological Chemistry*, 280(10):9243–9250.

Juan, D., Pazos, F., and Valencia, A. (2008). High-confidence prediction of global interactomes based on genome-wide coevolutionary networks. *Proceedings of the National Academy of Sciences*, 105(3):934 LP – 939.

Jung, G. (1991). Nisin and novel lantibiotics. In *Nisin and novel lantibiotics*, pages 1–34.

Kjos, M. (2012). *Class II bacteriocins: target recognition, resistance and immunity*. PhD thesis, Norwegian University of Life Science.

Kjos, M., Borrero, J., Opsata, M., Birri, D. J., Holo, H., Cintas, L. M., Snipen, L., Hernández, P. E., Nes, I. F., and Diep, D. B. (2011). Target recognition, resistance, immunity and genome mining of class II bacteriocins from Gram-positive bacteria.

Kjos, M., Salehian, Z., Nes, I. F., and Diep, D. B. (2010). An extracellular loop of the mannose phosphotransferase system component IIC is responsible for specific targeting by class IIa bacteriocins. *Journal of Bacteriology*, 192(22):5906–5913.

Kodama, Y., Mashima, J., Kosuge, T., Kaminuma, E., Ogasawara, O., Okubo, K., Nakamura, Y., and Takagi, T. (2018). DNA Data Bank of Japan: 30th anniversary. *Nucleic Acids Research*, 46(Database issue):D30–D35.

König, H. and Fröhlich, J. (2017). Lactic Acid Bacteria BT - Biology of Microorganisms on Grapes, in Must and in Wine. pages 3–41. Springer International Publishing, Cham.

Métivier, A., Pilet, M. F., Dousset, X., Sorokine, O., Anglade, P., Zagorec, M., Piard, J. C., Marion, D., Cenatiempo, Y., and Fremaux, C. (1998). Divercin V41, a new bacteriocin with two disulphide bonds produced by Carnobacterium divergens V41: Primary structure and genomic organization. *Microbiology*, 144(10):2837–2844.

Morton, J. T., Freed, S. D., Lee, S. W., and Friedberg, I. (2015). A large scale prediction of bacteriocin gene blocks suggests a wide functional spectrum for bacteriocins. *BMC Bioinformatics*, 16:381.

Muley, V. Y. and Ranjan, A. (2012). Effect of reference genome selection on the performance of computational methods for genome-wide protein-protein interaction prediction. *PLoS ONE*, 7(7).

Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similiarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453.

Neu, H. C. (1992). The Crisis in Antibiotic Resistance. *Science*, 257(5073):1064 LP – 1073.

Nissen-Meyer, J., Rogne, P., Oppegård, C., Haugen, H. S., and Kristiansen, P. E. (2009). Structure-function relationships of the non-lanthionine-containing peptide (class II) bacteriocins produced by gram-positive bacteria. *Current pharmaceutical biotechnology*, 10(1):19–37.

Ochoa, D. (2010). MirrorTree Server.

Ochoa, D., Juan, D., Valencia, A., and Pazos, F. (2015). Detection of significant protein coevolution. *Bioinformatics*, 31(13):2166–2173.

Ochoa, D. and Pazos, F. (2010). Studying the co-evolution of protein families with the Mirrortree web server. *Bioinformatics*, 26(10):1370–1371.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D., and Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(Database issue):D733–D745.

Pazos, F. and Valencia, A. (2001). Similarity of phylogenetic trees as indicator of proteinprotein interaction. *Protein Engineering, Design and Selection*, 14(9):609–614.

Pazos, F. and Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. *The EMBO Journal*, 27(20):2648–2655.

Pearson, K. (1901). LIII. ¡i¿On lines and planes of closest fit to systems of points in space¡/i¿. *Philosophical Magazine Series 6*, 2(11):559–572.

Piper, C., Cotter, P. D., Ross, R. P., and Hill, C. (2009). Discovery of Medically Significant Lantibiotics. *Current Drug Discovery Technologies*, 6:1–18.

Postma, P. W., Lengeler, J. W., and Jacobson, G. R. (1993). Phosphoenolpyruvate: Carbohydrate phosphotransferase systems of bacteria. *Microbiological reviews.*, 57(3):543–594.

Quadri, L. E. N., Sailer, M., Roy, K. L., Vederas, J. C., and Stiles, M. E. (1994). Chemical and genetic characterization of bacteriocins produced by ¡i¿Carnobacterium piscicola¡/i¿ LV17B. *Journal of Biological Chemistry*, 269(16):12204–12211.

R., K. T. (1993). Genetics of bacteriocins produced by lactic acid bacteria*. *FEMS Microbiology Reviews*, 12(13):39–85.

Rea, M., Ross, R., Cotter, P., and Hill, C. (2011). Classification of Bacteriocins from Gram-Positive Bacteria. In *Prokaryotic Antimicrobial Peptides SE - 3*, pages 29–53.

Sean R. Eddy, T. J. W. (2015). *HMMER Users Guide*. Howard Hughes Medical Institute.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, 30(14):2068–2069.

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Snipen, L. and Liland, K. H. (2015). micropan: An R-package for microbial pan-genomics. *BMC Bioinformatics*, 16(1):1–8.

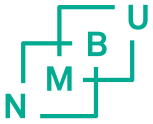
Snipen, L., Wassenaar, T. M., Altermann, E., Olson, J., Kathariou, S., Lagesen, K., Takamiya, M., Knøchel, S., Ussery, D. W., and Meinersmann, R. J. (2012). Analysis of evolutionary patterns of genes in Campylobacter jejuni and C. coli. *Microbial informatics and experimentation*, 2(1):8.

van Heel, A. J., de Jong, A., Montalbán-López, M., Kok, J., and Kuipers, O. P. (2013). BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic acids research*, 41(Web Server issue).

Ventola, C. L. (2015). The Antibiotic Resistance Crisis: Part 1: Causes and Threats. *Pharmacy and Therapeutics*, 40(4):277–283.

Willey, J. M. and van der Donk, W. A. (2007). Lantibiotics: Peptides of Diverse Structure and Function. *Annual Review of Microbiology*, 61(1):477–501.

Zaman, S. B., Hussain, M. A., Nye, R., Mehta, V., Mamun, K. T., and Hossain, N. (2017). A Review on Antibiotic Resistance: Alarm Bells are Ringing. *Cureus*.



Norges miljø- og biovitenskapelige universitet
Noregs miljø- og biovitskapelege universitet
Norwegian University of Life Sciences

Postboks 5003
NO-1432 Ås
Norway