# DESIGN AND CLASSIFICATION OF ANTIMICROBIAL AND ANTIBACTERIAL PEPTIDES

Nydia Paola Rondón Villarreal
Ingeniera de Sistemas

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Ingenierías Eléctrica, Electrónica y Telecomunicaciones
Doctorado en Ingeniería, Área Ingeniería Electrónica
Bucaramanga, Santander, Colombia
2016

# DESIGN AND CLASSIFICATION OF ANTIMICROBIAL AND ANTIBACTERIAL PEPTIDES

Nydia Paola Rondón Villarreal
Ingeniera de Sistemas

Trabajo de investigación presentado para optar al título de
Doctora en Ingeniería

Director
Daniel Alfonso Sierra Bueno
Doctor en Ingeniería Biomédica

Co-director
Rodrigo Gonzalo Torres Sáez
Doctor en Bioquímica y Biología Molecular

Universidad Industrial de Santander
Facultad de Ingenierías Fisicomecánicas
Escuela de Ingenierías Eléctrica, Electrónica y Telecomunicaciones
Doctorado en Ingeniería, Área Ingeniería Electrónica
Bucaramanga, Santander, Colombia
2016

*To God,*

*To my parents,*

*To my sister,*

*To my little nephews...*

# ACKNOWLEDGMENTS

Thank you God for giving me health, wisdom, guidance, and everything that I need to make my dreams come true.

I would also like to thank my supervisors Dr. Daniel Alfonso Sierra Bueno and Dr. Rodrigo Torres Sáez for the invaluable support, lessons, patience, friendship, encouragement, and advices. I do not have enough words to thank them for their commitment during my Ph.D. studies. Thank you *Profes* for everything.

My gratitude also goes to Ph.D. Marcel Reinders and Ph.D. David Tax for giving me the opportunity to work with them at the Bioinformatics and Pattern Recognition Lab at TUDelft, during my research internship, and for all the support, lessons, guidance and the valuable contributions to the development of the classifiers. Many thanks to all my colleagues from this lab for their friendship, their help and the good talks with coffee and orange juice. And also, many thanks to my good friends Sasheeka, Tania, Blanca, José and Tom for their friendship and the good moments that we lived together. Without you guys my internship in the NL would not have been as wonderful as it was, and you are always welcome home.

I express my deep gratitude to my parents and my sister for their love and support in all the decisions that I have made, including the one of doing a Ph.D. Thank you very much for always being there for me, and I hope to be always there for you.

I would like to thank my dearest friends Yulieth, Cafa and Dario, for their unconditional friendship and support, and for have walked with me this path. Thank you guys for being there at the good and not so good moments that we had during this long journey. I wish you the best and cheers for all the good moments that are waiting for us.

Finally, I thank my friends Caro and Juli, my family and all the persons that were there for me during all this time of my Ph.D. studies.

# Contents

# Figure List

# Table List

# RESUMEN

**TÍTULO:** DISEÑO Y CLASIFICACIÓN DE PÉPTIDOS ANTIMICROBIANOS Y ANTIBACTERIANOS [1].

**AUTOR:** NYDIA PAOLA RONDÓN VILLARREAL [2]

**PALABRAS CLAVES:** PÉPTIDOS ANTIBACTERIANOS, CLASIFICACIÓN, DISEÑO.

**DESCRIPCIÓN:**

Uno de los problemas de salud pública más importantes es la resistencia a los antibióticos que poseen bacterias patógenas de gran impacto en la salud humana. El problema es tan importante que puede afectar a la medicina moderna, como por ejemplo, en el área de cirugías especializadas, debido al gran riesgo de adquirir una bacteria super resistente intrahospitalaria, que no pueda ser tratada con los antibiticos existentes. La situacin es bastante desalentadora, la resistencia a los antibiticos está creciendo a tasas alarmantes y el número de nuevos antibióticos desarrollados y probados ha disminuido en las últimos décadas, básicamente por razones económicas y de regulación. En este sentido, múltiples empresas farmacéuticas han abandonado la investigación y el desarrollo de nuevos compuestos antimicrobianos. Sin embargo, en los últimos años, un buen número de investigadores se ha enfocado en el desarrollo de nuevos antibióticos. Entre estos, los péptidos antimicrobianos (PAMs) han aparecido como una solución prometedora para combatir estas bacterias super resistentes. Por esta razón, múltiples esfuerzos teóricos se han llevado a cabo en el desarrollo de nuevas herramientas computacionales para el diseño racional de péptidos que sean mejores y más efectivos.

En esta tesis, se proponen dos estrategias para diseñar nuevos péptidos antibacterianos potenciales. Adicionalmente, la toxicidad de los péptidos también fue considerada en una de las estrategias propuestas. Los resultados han sido bastante satisfactorios. Múltiples péptidos que fueron diseñados en esta tesis fueron sintetizados y probados a nivel experimental y han mostrado actividad contra tres bacterias resistentes a los antibióticos. Adicionalmente, se realizaron pruebas de toxicidad a los péptidos más activos, y resultaron ser no tóxicos en eritrocitos de carnero y en células de tejido de pulmón de la línea A549.

[1] Trabajo de grado de doctorado

[2] Facultad de ingenierías fisicomecánicas. Escuela de ingenierías eléctrica, electrónica y de telecomunicaciones (E3T). Director: Daniel Alfonso Sierra Bueno, Ph.D., co-director: Rodrigo Gonzalo Torres Sáez, Ph.D.

# ABSTRACT

**TITLE:** DESIGN AND CLASSIFICATION OF ANTIMICROBIAL AND ANTIBACTERIAL PEPTIDES [1].

**AUTHOR:** NYDIA PAOLA RONDÓN VILLARREAL [2]

**KEYWORDS:** ANTIBACTERIAL PEPTIDES, CLASSIFICATION, DESIGN.

**DESCRIPTION:**

One of the most important public health issues is the microbial and bacterial resistance to conventional antibiotics by pathogen microorganisms. This issue is so serious that modern medicine could be affected, e.g. in the area of advanced surgeries, due to the great risk of acquiring intrahospitalary multidrug-resistant bacterial infections that cannot be controlled by existing medications. The situation is even worse, antimicrobial resistance is increasing at an alarming rate and the number of new antibiotics developed and approved has decreased in the last decades, basically for economic and regulatory obstacles. In this regard, multiple pharmaceutical industries have abandoned the research and development of new antimicrobial compounds. However, in recent years, many researches have been focused in the development of new antibiotics. Among these, antimicrobial peptides (AMPs) have raised as a promising alternative to combat antibiotic-resistant microorganisms. For this reason, many theoretical efforts have been done in the development of new computational tools for the rational design of both better and effective AMPs.

In this thesis, two strategies to design new potential antibacterial peptides are proposed. Moreover, the toxicity of the peptides was also considered in one of the proposed strategies. The results have been highly satisfactory. Peptides that were designed in this thesis have been tested experimentally and they have been active against three strains of multidrug-resistant bacteria. Additionally, toxicity tests were performed for the most active peptides, and they are non-toxic neither on lamb erythrocytes nor on lung tissue cell line A549.

[1] Doctoral dissertation: Ph.D. on Engineering

[2] Facultad de ingenierías fisicomecánicas. Escuela de ingenierías eléctrica, electrónica y de telecomunicaciones (E3T). Advisor: Daniel Alfonso Sierra Bueno, Ph.D., co-advisor: Rodrigo Gonzalo Torres Sáez, Ph.D.

# Chapter 1

# Introduction

Antimicrobial resistance is one of the most important public health problems around the world, especially in developing countries where economical, biological, pharmacological and cultural aspects increase this problem [1]. This issue is so serious that modern medicine could be affected, e.g. in the area of advanced surgeries, due to the great risk of acquiring intrahospitalary multidrug-resistant bacterial infections that cannot be controlled by existing medications [2,3].

According with the CDC urgent solutions are required for at least 15 microorganisms (See Table 1.1). The situation is extremely critical for three of them, *Clostridium difficile*, Carbapenem-resistant *Enterobacteriaceae* (CRE) and Drug-resistant *Neisseria gonorrhoeae.* Only the *C. difficile* is responsible for more than 250,000 infections, 14,000 deaths and more than $ 1 billion in excess medical costs per year in the United States [4]. It is estimated that the yearly costs of antimicrobial resistance to the U.S. health system is between $21 to $34 billion dollars [5].

Similarly, the European Center for Disease, Prevention and Control (ECDC) estimated in 2009 that each year 25,000 deaths in Europe are caused by resistant bacteria. Moreover, it was estimated that at least EUR 1.5 billion are spent in health care costs and productivity losses associated with these bacterial infections. Since the antibacterial resistance is increasing, it is likely that these numbers are higher nowadays [6].

In 2014, the World Health Organization (WHO) published the antimicrobial resistant global report on surveillance, pointing out that urgent actions are required to control these microorganisms. The data reported by WHO show that estimated proportion of resistance for bacteria *Escherichia coli, Klebsiella pneumoniae, Staphylococcus aureus, Salmonella* and *Shigella* species is over the 80% in at least one of the world regions (See Table 1.2). However, WHO remarks that the given proportions of resistance in microorganisms should be treated as indicators and not as measures due to the lack of agreed global standards for antibacterial surveillance [5].

The situation is even worse, antimicrobial resistance is increasing at an alarming rate and the number of new antibiotics developed and approved has decreased in the last decades, basically for economic and regulatory obstacles. In this regard, multiple pharmaceutical industries have abandoned the research and development of new antimicrobial compounds [7–9]. High costs of production (US $ 2.6 billion) [10], and the long time required for the development and posterior approval of the new medicine (more than 10 years) [10] have affected dramatically the arising of both new and more effective antibiotics. Moreover, the short duration of the treatment with antimicrobial agents (no more than 10 days), in comparison with the medicine for chronic diseases that would be used in treatments for more than one year, restrict their commercial market [11–13]. Additionally, the approval for a new antibacterial medicine is often an obstacle, due to an inflexible regulatory pathway, and difficulties such as differences in clinical trials requirements among countries, bureaucracy, absence of clarity, among others [7].

Table 1.1: Classification of dangerous microorganisms according with the CDC. Data per year in the United States.

| Microorganisms | No. I.R.S.[a] | No. Deaths[b] | A.M.C.[c] |
|---|---|---|---|
| Threat level of Urgent | | | |
| *Clostridium Difficile* | 250,000 | 14,000 | 1,000,000,000 |
| Carbapenem-resistant *Enterobacteriaceae* | 9,000 | 600 | - |
| Drug-resistant *Neisseria Gonorrhoeae* | 246,000 | - | - |
| Threat level of Serious | | | |
| Multidrug-resistant *Acinetobacter* | 7,300 | 500 | - |
| Drug-resistant *Campylobacter* | 1,300,000 | 120 | - |
| Fluconazole-resistant *Candida* | 3,400 | 220 | 6,000 - 29,000 f.e.i.[d] |
| Extended spectrum $\beta$-lactamase producing *Enterobacteriaceae* | 26,000 | 1,700 | 40,000 f.e.i |
| Vancomycin-resistant *Enteoroccus* | 20,000 | 1,300 | - |
| Multidrug-resistant *Pseudomonas Aeruginosa* | 6,700 | 440 | - |
| Drug-resistant non-typhoidal *Salmonella* | 100,000 | 450 | 365,000,000 |
| Drug-resistant *Salmonella Serotype Typhi* | 3,800 | - | - |
| Drug-resistant *Shigella* | 27,000 | 40 | - |
| Methicillin-resistant *Staphylococcus Aureus* | 80,461 | 11,285 | - |
| Drug-resistant *Streptococcus Pneumoniae* | 1,200,000 | 7,000 | 96,000,000 |
| Drug-resistant *Tuberculosis* | 1,042 | - | - |
| Threat level of Concerning | | | |
| Vancomycin-resistant *Staphylococcus Aureus* | - | - | - |
| Erythromycin-resistant group A *Streptococcus* | 1,300 | 160 | - |
| Clindamycin-resistant group B *Streptococcus* | 7,600 | 440 | - |

[a]Number of infections with resistant strains. [b]Number of Deaths. [c]Associated medical costs ($U.S. dollars). [d]f.e.i.=For each infection.

Table 1.2: Proportions of resistance for bacteria of international concern[a]

| Bacteria | AFR[b] | AMR[c] | EMR[d] | EUR[e] | SEAR[f] | WPR[g] |
|---|---|---|---|---|---|---|
| *Escherichia coli*[h] | 2 - 70 | 0 - 48 | 22 - 63 | 3 - 82 | 16 - 68 | 0 - 77 |
| *Escherichia coli*[i] | 14 - 71 | 8 - 58 | 21 - 62 | 8 - 48 | 32 - 64 | 3 - 96 |
| *Klebsiella pneumoniae*[j] | 8 - 77 | 4 - 71 | 22 - 50 | 2 - 82 | 34 - 81 | 1 - 72 |
| *Klebsiella pneumoniae*[k] | 0 - 4 | 0 - 11 | 0 - 54 | 0 - 68 | 0 - 8 | 0 - 8 |
| MR *Staphylococcus aureus*[l] | 12 - 80 | 21 - 90 | 10 - 53 | 0.3 - 60 | 10 - 26 | 4 - 84 |
| Nontyphoidal *Salmonella*[m] | 0 - 35 | 0 - 96 | 2 - 49 | 2 - 3 | 0.2 - 4 | 0 - 14 |
| *Shigella* species[n] | 0 - 3 | 0 - 8 | 3 - 10 | 0 - 47 | 0 - 82 | 3 - 28 |
| *Neisseria gonorrhoeae* | 0 - 12 | 0 - 31 | 0 - 12 | 0 - 36 | 0 - 5 | 0 - 31 |

[a]Data from national sources. [b]African Region. [c]Region of the Americas. [d]Eastern Mediterranean Region. [e]European Region. [f]South-East Asia Region. [g]Western Pacific Region. [h]*Escherichia coli* resistance to third-generation cephalosporins. [i]*Escherichia coli* resistance to fluoroquinolones. [j]*Klebsiella pneumoniae* resistance to third-generation cephalosporins. [k]*Klebsiella pneumoniae* resistance to carbapenems. [l]methicillin-resistant *Staphylococcus aureus.* [m]Nontyphoidal *Salmonella* resistance to fluoroquinolones. [n]*Shigella* species resistance to fluoroquinolones. [o]*Neisseria gonorrhoeae* decreased susceptibility to third-generation cephalosporins.

According with Brown and Wright [14] the post-antibiotic age has arrived. Pathogenic bacteria that are resistant to multiple or all available antibiotics are isolated frequently. Hence, new antibiotics are urgently needed, and despite multiple efforts, only two new antibiotics (telavancin and ceftaroline) have been approved since 2009 [15]. In this sense, most of the available antibiotics were discovered in the golden era: sulfonamides (sulfanilamide), $\beta$-Lactams (penicillins, cephalosporins, carbapenems), aminoglycosides (spectinomycin, kanamycin, neomycin), tetracyclines (tetracycline, doxycycline), chloramphenicols (chloramphenicol), macrolides (erythromycin, clarithromycin), glycopeptides (vancomycin, teicoplanin), oxazolidinones (linezolid), ansamycins (rifamycin), quinolones (ciprofloxacin), streptogramins (pristinamycin). Years later, the medicinal chemistry era started, and the development of new antibiotics was focused on the creation of synthetic versions of the natural antibiotics. Therefore, most of the existing antibiotics tend to target the bacterial cell wall, DNA or ribosomes, and are derived from natural sources. Later, by the 1990s the resistance era began and the emphasis was on target-based drug discovery to find broad-spectrum agents. Unfortunately, this model has failed to provide new antibiotics, and it is important to understand the failures of target-based approaches in order to obtain better results and new potential antibiotics [14].

The message is clear, new strategies to stimulate the research and development of new antibacterials are required. Therefore, multiple organizations, including IDSA (Infectious Diseases Society of America), have proposed The 10 x '20 Initiative with the aim of developing 10 new antibiotics by the year 2020. These organizations are encouraging the pharmaceutical companies, scientists, politicians and the whole humanity in the search of new solutions for the antimicrobial resistance issue, especially for the multidrug-resistant bacteria since they represent a great risk for humankind [2]. Moreover, there are other initiatives that are also playing an important role in the fight against antibacterial resistance. Among them are: the Joint Programming Initiative on Antimicrobial Resistance supported by Canada and 18 European countries, World Alliance Against Antibiotic Resistance (WAAAR, France), Antibiotic Union (UK), ReAct(Sweden), and the Antibiotic Resistance Initiative (ISGlobal, Spain) [9].

In the fight against super bacteria, antimicrobial peptides (AMPs) have appeared as a promising solution to this important public health problem. In the last years, the scientific community has been using computational tools to design new AMPs [16]. The main objective of these *in silico* [1] models is to search for peptides that are likely to present antimicrobial activity, thus reducing the cost and the time spent in the synthesis of peptides without any biological activity or with high toxicity for human beings.

Therefore, the main goal of this thesis was to propose a methodology to design new potential non-toxic antibacterial peptides, based on the **Hypothesis** that

It is possible to use genetic algorithms, classification methods, string kernel methods, and peptide descriptors to design effective non-toxic antibacterial peptides.

One of the important aspects to consider in the design of new antibacterial peptides was to reduce the number of false positives, with the aim to avoid the synthesis of peptides that would not possess antibacterial activity or peptides that would be toxic.

[1] *in silico* means performed on computer

Two strategies were proposed. Both of them involve the creation of a genetic algorithm that allows the design of peptides, that hold with established ranges for physicochemical properties such as charge, isoelectric point, hydrophobicity and instability index. These ranges are established by the user, and both genetic algorithms generate the number of candidate peptides that the user desires. Although the genetic algorithms are similar in their design, there is an important difference between them. The optimization function of the genetic algorithm in the second strategy includes an estimation of the probability that a peptide will possess antibacterial activity, and an estimation of the probability of being a toxic peptide. These estimations of the probabilities are given by classifiers created using logistic classifier and a larger dataset that the one that was used in the first strategy. In contrast, the genetic algorithm of the first strategy does not consider neither the antibacterial activity nor the toxicity of the peptides during the design process of the peptides using the genetic algorithm. Therefore, the first strategy requires additional steps after the design of the peptides with the genetic algorithm, i.e., the designed peptides should be classified as antimicrobial and as antibacterial peptides before the analysis of their secondary structure. In this sense, we believe that the design of new non-toxic antibacterial peptides is more efficient using the strategy 2. A graphical summary of the strategies designed in this thesis is shown in Figure 1.1.

The obtained results are highly satisfactory. More than 18 peptides designed in this thesis have been tested experimentally and they have been active against *Staphylococcus Aureus*, *Pseudomona aeruginosa*, and *Scherichia coli O157:H7* strains. Additionally, toxicity tests were performed for the most active peptides, and they are non-toxic neither on lamb erythrocytes nor on lung tissue cell line A549 [1]. Two of these peptides (GIBIM-P6 and GIBIM-P5F8W), and the proposed strategies are in the process of being patented.

Finally, parts of this thesis have been published in [17, 18] and are going to be published in [19].

[1] The experimental tests of the peptides have been performed by undergrad, master and PhD students of chemistry at the GIBIM Lab - Universidad Industrial de Santander

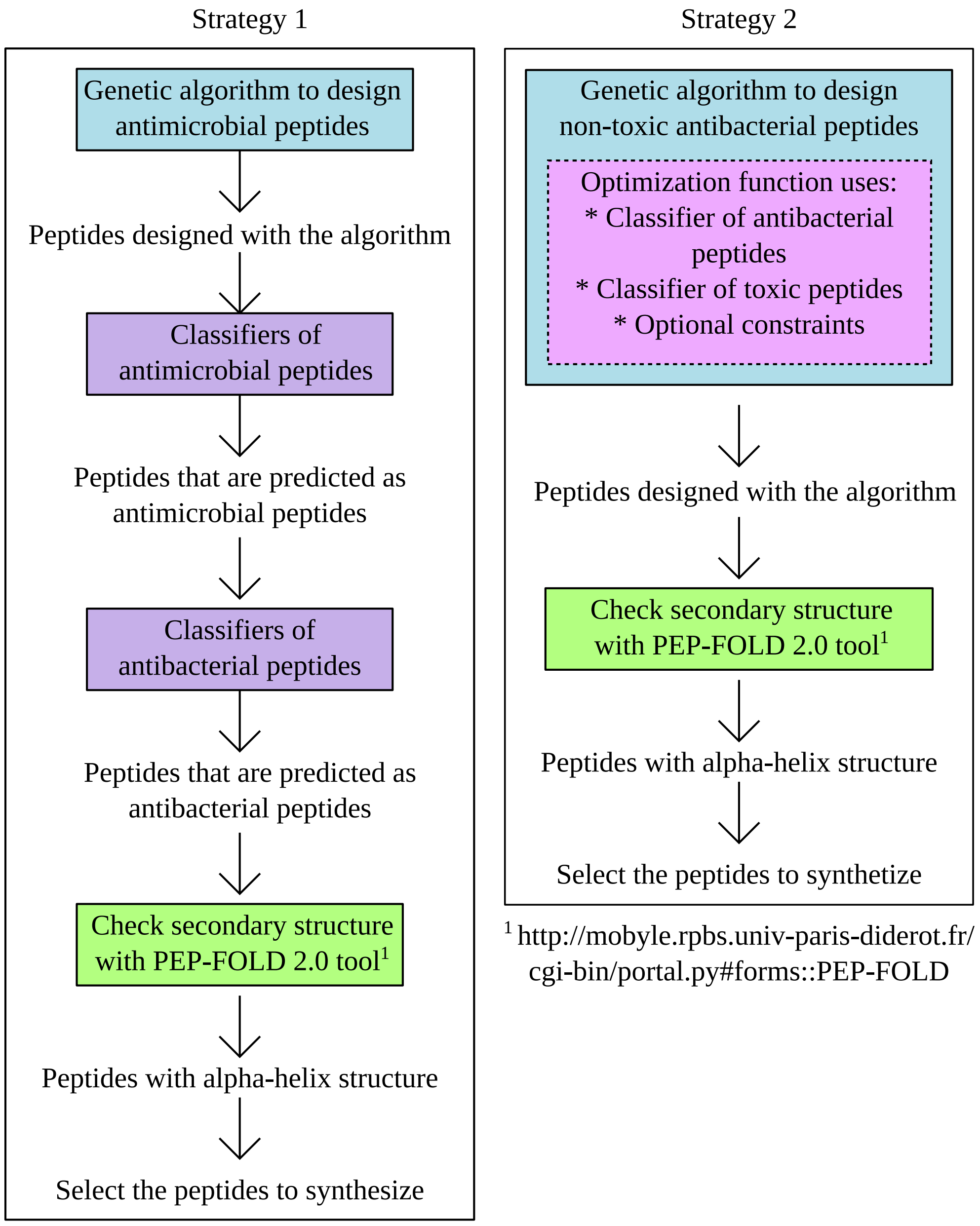

Strategy 1
Genetic algorithm to design antimicrobial peptides
Peptides designed with the algorithm
Classifiers of antimicrobial peptides
Peptides that are predicted as antimicrobial peptides
Classifiers of antibacterial peptides
Peptides that are predicted as antibacterial peptides
Check secondary structure with PEP-FOLD 2.0 tool[1]
Peptides with alpha-helix structure
Select the peptides to synthesize
Strategy 2
Genetic algorithm to design non-toxic antibacterial peptides
Optimization function uses:
* Classifier of antibacterial peptides
* Classifier of toxic peptides
* Optional constraints
Peptides designed with the algorithm
Check secondary structure with PEP-FOLD 2.0 tool[1]
Peptides with alpha-helix structure
Select the peptides to synthetize
[1] http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::PEP-FOLD

Figure 1.1: Graphical summary of the strategies designed in this thesis.

# Chapter 2

# Rational design of antimicrobial and antibacterial peptides

## 2.1 Antimicrobial peptides

The antimicrobial peptides (AMPs) are small host defense proteins that are part of the immune system of multicellular organisms such as plants and animals. In general, they possess a positive net charge and hydrophobic percentages greater than 30% [20, 21]. The typical length of these peptides is not well established and there are different opinions in this regard. Some authors state that their length is less than 50 amino acids [21–23], while others claim that their length is less than 100 [24–26].

Additionally, these peptides are active against multiple microorganisms such as virus, bacteria, parasites, among others [21, 25, 27]. Due to their several advantages over small molecules, such as high specificity, high penetration, ease of manufacturing, the antimicrobial peptides have emerged as a promising solution for the threat of antibacterial resistance [28]. Moreover, these peptides are able to destroy the bacterial membrane or essential component inside the cell. Therefore, it is less likely that bacteria develop AMP resistance, since they would require the complete alteration of the membrane and/or bypassing of several biochemical pathways [29].

However, AMPs generally have low stability *in vivo* and are degraded by both endogenous human proteases and proteases secreted by invading microbes [29]. Additionally, problems such as toxicity and high cost of production are important drawbacks for therapeutic applications of these peptides [29]. Therefore, the main goal is to obtain new peptides that possess high antibacterial activity, low toxicity and low propensity to proteolytic degradation.

On the other hand, it is important to mention that the function of a peptide depends on its primary structure, i.e., its amino acid sequence. Theory suggests that the amino acid sequence of a peptide plays an important role in determining its three-dimensional structure, and thereby its function. If two proteins have a different function, then these two proteins have different amino acid sequences. Moreover, multiple genetic diseases are related to the production of defective proteins. These proteins are defective because their primary structures have been changed by only one single amino acid alteration, and this alteration may lead to a change in the function of the protein. Moreover, proteins that display similar functions in different species often have similar amino acid regions that are essential to their function, and therefore, these regions are conserved [30]. In the same way, it is likely that peptides that display antibacterial activity also display conserved regions, in their amino acid sequences, that are responsible for the antibacterial activity.

### 2.1.1 Mechanisms of action of antimicrobial peptides in bacteria

Antibacterial peptides possess interesting advantages in comparison with conventional antibiotics, such as the speed in killing bacteria and the fact that they are not hindered by the resistance mechanisms developed by bacteria [31].

Studies have shown that all peptides interact with membranes and two different mechanisms have been identified: membrane-disruptive (barrel stave, toroidal, carpet and micellar aggregate models) and non-membrane disruptive (intracellular targets) [16].

Bacterial cell membranes contain high proportion of negatively charged phospholipid headgroups such as phosphatidylglycerol (PG), cardiolipin (CL), or phosphatidylserine (PS). Hence, the initial contact between bacterial membranes and cationic peptides occurs through electrostatic interactions [27].

In the **barrel stave** model, a relatively small number of individual peptides are perpendicularly inserted and aggregated in a *barrel-like* ring inside the membrane, leading to a transmembrane pore or channel with a cylindrical structure. This type of interaction is typical for small number of peptides. In the **toroidal pore** model, peptides are inserted from the membrane surface into the hydrophobic part of the lipid bilayer and the pore wall is formed by both, hydrophilic regions of peptide molecules and lipid head groups. The **carpet** mechanism occurs when there are strong electrostatic interactions between peptides and negatively charged phospholipid polar head groups. The peptides bind to the membrane carpeting the phospholipid bilayer. When some critical threshold concentration of peptide is reached, the rupture of the membrane takes place followed by micelle formation. In the **aggregate** model the peptides are clustered, after binding to the phospholipid head groups, into unstructured aggregates that allow the pore formation that span the membrane for short periods. This mechanism allows antimicrobial peptides to enter the intracellular space to perform their killing activities. Finally, there is an increasing evidence that the formation of ion channels, transmembrane pores and membrane rupture are not the only mechanisms of microbial killing, indicating that antimicrobial peptides have other **intracellular targets** [16, 27].

## 2.2 Rational design of antimicrobial peptides

In the rational design of antimicrobial peptides, *in silico* models are used or created in order to identify peptides that could become new antimicrobial medicines. In this sense, multiple strategies have been used in the design of antimicrobial peptides. Among them, it is possible to find three major approaches: improvement of the existing peptides, regression and classification.

The information to create these *in silico* models is available in multiple databases of antimicrobial peptides. Table 2.1 shows a summary of these databases.

Table 2.1: Databases for antimicrobial peptides*.

| **Database** | **Peptides** | **Availability** |
|---|---|---|
| AMSDb [32] | Plant/Animal AMPs | Inactive |
| SAPD [33] | Synthetic antibiotic peptides | Inactive |
| Peptaibol [34] | Peptaibols | http://www.cryst.bbk.ac.uk/peptaibol |
| APD [35,36] | AMPs | http://aps.unmc.edu/AP/main.php |
| DAMPD (updated version of ANTIMIC) [37] | AMPs | http://apps.sanbi.ac.za/dampd/index.php |
| PenBase [38] | Penaeidins | http://penbase.immunaqua.com |
| CyBase [39] | Cyclic proteins | http://www.cybase.org.au/index.php?page=welcome |
| BAGEL2 [40] | Bacteriocins | http://bagel2.molgenrug.nl/ |
| AMPer [41] | Gene-coded AMPs | http://marray.cmdr.ubc.ca/cgi-bin/amp.pl. |
| BACTIBASE [42,43] | Bacteriocins | http://bactibase.pfba-lab-tun.org/main.php |
| Defensins [44] | Defensins | http://defensins.bii.a-star.edu.sg/ |
| RAPD [45] | Recombinant AMPs | http://faculty.ist.unomaha.edu/chen/rapd/index.php |
| phytAMP [46] | Plant AMPs | http://phytamp.pfba-lab-tun.org/about.php |
| CAMP [47] | AMPs | http://www.bicnirrh.res.in/antimicrobial/index.php |
| YADAMP [48] | AMPs | http://yadamp.unisa.it/default.aspx |
| DADP [49] | Anuran defense peptides | http://split4.pmfst.hr/dadp/? |
| THIOBASE [50] | Thiopeptides | http://db-mml.sjtu.edu.cn/THIOBASE/ |
| EnzyBase [51] | Cleaving enzymes | http://biotechlab.fudan.edu.cn/database/EnzyBase/home.php |
| LAMP [52] | AMPs | http://biotechlab.fudan.edu.cn/database/lamp/ |
| MilkAMP [32] | Milk AMPs | http://milkampdb.org/home.php |
| DBAASP [53] | AMPs | http://dbaasp.org/home.xhtml |
| BaAMPs [54] | Biofilm-active AMPs | http://www.baamps.it/ |

* Adapted from [55]

### 2.2.1 Peptide descriptors

Generally speaking, peptide descriptors can be viewed as numerical values that give information about the structure, the composition, the topology and the chemical features of a peptide. These descriptors are widely used in the rational design of antimicrobial peptides and they can be organized in four categories: 2D Quantitative Structure-Activity Relationship (QSAR) descriptors, 3D QSAR descriptors, inductive descriptors and other descriptors.

The descriptors of the first category do not need information about the three dimensional orientation of the compound, because they measure the topological and geometrical properties of the molecule. In this category the descriptors can be grouped in constitutional descriptors, electrostatic and Quantum-chemical descriptors, topological descriptors, geometrical descriptors, fragment-based descriptors and molecular fingerprints [56]. In contrast, the 3D QSAR descriptors require the conformation of the compound experimentally or by molecular mechanics. They give numerical information about the compound structure and can be organized in the following groups of descriptors: comparative molecular field analysis, comparative molecular similarity indices analysis, comparative molecular moment analysis, weighted holistic invariant molecular descriptors and grid-independent descriptors [56].

The third category is comprised of those descriptors that allow the quantification of inductive and steric interactions between any substituent and reaction centre, the partial atomic charges, analogues of chemical hardness-softness and electronegativity. They are expressed in terms of the parameters of bound atoms such as their electronegativities, the covalent radii and the intramolecular distances [57].

The last category is created by those descriptors that are not so common in the rational design of antimicrobial peptides. Some examples of these descriptors are: distance between the first and second arginines [58], contact energy between neighboring amino acids [59, 60], and internal dipole moment [61], among others.

### 2.2.2 Improvement of the existing peptides

This approach consists in the modification of the existing peptides in order to obtain better versions of them. The existing peptides are considered as a template that will be changed in a small proportion in order to look for a significant improvement of the antimicrobial activity or a significant reduction of the toxicity. These type of studies have allowed the identification of specific properties of AMPs that could affect the antimicrobial activity, the stability and the toxicity of the peptide. Section 2.2.5 shows a summary of the main properties that have been addressed in these studies. For example, in the study of Kim *et. al* [29], two peptides were designed with enhanced stability and cell specificity by systematic amino acid arrangement. This study used as design criteria the amphipathicity of the peptides when folded into $\alpha$-helical structures, the structural features that are important for effective antimicrobial activity and selectivity, and the reduction of protease-scissile sites as much as possible.

Multiple studies have shown that a single amino acid substitution might enhance antimicrobial activity, reduce the toxicity or both [24,62,63]. In this sense, in the study of Tan *et. al* [24] six amino acids (K, E, G, S, A, and L) were individually used to replace the valine (V) at the sixteenth location of the non-polar face of V13K. The designed peptides showed great antimicrobial activity against Gram-negative bacteria and weak hemolytic activity against human red blood cells. Zhu *et. al.* [62] show that tryptophan at the hydrophobic face has a significant role in transforming an amphipathic peptide into a *P. aeruginosa*-targeted AMP. Similarly, the study of Wu *et. al* [63] proposes an amino acid-based prediction method to improve the activity of antimicrobial peptides by the substitution of one or two amino acids. These substitutions were performed considering the information of the amino acid activity contribution matrix that they developed. This matrix has an activity contribution value for each amino acid in each position of the model peptide. Their designed peptides possessed higher antimicrobial activities than the model peptide.

Additionally, this approach also includes those studies that have used optimization algorithms in order to improve some properties of antimicrobial peptides. For example, Maccari *et. al.* [64] introduced a method for virtual screening of antimicrobial peptides with natural and non-natural amino acids. They used Quantitative structure-activity relationship (QSAR) descriptors to code the peptides and trained two statistical models. The first model represents antimicrobial physicochemical properties, while the second model accounts for the all-helix conformation of the peptide. These models were used as fitness functions for a multi-objective evolutional algorithm. The results were satisfactory, they designed two ab-initio natural peptides and optimized the well-known Cecropin-Mellitin alpha helical antimicrobial peptide, by the reduction of its size while preserving its activity. In the study of Fjell *et. al* [65], a genetic algorithm was designed to generate candidate antibacterial peptide sequences. They found that this approach dramatically lowers the number of peptides that must be evaluated experimentally by the identification of those peptides with high potency. Additionally, they found that the results are dependent on the starting population and that the most reliable predictions will be obtained for peptides that are similar to those ones that were used to build the model. Despite these limitations, they reported several novel peptides that were active against important pathogens. The algorithm created by Juretić *et. al.* [66] allows the design of antimicrobial peptides with a high therapeutic index. They used as reference the gram-negative bacteria *Escherichia coli*, and a total of 73 frog-derived AMPs were collected with their corresponding MIC (Minimal Inhibitory Concentration) values. Their results were satisfactory, one peptide was designed and tested experimentally, and its therapeutic index was higher than the best AMP present in their dataset. However, this study is limited to frog-derived AMPs that are active against *E. coli*.

Finally, due to the huge number of possible amino acids combinations, stochastic optimization methods become a useful tool to perform directed random searches in large problem spaces [64], such as the rational design of antimicrobial peptides.

### 2.2.3 Regression models

The second approach uses regression models in order to create predictors of important characteristics of peptides such as the Minimal Inhibitory Concentration [61, 67], antimicrobial potency [68, 69], and antibacterial activity [22, 70], among others. A summary of these regression applications is shown in Table 2.2.

Table 2.2: Studies of regression in the rational design of Antimicrobial Peptides

| **Tech.**[a] | **Size input dataset** | **Pred. feat.** [b] | **Val. Tech.** [c] | $r^2$ [d] | **Other** [e] |
|---|---|---|---|---|---|
| ANN - [70] | 933(A),<br>500(B),<br>1,433(A+B) | 3 | 3 | - | 0.87±0.10,<br>0.83±0.12,<br>0.80±0.09<br>(AROC) |
| ANN - [68] | 189 | 4 | 4 | 0.85 | 0.72 $q^2$ |
| MLRA - [61] | 37 | 1 | 1 | 0.68 - 0.72 | 0.199 - 0.2230 (SEE) |
| PLS - [67] | 33 | 1 | 1 | 0.975, 0.972 | 0.742, 0.737 ($q^2$) |
| PLS - [69] | 58 | 2 | 2 | 0.73 | 0.61 ($q^2$) |

[a] Technique: ANN= Artificial neural networks, MLRA=Multiple linear regression analysis, PLS=Partial least squares. [b] Predicted feature: 1=Minimal inhibitory concentration, 2=antimicrobial potency in suicide expression system, 3=antibacterial activity, 4=antimicrobial potency. [c] Validation technique: 1=Independent test set, 2=Leave-one-out cross-validation, 3= 10-fold cross-validation, 4= leave-20%-out cross-validation. [d] $r^2$=coefficient of determination. [e] SEE= Standard error of estimation, $q^2$= cross-validated correlation, AROC= Area under the ROC curves (mean±standard deviation).

### 2.2.4 Classification models

The third approach deals with the recognition of patterns inside the reported antimicrobial peptides in order to differentiate them from those that do not present any biological activity, or to discriminate peptides that possess activity against different microorganisms. In general, the classification of antimicrobial or antibacterial peptides is a binary process with an imbalanced initial dataset. However, there are studies dealing with multi class classification [71]. Table 2.3 presents a summary of classification studies.

On the other hand, there are also studies that deal with the classification of toxic peptides. For example, the study of Gupta *et. al.* [28] proposes a hybrid method for the prediction of toxic peptides. The peptides are represented by motifs and their dipeptide composition. In the

Table 2.3: Studies of classification in the rational design of Antimicrobial Peptides

| **Tech.** [a] | **Input dataset** [b] | **Val. Tech.** [c] | **Sn %** [d] | **Sp %** [e] | **Acc %** [f] | **Mcc** [g] |
|---|---|---|---|---|---|---|
| ANN [68] | 1,157/991 | 2 | - | - | 90 | - |
| ANN [72] | 436/436 | 1 | 88.17 | 88.17 | 88.17 | - |
| DA [47] | 2,578/4,011 | 4 | - | - | 87.5 | 0.74 |
| NNA [73] | 2,752/10,014 | 3 | 80.23 | 94.59 | 93.31 | 0.7312 |
| QM [72] | 436/436 | 1 | 90.02 | 90.72 | 90.37 | - |
| RF [47] | 2,578/4,011 | 4 | - | - | 93.2 | 0.86 |
| SVM [74] | 146/146 | 4 | 75.36 | 97.3 | 83.02 | - |
| SVM [47] | 2,578/4,011 | 4 | - | - | 93.2 | 0.86 |
| SVM [72] | 436/436 | 1 | 92.11 | 92.11 | 92.11 | - |
| SVM [75] | 861/861 | 1 | 90.59 | 93.69 | 92.14 | 0.843 |

[a] Technique: ANN=Artificial neural networks, DA=Discriminant Analysis, NNA=Nearest neighbor algorithm, QM= Quantitative matrices, RF=Random Forests, SVM=Support vector machines. [b] Input dataset: Number of positive samples/Number of negative samples. [c] Validation Technique: 1= 5-fold cross-validation, 2= Independent test set, 3=Jackknife test, 4= 10-fold cross-validation [d] Sn=sensitivity. [e] Sp=specificity. [f] Acc=accuracy. [g] Mcc=Matthew's correlation coefficient.

first step various motifs are searched in the query peptides, and if any of the motifs of toxic peptide is present, then its SVM score is increased by the value of 5. The final score is used for the prediction. This model achieved an accuracy around 98%.

### 2.2.5 Properties that might affect the antimicrobial activity and the toxicity of the peptides

In the rational design of antimicrobial and antibacterial peptides it is important to consider the theoretical aspects that have been published based on the experimental results. In this sense, there are multiple physicochemical properties, that can be tuned up by modifications in the amino acid content, that play a crucial role in the activity and toxicity of peptides. Among them, the most well known are size, charge, and hydrophobicity [29, 76].

The **length** of a peptide can affect both, the antibacterial activity and the toxicity. The length of $\alpha$-helical peptides should be at least 22 amino acids in order to transverse the lipid bilayer of bacteria in the barrel-stave model, while for $\beta$-sheet peptides the length should be at least 8 amino acids. However, long peptides could be more toxic than short peptides, like for example a shortened melittin peptide exhibited at least 300 times less toxicity to rat erythrocytes, compared with the original form [76].

It is well known, that a positive net **charge** is essential for the initial interaction with negatively charged cell membranes [76]. The charge value of the known antimicrobial peptides varies between -12 and +30. Considering that the majority of them (97.4%) have a net charge between -5 and +10, this might be an useful range, for this physicochemical property, in the rational design of antimicrobial peptides [55]. However, high values of positive charge might be related with an increment in the toxicity of the peptides.

The **hydrophobicity** also affects the antimicrobial activity of a peptide. In most cases, an increase in the hydrophobicity on the positively charged side can increase its antimicrobial activity. However, it seems that there is an optimal value of hydrophobicity for each peptide, beyond which its activity decreases rapidly [76]. Moreover, an increase in the levels of hydrophobicity might be strongly correlated with mammalian cell toxicity [29]. Therefore, in the rational design of antimicrobial peptides, the hydrophobicity should be selected considering an optimal window [76].

Moreover, the **amino acid composition** of the AMPs is an important feature to consider in the design of effective and non-toxic peptides. The following list shows interesting facts that have been addressed in multiple studies in this regard.

- Leucines (L), glycines (G) and lysines (K) are the most frequent amino acids in all the 2,329 peptides in the APD database, according with Wang [55].
- Proline (P) and glycine (G) are not preferred in the design of $\alpha$-helical peptides, since they have lower helix-forming propensities compared to other amino acids. In fact, it was found that a higher proline content reduced the capability of a peptide to permeabilize *E. coli* cell membrane [76].
- Amino acids with long alipathic side chains (such as valine (V), leucine (L), isoleucine (I) and glutamine (Q)) should be used to increase both the hydrophobic and polar face depths, which are important in modulating membrane interaction and antimicrobial activity [29].
- A reduction of the cytotoxicity (of a human AMP) was achieved by removing asparagine (N) and glutamine (Q) residues and adding two units of arginine (R) (a more positively charged residue) [76].
- To reduce potential cross linking or oxidation, Giuseppe *et. al* [77] excluded cysteine (C) and methionine residues in the synthesis phase.
- The use of lysine (K) instead of arginine (R) at the scissile site might impede tryptic digestion, because trypsin has up to 10-fold greater affinity for arginine than for lysine [29].
- The hydrophobicity of peptides can be increased by replacing all the valine (V) residues with isoleucine (I) or leucine (L) residues [29].
- Peptides that do not contain either non-natural or chemically modified amino acids, can be produced in a cost-effective manner in biological expression systems [29].
- It has been reported that replacing tryptophan (W) and phenylalanine (F) residues with less hydrophobic amino acids or interrupting the hydrophobic patches with basic amino acids decreases the toxicity of AMPs [29].
- The substitution of a threonine (T) residue on the middle position of the hydrophobic face with a tryptophan (W) residue, transforms an amphipathic peptide into a *P. aeruginosa*-targeted antimicrobial peptide [62].
- In the study of Gupta *et. al* it was observed that cysteine (C), histidine (H), asparagine (N) and proline (P) residues are abundant as well as preferred at various positions in toxic peptides [28].
- Cysteine (C) was preferred at almost all positions in toxic peptides according with the study of Gupta *et. al* [28]
- Proline (P), glycine (G), arginine (R) and serine (S) were found to be preferred at few positions at N-terminus, while valine (V), asparagine (N) and histidine (H) were preferred at few positions at C-terminus of non-toxic peptides [28].
- Methionine (M), leucine (L), phenylalanine (F) and isoleucine (I) were preferred at various positions at N-terminus, while leucine (L), glycine (G) and lysine (K) were preferred at various positions at C-terminus of non-toxic peptides [28].

- Composition of proline (P), asparagine (N) and histidine (H) was found to be higher in toxic peptides in comparison to non-toxic peptides [28].
- Resistance to proteolytic enzymes could be overcome by making peptides with only D-amino acids [78].
- Albada *et. al* have shown that systemic l-to-d exchange of amino acid could decrease the hemolytic potency of peptides without compromising their therapeutic activity [79].

Finally, it is important to mention that all the properties that are correlated with the antimicrobial activity, the toxicity and stability, should be considered together since the change in one of them, might alter other properties [76].

## 2.3 Current limitations in the rational design of antimicrobial peptides

In the rational design of antimicrobial peptides there are multiple limitations to deal with. Among them, the most important are:

- Lack of benchmark datasets: Due to the existence of multiple peptide databases that are growing day by day, there is not a benchmark dataset to work with. In the literature, most of the studies create their own dataset. Therefore, fair comparisons between different methods are difficult.
- Lack of a negative dataset: In most of the studies the negative datasets are build with random sequences from UniProt [80] that are label as non-AMP, non-membrane and non-secretory proteins. Although it is expected that these random sequences do not possess antimicrobial activity, experimental results are not available to probe this hypothesis.
- Few experimental structure data available: The 3-Dimensional structure has been obtained experimentally for only a few number of antimicrobial peptides. Therefore, it is difficult to obtain relationships between the structure and the antimicrobial activity of the peptides.
- Unanswered questions about AMPs: One of the major limitations in the rational design of antimicrobial peptides is the inability to describe their mechanism of action in physical-chemical terms and the lack of explicit, molecular, structure-function relationships [81].
- Models considering just one aspect: Most of the available models consider either the antimicrobial activity or the toxicity but not both. However, potential peptides should be very active with a low toxicity value, and the models should consider both aspects at the same time, as the improvement of the antimicrobial activity might lead to an increment of the toxicity or viceversa. Therefore, in this thesis we propose a strategy to generate potential antibacterial peptides with a low toxicity probability. Our models use peptide descriptors related with the primary structure of the peptides, considering the importance of the amino acid sequence in the final function of the peptide. Moreover, the models are based on classification performances and restrictions of the most important physicochemical properties according with the literature.

## 2.4 Overcoming one limitation

In this thesis, two strategies to design antibacterial peptides are proposed. The second strategy deals with the design of non-toxic antibacterial peptides. Therefore, this thesis helps to overcome one of the limitations in the rational design of antibacterial peptides.

The design of the proposed strategies involved different stages, as shown in Figure 2.1. The following chapters contain the details of these stages.

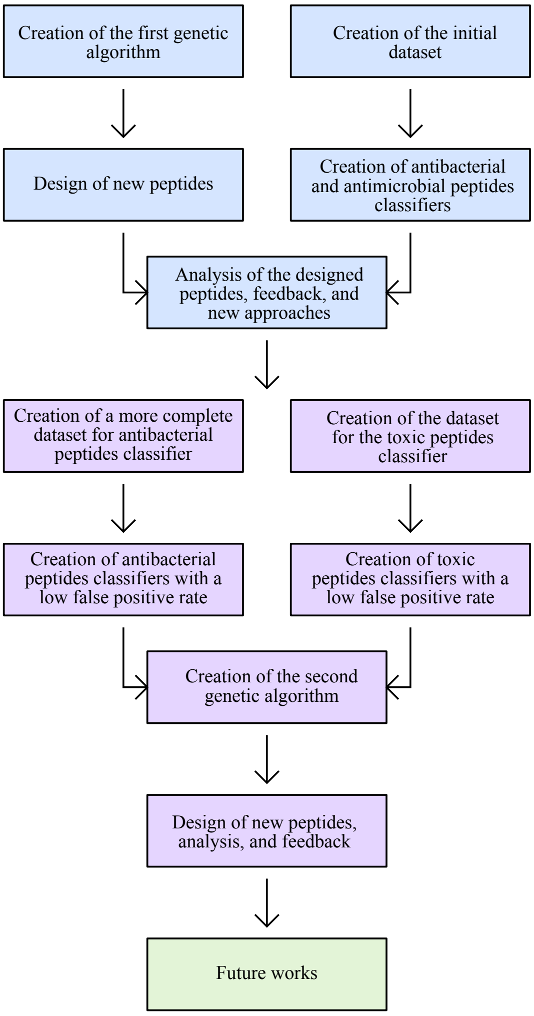

Creation of the first genetic algorithm
Creation of the initial dataset
Design of new peptides
Creation of antibacterial and antimicrobial peptides classifiers
Analysis of the designed peptides, feedback, and new approaches
Creation of a more complete dataset for antibacterial peptides classifier
Creation of the dataset for the toxic peptides classifier
Creation of antibacterial peptides classifiers with a low false positive rate
Creation of toxic peptides classifiers with a low false positive rate
Creation of the second genetic algorithm
Design of new peptides, analysis, and feedback
Future works

Figure 2.1: Stages performed in the design of the proposed strategies.

# Chapter 3

# Strategy 1 to design potential antibacterial peptides

The proposed strategy to design new antibacterial peptides by using genetic algorithms and classification methods is shown in Figure 3.1. The first step is to use the genetic algorithm to design a number $n$ of peptides using as input the sequences of the antibacterial peptides reported in the APD and CAMP databases. Later, the $n$ peptides should be tested in the antimicrobial peptides classifier and the peptides that are classified as antimicrobial peptides should be saved. Then, these peptides should be tested in the antibacterial peptides classifier and those that are classified as antibacterial peptides should be analyzed in the PEP-FOLD 2.0 tool [82], in order to select those peptides that possess a predicted alpha-helix structure. Finally, those peptides are candidate peptides to be synthesized.

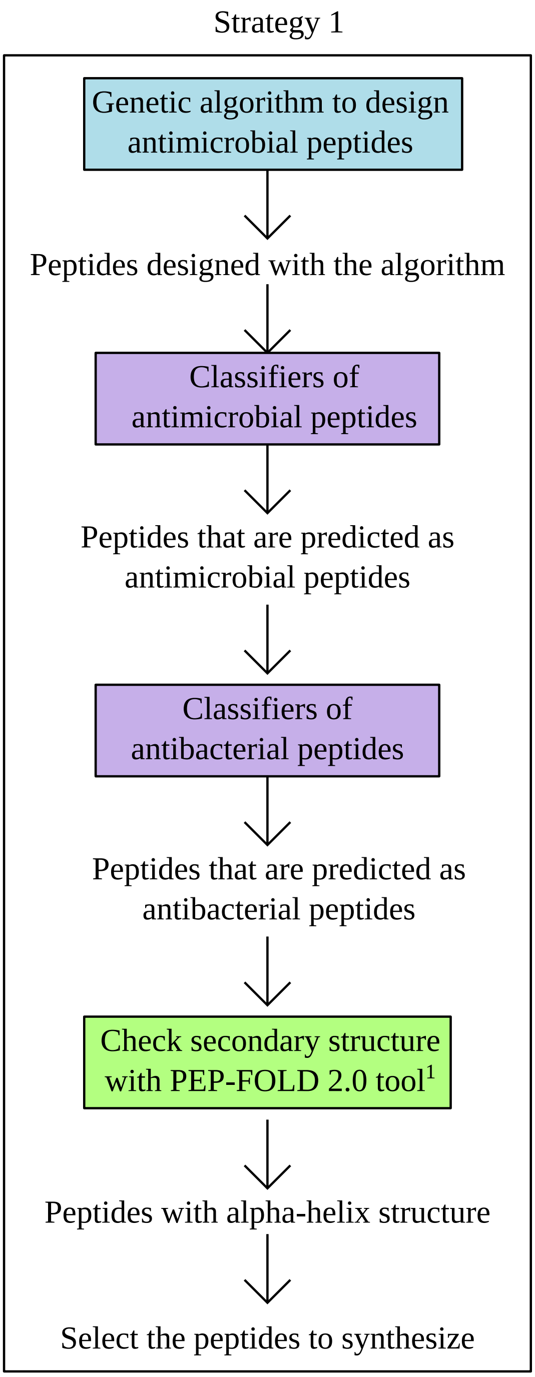

Strategy 1
Genetic algorithm to design antimicrobial peptides
Peptides designed with the algorithm
Classifiers of antimicrobial peptides
Peptides that are predicted as antimicrobial peptides
Classifiers of antibacterial peptides
Peptides that are predicted as antibacterial peptides
Check secondary structure with PEP-FOLD 2.0 tool[1]
Peptides with alpha-helix structure
Select the peptides to synthesize

Figure 3.1: Workflow of the proposed strategy 1.

## 3.1 Genetic algorithm DEPRAMPs 1.0

New antibacterial peptides could lead to the discovery of new antibiotics that have activity against multidrug-resistant bacteria. Therefore, strategies that allow the design of new potential antibacterial peptides should be created. In this strategy, we designed and developed a genetic algorithm (DEPRAMPs 1.0) that allows the generation of a desired number of peptides, of a given length, that satisfy the established ranges for each one of the physicochemical descriptors selected in this thesis: charge, hydrophobicity, isoelectric point, and instability index (See Table 3.1). These descriptors were selected considering their crucial role in the antibacterial activity, toxicity and stability of the peptides. The user can modify the default ranges. However, we recommend to use these ranges due to they were established considering the literature review, and they have given good results in the experimental tests performed at the GIBIM Lab.

Table 3.1: Suggested ranges for the physicochemical properties

| Property | Default range |
|---|---|
| Charge | $[2, 8]$ |
| Hydrophobicity | $[-1.5, 1.5]$ |
| Isoelectric point | $[7, 12]$ |
| Instability index | $< 40$ |

## 3.2 Overview of DEPRAMPs 1.0 algorithm

The optimization problem to solve is given by the Equation 3.1,

$$\begin{aligned} & Maximize\ fitness(p) \\ & \quad subject\ to: \\ & \quad h_1(p) : x_{1_{min}} < x_1(p) < x_{1_{max}} \\ & \quad h_2(p) : x_{2_{min}} < x_2(p) < x_{2_{max}} \\ & \quad h_3(p) : x_{3_{min}} < x_3(p) < x_{3_{max}} \\ & \quad h_4(p) : \qquad\quad x_4(p) < x_{4_{max}} \end{aligned} \tag{3.1}$$

where, $p$ is the amino acid sequence of a peptide, $fitness(p)$ is the fitness of the peptide $p$ and is given by the Equation 3.2, $x_1(p), x_2(p), x_3(p)$, and $x_4(p)$ are the charge, isoelectric point, hydrophobicity and instability index of the peptide $p$, respectively. The range of possible values for each of these physicochemical properties is given by the user, the minimum allowed value is $x_{i_{min}}$ and the maximum is $x_{i_{max}}$ where the value of $i = \{1, 2, 3, 4\}$ indicates the desired property.

$$\begin{aligned} fitness(p) &= (\textstyle\sum_{i=1}^{i=3} \left(p_{i_{low}}(p)^2 + p_{i_{up}}(p)^2\right) + p_4(p)^2) * -1 \\ p_{i_{low}}(p) &= \begin{cases} 0 & if \quad (x_{i_{min}} - x_i(p)) < 0 \\ \frac{x_{i_{min} - x_i(p)}}{x_{i_{min}}} & otherwise \end{cases} \\ p_{i_{up}}(p) &= \begin{cases} 0 & if \quad (x_i(p) - x_{i_{max}}) < 0 \\ \frac{x_i(p) - x_{i_{max}}}{x_{i_{max}}} & otherwise \end{cases} \\ p_4(p) &= \begin{cases} 0 & if \quad (x_4(p) - x_{4_{max}}) < 0 \\ \frac{x_4(p) - x_{4_{max}}}{x_{4_{max}}} & otherwise \end{cases} \end{aligned} \tag{3.2}$$

Figure 3.2 shows a graphical representation of the fitness function. The design of this function was based in the penalty methods that allow to transform a constrained optimization problem into an unconstrained optimization problem. Moreover, since the main objective is to design peptides that posses their physicochemical properties values inside a desired range of values, the designed function takes into account the distance that exists between the closest limit of the desired interval and the current property value, and assigns a proportional value to this distance. If the current value belongs to the desired interval, then a value of zero is assigned. Finally, the sum of the squared values obtained for each constraint is multiplied by -1 to allow the use of this function as the fitness value of the genetic algorithm. In this type of algorithms, the individuals that posses the higher fitness values are the ones that are going to survive in each generation. In this sense, the best fitness value for the designed function is equal to zero, i.e., the value for all the physicochemical properties complies with the desired constraints.

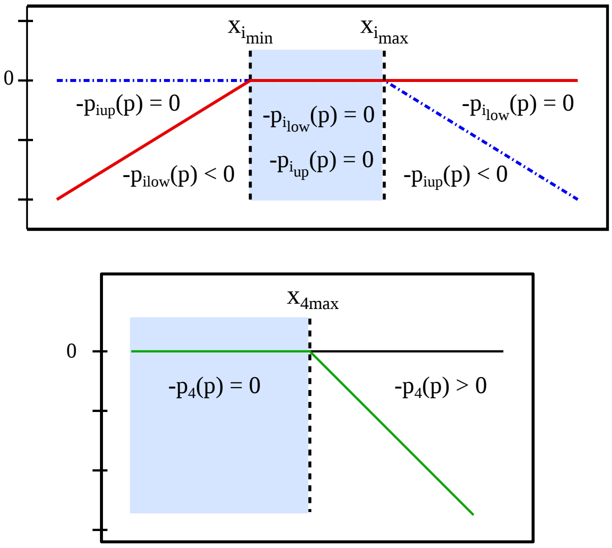


Figure 3.2: Graphical representation of the fitness function used in the genetic algorithm. Blue regions correspond to those values of $x_i$ that comply with the desired constraints. Therefore, the best value for the fitness function is equal to zero.

The workflow of the designed genetic algorithm is shown in Figure 3.3. The algorithm starts with the generation of the population, followed by the fitness evaluation for all the individuals. Later, the offspring is generated and a percentage of the population is replaced with the new individuals. The peptides with a low fitness value are the ones that are replaced by the offspring. In each iteration, the diversity of the population is measured and if its value is lower than the desired threshold, then a fixing strategy is used, and new individuals are created to replace the worse individuals in the current population. The stopping criteria is given by the number of generations or by the number of desired peptides. In each iteration, the peptides in the current population are evaluated in order to determine the number of candidate peptides in the current population. A candidate peptide is a peptide that complies with the constraints $h_1, h_2, h_3$, and $h_4$. If the number of candidate peptides is equal or higher than the desired number of candidate peptides ($desNum$), then the genetic algorithm stops and the candidate peptides are printed.

The parameters that should be set by the user are shown in Table 3.2 with their corresponding default values. The user can modify them, although some suggestions should be considered:

- The size of the peptide should be enough to guarantee antibacterial activity and stability of the peptide. According with Bahar and Ren [76], the length of *alpha*-helical peptides should be at least 22 amino acids in order to transverse the lipid bilayer of bacteria in the

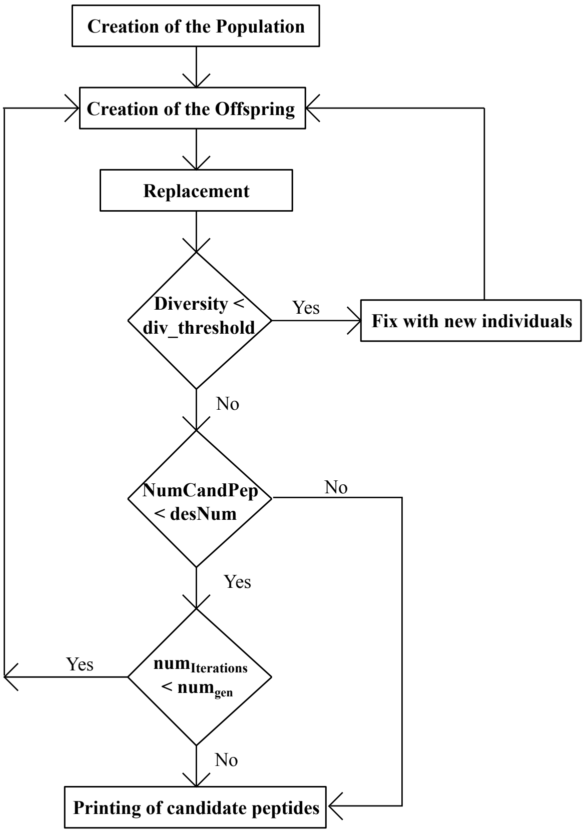

Creation of the Population
Creation of the Offspring
Replacement
Diversity < div_threshold
Yes
Fix with new individuals
No
NumCandPep < desNum
No
Yes
$num_{Iterations} < num_{gen}$
Yes
No
Printing of candidate peptides

Figure 3.3: Workflow of the genetic algorithm DEPRAMPs 1.0.

barrel-stave model. However, the antimicrobial peptide Lactoferrampin is active against multiple bacteria and its length is 17 amino acids. Since, the Lactoferrampin was the first peptide synthesized at the GIBIM Lab, we decided to perform the first experimental tests for peptides with 17 amino acids. Currently new experiments are being performed for peptides with 15 amino acids.

- It is recommended that the size of the population would around four times the number of candidate peptides in order to reduce the computational time of the simulation.
- The ranges of the physicochemical properties should not be changed if the user is new in the rational design of antimicrobial and antibacterial peptides. In that case, the default values are a good option.

### 3.2.1 Creation of the population

In this step, the genetic algorithm creates as many peptides, of the desired length, as indicated by the user ($desNum$). The procedure to create these peptides is illustrated in figure 3.4. All the antibacterial peptides reported in the APD and CAMP databases are joined in a unique sequence. Then, the genetic algorithm selects a random number between 1 and the desired length of the peptide minus two, i.e. $1 <= r <= desLength - 1$. With the value of $r$ the creation of the peptides starts. The first peptide will be created with the amino acids comprehend between the position $r + 1$ and the position $r + desLength$, then the next peptide will be created with the amino acids between the position $r + desLength + 1$ and the position $r + 2 * desLength$, and so on. If the end of the unique sequence is reached before the total number of peptides have been created, then, the procedure starts again with a different value of $r$. This procedure

Table 3.2: Parameters of the designed genetic algorithm

| Parameter | Default value |
|---|---|
| Size of the peptides ($desLength$) | 17 |
| Number of candidate peptides ($desNum$) | 100 |
| Size of the population ($sizePop$) | 500 |
| Number of generations ($num_{gen}$) | 1000 |
| Percentage of crossover ($per_{cross}$) | 0.8 |
| Percentage of mutation ($per_{mut}$) | 0.2 |
| Percentage of replacement in each generation ($per_{rep}$) | 0.7 |
| Percentage of replacement for diversity ($per_{rep_{div}}$) | 0.5 |
| Diversity threshold ($div_{threshold}$) | 0.6 |
| Desired antibacterial probability ($desProb_{abp}$) | 0.99 |
| Desired toxicity probability ($desProb_{tox}$) | 0.01 |
| Minimum value for charge ($x_{1_{min}}$) | 2 |
| Maximum value for charge ($x_{1_{max}}$) | 8 |
| Minimum value for hydrophobicity ($x_{2_{min}}$) | -1.5 |
| Maximum value for hydrophobicity ($x_{2_{max}}$) | 1.5 |
| Minimum value for isoelectric point ($x_{3_{min}}$) | 7 |
| Maximum value for isoelectric point ($x_{3_{max}}$) | 12 |
| Maximum value for instability index ($x_{4_{max}}$) | 40 |
| Hard constraints ($hardCons$) | No |

is repeated until all the peptides have been created.

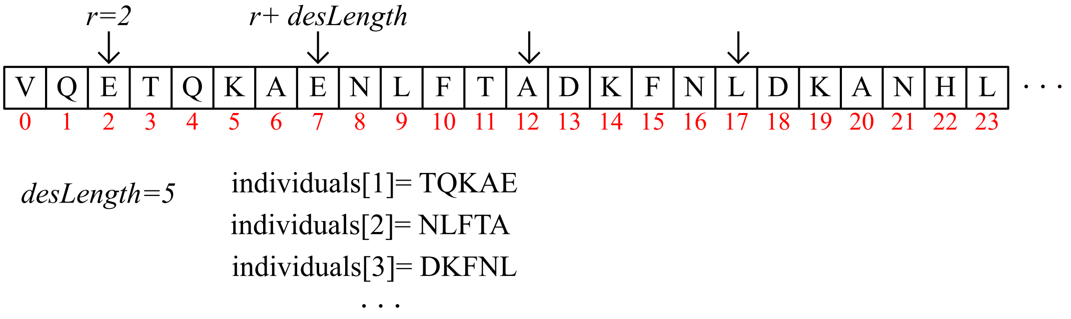


Figure 3.4: Creation of the initial population. In this figure the parameter $desLength$ was set to 5 just for graphical issues. The suggested value is 17 amino acids.

### 3.2.2 Creation of the offspring

In this step, the peptides that will replace the worse individuals in the population are created. The number of replacements is given by the size of the population ($sizePop$) and the percentage of replacement in each generation ($per_{rep}$).

The creation of the offspring is illustrated in Figure 3.5. The first step is the selection of the parents. Two parents are selected to create one child, hence, the number of selected parents is twice the number of desired children. The parents are selected by roulette selection.

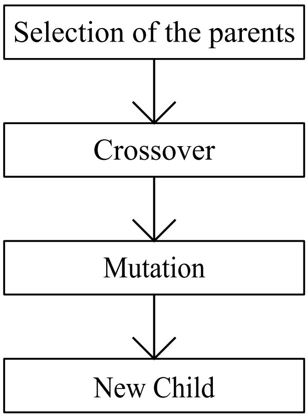


Figure 3.5: Steps in the creation of the offspring.

Once all the parents have been selected, the children are created. The first child is created with the first two parents, the second child with the following parents, and so on. For each child, a random number between 0 and 1 is created and if the number is lower than the percentage of crossover $per_{cross}$, then the crossover takes place. This procedure is shown in Figure 3.6. A random number $r_c$ between 1 and $desLength - 1$ is created. This number indicates the part that the child will get from each parent. The first part is from the first parent and the second part from the second parent. If there is no crossover, then the child is a copy of the first parent.

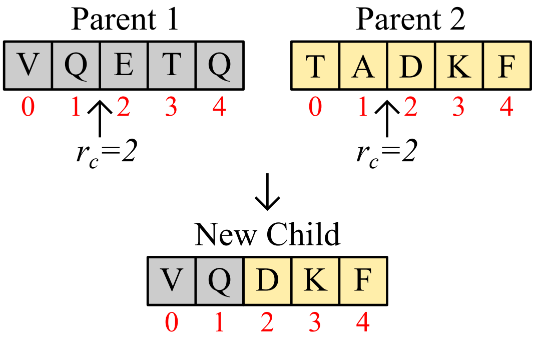


Figure 3.6: Graphical illustration of the crossover of two parents to create one child.

Once the child has been created, the mutation might take place. A random number $r_m$ between 0 and 1 is generated. If $r_m < per_{mut}$ then a random mutation of one amino acid is performed. The position of the mutation and the new amino acid are chosen randomly, without considering any physicochemical property.

### 3.2.3 Replacement

The designed genetic algorithm replaces a percentage of the population ($per_{rep}$) in each generation. The individuals to replace are those that have the worse fitness values. The genetic algorithm calculates the phenotypic diversity of the population in each new generation. This measure is obtained by the Equation 3.3, where $unique_{ind}$ is the number of unique peptides, and $total_{ind}$ is the total number of peptides in the population. If the diversity value is lower than the established threshold ($div_{threshold}$), then a correction action takes place, and new individuals are created to replace a percentage ($per_{rep_{div}}$) of the current population. The individuals that are replaced are those that have the lowest fitness values. The correction action is performed as many times as required until one of the two stop criterion is reached.

$$diversity = \frac{unique_{ind}}{total_{ind}} \tag{3.3}$$

### 3.2.4 Stop criteria

For the design of new antibacterial peptides a new stop criterion was designed. This criterion comprehends the number of candidate peptides that exist in the current population. If the number of unique candidate peptides is equal or higher than the number of desired peptides, then the algorithm stops and the candidate peptides are printed. Additionally, the stop criterion given by the number of generations was also implemented in this genetic algorithm.

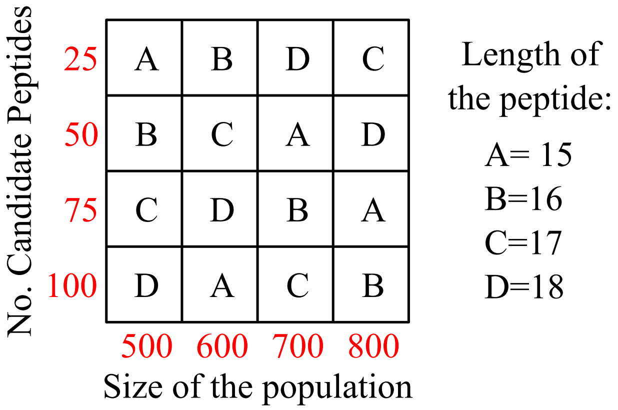
### 3.2.5 Printing of candidate peptides

At the end of the simulation, the candidate peptides are printed in two files. The first one corresponds to the fasta file of the sequences. The second file is .info and it contains the information of the designed peptides such as peptide sequence, charge, hydrophobicity, isoelectric point and instability index.

## 3.3 Simulations performed

In order to determine a suitable factor between number of candidate peptides and size of the population, we performed simulations with the number of candidate peptides set to 25, and the size of the population set to 1, 2, 3, 4, 6, and 8 times the number of candidate peptides. Five simulations were performed for each value. For these simulations, the number of generations was set to 2000.

Additionally, we performed multiple simulations of the genetic algorithm using different values for the parameters: size of the peptides ($desLength$), number of candidate peptides ($desNum$) and size of the population ($sizePop$). For this purpose, we used a Latin Square (see Figure 3.7) with these parameters and 16 different simulations were designed. Each of the simulations was run 3 times.



| No. Candidate Peptides | 500 | 600 | 700 | 800 |
|---|---|---|---|---|
| 25 | A | B | D | C |
| 50 | B | C | A | D |
| 75 | C | D | B | A |
| 100 | D | A | C | B |

Figure 3.7: Latin square of the performed simulations.

### 3.3.1 The genetic algorithm is quite fast if the correct size of the population is set

Table 3.3 shows the average time and number of generations that the genetic algorithm spent in the design of 25 candidate peptides when the size of the population was set to different values. The simulations for a population of size 25, 50 and 75 peptides were not satisfactory. The genetic algorithm was not able to design the desired amount of peptides and it stopped after 2000 generations.

Table 3.3: Average time and number of generations spent by the genetic algorithm when designing 25 peptides. The standard deviation is shown in parenthesis. The simulations were performed in a computer with 4 GB of RAM memory, and an Intel® Core$^{TM}$ i5-480M processor with 2 cores and 2.67GHz of frequency.

| Size population | Time in seconds | Number generations |
|---|---|---|
| 100 | 562.287 (107.758) | 102 (28) |
| 150 | 14.761 (0.343) | 1 (0) |
| 200 | 17.388 (2.027) | 1 (0) |

It is interesting that if the correct size of the population is used, the genetic algorithm only requires one generation to obtain the desired candidate peptides. Although the number of generations is the same in different runs, the computational time is different. This situation might be caused by the random nature of the crossover and mutation processes when the offspring is created.

### 3.3.2 The designed peptides comply with the established constraints

Figure 3.8 shows the boxplots of the charge, isoelectric point, hydrophobicity and instability index for the peptides that were designed with the genetic algorithm. The established ranges for these physicochemical properties were fulfilled for all the peptides. Hence, we can conclude that the genetic algorithm works satisfactorily and that the number of generations does not affect the obtained results, i.e., even with just one generation the genetic algorithm is able to design the desired new candidate peptides.

## 3.4 Classification of antimicrobial and antibacterial peptides by using kernel methods and peptide descriptors

The majority of studies related to the classification of antimicrobial and antibacterial peptides have used QSAR and machine learning techniques [47,68]. In this section, we proposed to use string kernel methods in conjunction with support vector machines to classify antimicrobial and antibacterial peptides. We believed that the order of the amino acids in peptide sequences might give enough information to classify them according to their biological activity. This assumption is based on the studies of Wang *et. al* [73], Chen and Luo [83], and Lata *et. al.* [72] that have obtained good results using the information of the amino acid composition, dipeptide composition and/or pseudo-amino acid composition. Therefore, the p-spectrum and mismatch kernels were used to perform the classification processes. Additionally, we performed simulations using local and global sequence alignment scores as kernel values. Finally, classifiers using peptide descriptors were also created. The obtained results showed that the order of the amino acids in

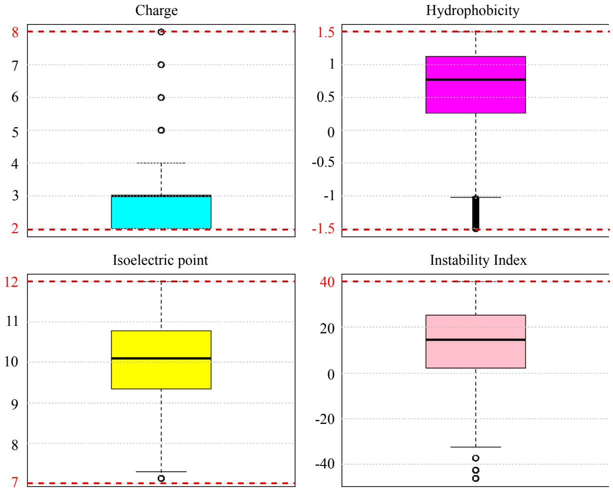


Figure 3.8: Boxplots of the physicochemical properties of the peptides design by the genetic algorithm.

a peptide is an important feature to take into account in the classification processes of antimicrobial and antibacterial peptides.

### 3.4.1 Creation of the working dataset

In the context of classification processes of antimicrobial and antibacterial peptides, the selection of the data can be done using datasets that appear in literature such as [67, 73, 84] or creating a new dataset with the information given in the international databases of antimicrobial peptides, as multiple studies have done [22, 68, 70, 72, 75, 83, 85]. The creation of the dataset of non-antimicrobial peptides presents some difficulties due to the absence of a non-antimicrobial peptides database. However, the most common strategy is to create this dataset with random sequences from UniProt [80] that are label as non-AMP, non-membrane and non-secretory proteins [73].

Building an antibacterial peptides classifier requires a training set of known antibacterial peptides and non-antibacterial peptides. The set of non-antibacterial peptides, however, consists of peptides that do not show activity against any microorganism and peptides that possess activity against microorganisms but bacteria. Hence (from a pattern recognition point of view) the negative class consists of two subtypes. Examples of classifiers that use the negative subtype of active peptides can be found in [86, 87] while examples of classifiers that use the negative subtype of non-active peptides are [72, 74, 75]. Due to the lack of a benchmark dataset for the classification of antibacterial peptides, different strategies have been used to build training sets. For this reason, a fair comparison between the antibacterial peptides classifiers found in the literature is difficult. However, reported accuracies vary between 85% and 95% [35,72,74,75,86].

For this part of the thesis, the positive dataset (further denoted as Abps) includes those peptides, from the APD database [36], that only reports activity against bacteria. The peptides that were active against bacteria and other microorganisms were not used to create the dataset. The total number of antibacterial peptides was 1,008. The negative dataset includes 192 peptides that do not report activity against bacteria (further denoted as non-Abps), and 10,014 peptides obtained by Wang *et. al* [73] from non-secretory proteins, and they are used as non-antimicrobial peptides. Figure 3.9 shows a graphical representation of the working dataset.

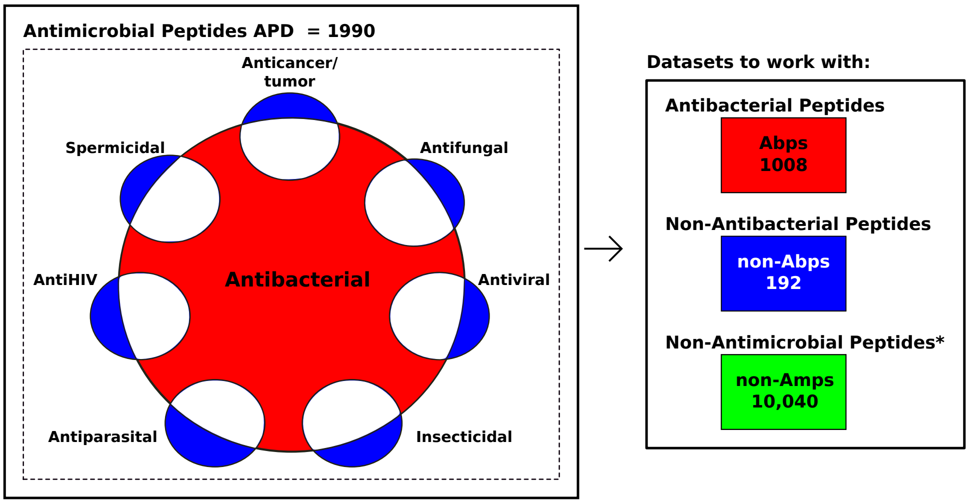


Figure 3.9: Graphical representation of the working dataset.

### 3.4.2 Common strategies to deal with the imbalanced problem

In the context of binary classification, the imbalanced problem appears in those applications where the number of samples in one class is much higher than the number of samples of the other class. In these situations, the classifiers will label the samples of the majority class with high precision while the samples of the minority class will be classified wrongly [88, 89].

In general, there are three approaches to deal with the imbalanced situation: internal, external and cost-sensitive learning. The internal approach comprehends those methods that modify the learning algorithm to consider the imbalanced problem, while the external refers to those methods that resample the initial dataset in order to obtain a more balanced situation. The cost-sensitive learning approach incorporates both internal and external level, i.e., resampling of the dataset and modifications of the learning algorithm [90].

In the case of external approaches, two categories can be proposed. The first one studies the best data to include them in the training set, and the second category concentrates in the study of the best proportion of positive and negative examples to include in the training set [89]. Some of the methods at the external level are random undersampling, random oversampling, synthetic minority oversampling technique, selective preprocessing of imbalanced data, among others [90].

The internal approaches have the disadvantage of being algorithm specific. This is problematic when it is desired to use a different algorithm considering that depending on the characteristics of the samples, some algorithms achieve better results than others. On the other hand, the main drawback of the cost-sensitive methods consists in the assignation of the misclassification costs, which are usually unavailable [90].

Finally, there is a new category that comprehends ensemble-based methods to deal with the imbalanced issue. These methods combine the ensemble learning techniques with one of the above approaches, specifically, external and cost-sensitive learning. Among the ensemble techniques that have been used to deal with the imbalanced problem are AdaCost, RareBoost, AdaC1, AdaC2, AdaC3, SMOTEBoost, RUSBoost, DataBoost-IM, overbagging, and underbagging, among others. For more information of these methods please refer to [90].

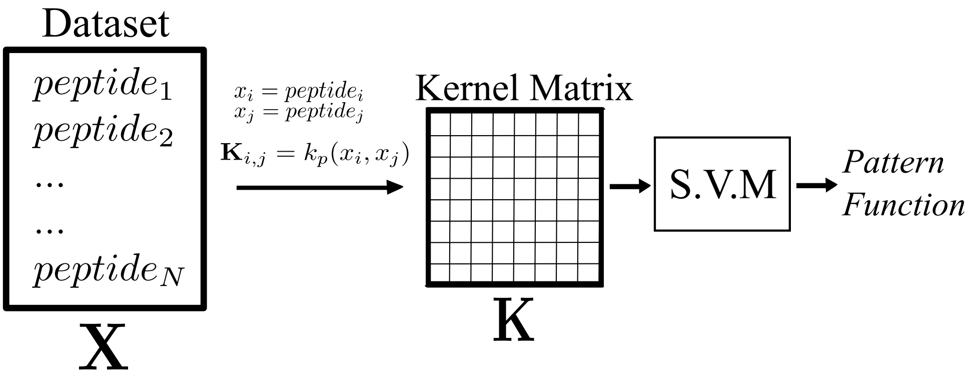


Figure 3.10: General diagram for the creation of an antimicrobial or antibacterial peptides classifier using kernel methods.

### 3.4.3 Kernel methods

In multiple situations, the classification problems present data that cannot be differentiated with linear relations. In these cases, the usage of kernel methods allows to perform a linear classification process by the mapping of the data into an $N$-dimensional space of order $N$ bigger than the order of the initial space. This can be done because the classes in this new space, called feature space, are usually linearly separable [91].

In the kernel methods, there are two important elements: the kernel function and the kernel matrix. The kernel function is defined by (3.4) [92]

$$k(\mathbf{x}, \mathbf{z}) = \langle \phi\left(\mathbf{x}\right), \phi\left(\mathbf{z}\right) \rangle . \tag{3.4}$$

where $\mathbf{x}, \mathbf{z}$ are elements of any set and their image $\phi(\mathbf{x})$ is a vector in $\mathbf{R}^N$. This kernel function is used to obtain the kernel matrix, defined in (3.5), that contains the inner product of all pairs of data points in the feature space [92].

$$\mathbf{K}_{i,j} = \langle \phi\left(\mathbf{x}_i\right), \phi\left(\boldsymbol{x}_j\right) \rangle = k\left(\mathbf{x}_i, \mathbf{x}_j\right) . \tag{3.5}$$

where $\mathbf{x}_i$, $\mathbf{x}_j$ are elements of any set and their image $\phi(\mathbf{x}_i)$, is a vector in $\mathbf{R}^N$ and $\mathbf{K}_{i,j}$ is the element in the row $i$ and column $j$ of the kernel matrix. It is important to mention that the kernel function calculates the inner product of the images of two elements in the feature space without explicitly computing the mapping of these elements.

The selection of the kernel function should consider the type of the input data, and the selection of the learning algorithm depends of the process that is required: classification, prediction or clustering. One of the algorithms used in classification processes is support vector machines [92], which was selected in this thesis as the learning algorithm used in the classification of antimicrobial and antibacterial peptides.

A graphical representation of the creation of an antimicrobial or antibacterial peptides classifier using kernel methods with support vector machines is shown in Figure 3.10. The first step consists in the creation of the kernel matrix, followed by the calculation of the pattern function through the application of the learning algorithm, which in this thesis is support vector machines.

### P-spectrum kernel

For a sequence $S$, its spectrum of order $p$ corresponds to the histogram of all their contiguous substrings of length $p$. The kernel based on this spectrum, allows the comparison between two sequences, calculating the number of substrings of length $p$ they have in common [92].

In this work the $p$-spectrum was calculated using an adaptation of the $p$-spectrum recursion algorithm defined in [92]. The algorithm was modified with the function *isEqual* and in the superior limit of the sum as shown in (3.6).

$$k_p(s,t) = \sum_{i=1}^{|s|-p+1} \sum_{j=1}^{|t|-p+1} isEqual(s(i:i+p-1), t(j:j+p-1)) \tag{3.6}$$

where $k_p(s,t)$ is the $p$-spectrum for sequences $s$ and $t$, $|s|$ is the length of the sequence $s$, $s(i:i+p-1)$ is the subsequence of $s$ that starts in position $i$ and ends in position $i+p-1$ and $isEqual(a,b)$ is the function defined by (3.7)

$$isEqual(a,b) = \begin{cases} 1 \quad if & a=b \\ 0 & otherwise \end{cases} \tag{3.7}$$

### Mismatch kernel

The mismatch kernel is very useful in bioinformatics applications because it allows some degree of mismatching between the subsequences that two sequences share. The mismatch kernel $K_{(k,m)}(s,t)$ allows the comparison between sequences $s$ and $t$, calculating the number of $k$-length substrings, that strings $s$ and $t$ share, differing by at most $m$ mismatches. The following definitions are used in the formal definition of the mismatch kernel [93].

- $k$-mer: subsequences of length $k$.
- $A$: alphabet of the sequences of size $|A| = l$. In our case, the alphabet are the common amino acids, therefore $l = 20$.
- $(k,m)(\alpha)-$ neighborhood: It is composed by the set of all $k$-length subsequences $\beta$ from $A$ that differ from $\alpha$ by at most $m$ mismatches. It is denoted as $N_{(k,m)}(\alpha)$.
- $\Phi_{(k,m)}(\alpha)-$ Feature map given by the equation 3.8: if $\alpha$ is a $k-mer$,

  $$\Phi_{(k,m)}(\alpha) = (\phi_\beta(\alpha))\beta \in A^k \tag{3.8}$$

  where $\phi_\beta(\alpha) = 1$ if $\beta$ belongs to $N_{(k,m)}(\alpha)$, and $\phi_\beta(\alpha) = 0$ otherwise.
- For a sequence $s$, the map comprehends the summation of the feature vectors for all the $k$-mers in $s$:

  $$\Phi_{(k,m)}(s) = \sum_{k-mers\alpha ins} \Phi_{(k,m)}(\alpha) \tag{3.9}$$

Finally, the mismatch kernel $K_{(k,m)}(s,t)$ is the inner product in feature space of the feature vectors of the sequences $s$ and $t$, as given in the Equation 3.10

$$K_{(k,m)}(s,t) = \langle \Phi_{(k,m)}(s), \Phi_{(k,m)}(t) \rangle \tag{3.10}$$

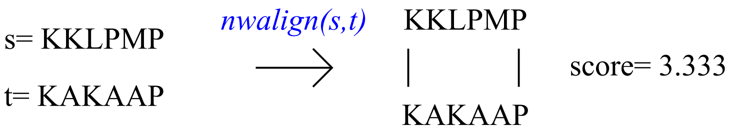


Figure 3.11: Alignment and score of sequences $s$ and $t$ using the Needleman-Wunsch algorithm.

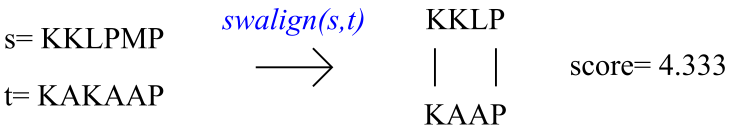


Figure 3.12: Alignment and score of sequences $s$ and $t$ using the Smith-Waterman algorithm.

### Needleman-Wunsch algorithm - Global alignment

The Needleman-Wunsch algorithm was developed to find similarities in the amino acid sequences of two proteins. This algorithm finds the maximum match, which may represent the largest number of amino acids of one protein that can be matched with the amino acids of another protein, allowing all possible interruptions in either of the protein sequences, or it may be a value that is a complex function of the relationship between the sequences [94].

The Matlab function *nwalign* [95] allows to perform a global alignment using the Needleman-Wunsch algorithm. Among the outputs of this function are the score and the alignment. For example, the score and alignment for the sequences $s$ =KKLPMP and $t$ =KAKAAP are shown in Figure 3.11.

In this thesis the score given by the global alignment between sequences $s$ and $t$, is used as the value of a kernel function $K_{global}(s,t)$.

### Smith-Waterman algorithm - Local alignment

The Smith-Waterman algorithm is used to find the pair of segments with maximum similarity between two sequences. An important aspect of this algorithm is that the similarity measure given by the algorithm allows deletions and insertions of arbitrary length [96].

The Matlab function *swalign* [95] is used to perform local alignments using the Smith-Waterman algorithm. In this thesis, the score given by this function is used as the value of a kernel function $K_{local}(s,t)$. Figure 3.12 shows the score and alignment obtained with this function for the sequences $s$ =KKLPMP and $t$ =KAKAAP.

## 3.4.4 Peptide descriptors

In the machine learning context, a peptide is usually represented by an $n$-dimensional vector, where $n$ is the number of peptide descriptors, also known as features, calculated for this chemical compound. In this thesis, two categories of feature sets were used. The first category includes features sets that contain physicochemical information of the peptides ($f_0$ - $f_{11}$). The second category includes features sets that consider the amino acid content and the order of the amino acids in the peptide sequence ($f_{12}$ - $f_{16}$). Table 3.4 shows the feature sets used and the tool that is used to obtain them.

Table 3.4: Feature sets and tools used in the creation of the classifiers.

| Feat. Set | No. of Features | Description | Tool |
|---|---|---|---|
| $f_0$ | 240 | Normalized Moreau-Broto autocorrelation | Propy [97] |
| $f_1$ | 240 | Moran autocorrelation | Propy [97] |
| $f_2$ | 240 | Geary autocorrelation | Propy [97] |
| $f_3$ | 147 | Composition, transition, distribution descriptors | Propy [97] |
| $f_4$ | 90 | Sequence order coupling numbers | Propy [97] |
| $f_5$ | 50 | Quasi-sequence-order descriptors | Propy [97] |
| $f_6$ | 21 | Hydrophobicity | SPiCE [98] |
| $f_7$ | 21 | Normalized VDW | SPiCE [98] |
| $f_8$ | 21 | Polarity | SPiCE [98] |
| $f_9$ | 21 | Polarizability | SPiCE [98] |
| $f_{10}$ | 21 | Charge | SPiCE [98] |
| $f_{11}$ | 19 | Autocorrelation | SPiCE [98] |
| $f_{12}$ | 20 | Amino acid composition | Propy [97] |
| $f_{13}$ | 400 | Dipeptide composition | Propy [97] |
| $f_{14}$ | 20 | N-terminal end amino acid count | SPiCE [98] |
| $f_{15}$ | 20 | C-terminal end amino acid count | SPiCE [98] |
| $f_{16}$ | 40 | Composition moment vector | Own code [99] |

### 3.4.5 Simulations performed

Ten different classifiers were created using the mentioned kernels, algorithms and peptide descriptors. Five classifiers allow the discrimination between antimicrobial and non-antimicrobial peptides, and the other five classifiers deals with the discrimination between antibacterial and non-antibacterial peptides. A 10-fold cross-validation process was performed for all the classifiers, i.e., all the classifiers were trained and tested using the same folds. This methodology allows a fair comparison between the classifiers.

For the antimicrobial peptides classifiers, a random subsampling of 400 antimicrobial peptides and 400 non-antimicrobial peptides was performed. In this sense, the initial dataset is composed by 800 peptides. Similarly, for the antibacterial peptides classifier, a random subsampling of 192 antibacterial peptides was performed. The final dataset for these classifiers was composed by 192 antibacterial and 192 non-antibacterial peptides.

We used support vector machines as the learning algorithm, using six different values for the penalization parameter $C = \{0.001, 0.01, 0.1, 1, 10, 100\}$. For the classifiers created using the Needleman-Wunsch and Smith-Waterman algorithms, the values for gap-open and gap-extension were set to 8 and the scoring matrix was BLOSUM50. The performance measures were sensitivity, specificity, false positive rate, false negative rate and precision.

### 3.4.6 Results and Discussion

#### Antimicrobial and antibacterial peptides classifiers created using string kernels are competitive

Tables 3.5 and 3.6 show the values for the sensitivity, specificity, false positive rate, false negative rate and accuracy, obtained with the different classifiers. In Table 3.5 the bold values are the best significant values according with ANOVA and Tukey tests (p_value $< 0.05$). In table 3.6 the red values are the worse significant values according with ANOVA and Tukey tests, i.e., the black color values shown in the table are the best significant values.

Table 3.5: Sensitivity, specificity, false positive rate, false negative rate and accuracy for the antimicrobial peptides classifiers created using peptide descriptors and kernel methods. The learning algorithm used is support vector machines with different $c$ values

| | Classifier | c=0.001 | c= 0.01 | c=0.1 | c=1 | c=10 | c=100 |
|---|---|---|---|---|---|---|---|
| Sen | Descriptors | 0.845(0.098) | 0.850(0.096) | 0.850(0.096) | 0.850(0.096) | 0.850(0.096) | 0.850(0.096) |
| | P-spectrum | **1.000(0.000)** | 0.823(0.062) | 0.820(0.057) | 0.793(0.070) | 0.790(0.069) | 0.790(0.069) |
| | Mismatch | 0.858(0.058) | 0.845(0.054) | 0.795(0.055) | 0.798(0.049) | 0.765(0.053) | 0.783(0.069) |
| | Swalign | 0.750(0.078) | 0.863(0.044) | 0.840(0.065) | 0.795(0.061) | 0.730(0.067) | 0.738(0.065) |
| | Nwalign | 0.825(0.041) | 0.855(0.052) | 0.848(0.064) | 0.765(0.063) | 0.755(0.087) | 0.753(0.075) |
| Sp | Descriptors | **0.868(0.062)** | **0.850(0.070)** | **0.850(0.070)** | **0.850(0.070)** | **0.850(0.070)** | **0.850(0.070)** |
| | P-spectrum | 0.000(0.000) | **0.918(0.070)** | **0.888(0.060)** | **0.860(0.053)** | **0.860(0.053)** | **0.860(0.053)** |
| | Mismatch | **0.885(0.047)** | **0.855(0.061)** | 0.813(0.059) | 0.765(0.080) | 0.778(0.067) | 0.783(0.069) |
| | Swalign | **0.958(0.039)** | **0.903(0.042)** | 0.848(0.063) | 0.808(0.076) | 0.710(0.065) | 0.703(0.079) |
| | Nwalign | **0.855(0.056)** | **0.855(0.056)** | 0.810(0.064) | 0.765(0.071) | 0.750(0.080) | 0.750(0.080) |
| Fpr | Descriptors | **0.133(0.062)** | **0.150(0.070)** | **0.150(0.070)** | **0.150(0.070)** | **0.150(0.070)** | **0.150(0.070)** |
| | P-spectrum | 1.000(0.000) | **0.083(0.070)** | **0.113(0.060)** | **0.140(0.053)** | **0.140(0.053)** | **0.140(0.053)** |
| | Mismatch | **0.115(0.047)** | **0.145(0.061)** | 0.188(0.059) | 0.235(0.080) | 0.223(0.067) | 0.218(0.069) |
| | Swalign | **0.043(0.039)** | **0.098(0.042)** | 0.153(0.063) | 0.193(0.076) | 0.290(0.065) | 0.298(0.079) |
| | Nwalign | **0.145(0.056)** | **0.115(0.056)** | 0.190(0.064) | 0.235(0.071) | 0.250(0.080) | 0.250(0.080) |
| Fnr | Descriptors | 0.155(0.098) | 0.150(0.096) | 0.150(0.096) | 0.150(0.096) | 0.150(0.096) | 0.150(0.096) |
| | P-spectrum | **0.000(0.000)** | 0.178(0.062) | 0.180(0.057) | 0.208(0.070) | 0.210(0.069) | 0.210(0.069) |
| | Mismatch | 0.143(0.058) | 0.155(0.054) | 0.205(0.055) | 0.203(0.049) | 0.235(0.053) | 0.213(0.050) |
| | Swalign | 0.250(0.078) | 0.138(0.044) | 0.160(0.065) | 0.205(0.061) | 0.270(0.067) | 0.263(0.065) |
| | Nwalign | 0.175(0.041) | 0.145(0.052) | 0.153(0.064) | 0.235(0.063) | 0.245(0.087) | 0.248(0.075) |
| Acc | Descriptors | **0.856(0.055)** | **0.850(0.061)** | **0.850(0.061)** | **0.850(0.061)** | **0.850(0.061)** | **0.850(0.061)** |
| | P-spectrum | 0.500(0.000) | **0.870(0.045)** | **0.854(0.043)** | **0.826(0.040)** | **0.825(0.039)** | **0.825(0.039)** |
| | Mismatch | **0.871(0.029)** | **0.850(0.038)** | 0.804(0.035) | 0.781(0.036) | 0.771(0.032) | 0.785(0.037) |
| | Swalign | **0.854(0.049)** | **0.883(0.033)** | **0.844(0.032)** | 0.801(0.053) | 0.720(0.053) | 0.720(0.051) |
| | Nwalign | **0.840(0.024)** | **0.870(0.048)** | **0.829(0.047)** | 0.765(0.039) | 0.753(0.051) | 0.751(0.052) |

In general, the classifiers created using string kernel methods have obtained a similar behavior that the one obtained using peptide descriptors. In the case of antimicrobial peptides classifiers there are some special situations. For example, for the sensitivity and false negative rate values, the best option is the one obtained with the p-spectrum kernel with $c = 0.001$. However, this classifier obtained the worse values for specificity and false positive rate. Therefore, a more accurate measure for the general performance of these classifiers is the accuracy, due to these classifiers were created using balanced datasets. In this sense, the classifier created with peptide descriptors obtained the best performance regarding the value for $c$. In the case of the string kernel methods, the value of $c$ has an influence in the performance obtained for the antimicrobial peptides classifiers. However, all the classifiers obtained a competitive performance when $c = 0.01$. On the contrary, in the antibacterial peptides classifiers the $c$ value only affects the classifiers created using the p-spectrum and mismatch kernels.

These results are interesting since the classifiers created using string kernels do not include any physicochemical information of the peptides. Therefore, it is feasible that the order of the amino acids inside the peptide sequences gives enough information to determine the presence or absence of any antimicrobial activity.

In addition, the obtained results using the 10-fold cross-validation technique show that the classifiers created are statistically stable due to the small variation that they present in the performance measures when the training dataset is changed. Moreover, there is not a significant difference in their performances and any of them could be use in the classification of antimicrobial and antibacterial peptides.

Table 3.6: Sensitivity, specificity, false positive rate, false negative rate and accuracy for the antibacterial peptides classifiers created using peptide descriptors and kernel methods. The learning algorithm used is support vector machines with different *c* values

| | Classifier | c=0.001 | c= 0.01 | c=0.1 | c=1 | c=10 | c=100 |
|---|---|---|---|---|---|---|---|
| Sen | Descriptors | 0.776(0.099) | 0.761(0.131) | 0.761(0.131) | 0.761(0.131) | 0.761(0.131) | 0.761(0.131) |
| | P-spectrum | 0.800(0.396) | 0.881(0.069) | 0.886(0.048) | 0.854(0.084) | 0.854(0.084) | 0.854(0.084) |
| | Mismatch | 0.808(0.066) | 0.797(0.055) | 0.755(0.092) | 0.741(0.096) | 0.756(0.082) | 0.730(0.104) |
| | Swalign | 0.855(0.080) | 0.897(0.072) | 0.828(0.071) | 0.765(0.077) | 0.765(0.077) | 0.765(0.077) |
| | Nwalign | 0.834(0.071) | 0.891(0.046) | 0.817(0.137) | 0.807(0.100) | 0.792(0.105) | 0.792(0.105) |
| Sp | Descriptors | 0.820(0.117) | 0.793(0.087) | 0.793(0.087) | 0.793(0.087) | 0.793(0.087) | 0.793(0.087) |
| | P-spectrum | 0.264(0.274) | 0.786(0.073) | 0.791(0.080) | 0.801(0.087) | 0.801(0.087) | 0.801(0.087) |
| | Mismatch | 0.771(0.148) | 0.757(0.138) | 0.740(0.097) | 0.724(0.075) | 0.708(0.101) | 0.698(0.096) |
| | Swalign | 0.734(0.067) | 0.812(0.045) | 0.828(0.079) | 0.812(0.062) | 0.812(0.062) | 0.812(0.062) |
| | Nwalign | 0.781(0.049) | 0.813(0.049) | 0.765(0.094) | 0.771(0.079) | 0.755(0.078) | 0.755(0.078) |
| Fpr | Descriptors | 0.180(0.117) | 0.207(0.087) | 0.207(0.087) | 0.207(0.087) | 0.207(0.087) | 0.207(0.087) |
| | P-spectrum | 0.736(0.274) | 0.214(0.073) | 0.209(0.080) | 0.199(0.087) | 0.199(0.087) | 0.199(0.087) |
| | Mismatch | 0.229(0.148) | 0.243(0.138) | 0.260(0.097) | 0.276(0.075) | 0.292(0.101) | 0.302(0.096) |
| | Swalign | 0.266(0.067) | 0.188(0.045) | 0.172(0.079) | 0.188(0.062) | 0.188(0.062) | 0.188(0.062) |
| | Nwalign | 0.219(0.049) | 0.187(0.049) | 0.235(0.094) | 0.229(0.079) | 0.245(0.078) | 0.245(0.078) |
| Fnr | Descriptors | 0.224(0.099) | 0.239(0.131) | 0.239(0.131) | 0.239(0.131) | 0.239(0.131) | 0.239(0.131) |
| | P-spectrum | 0.200(0.396) | 0.119(0.069) | 0.114(0.048) | 0.146(0.084) | 0.146(0.084) | 0.146(0.084) |
| | Mismatch | 0.192(0.066) | 0.203(0.055) | 0.245(0.092) | 0.259(0.096) | 0.244(0.082) | 0.270(0.104) |
| | Swalign | 0.145(0.080) | 0.103(0.072) | 0.172(0.071) | 0.235(0.077) | 0.235(0.077) | 0.235(0.077) |
| | Nwalign | 0.166(0.071) | 0.109(0.046) | 0.183(0.137) | 0.193(0.100) | 0.208(0.105) | 0.208(0.105) |
| Acc | Descriptors | 0.797(0.044) | 0.776(0.052) | 0.776(0.052) | 0.776(0.052) | 0.776(0.052) | 0.776(0.052) |
| | P-spectrum | 0.535(0.183) | 0.833(0.062) | 0.838(0.050) | 0.828(0.056) | 0.828(0.056) | 0.828(0.056) |
| | Mismatch | 0.790(0.095) | 0.777(0.079) | 0.748(0.046) | 0.732(0.059) | 0.732(0.066) | 0.713(0.068) |
| | Swalign | 0.794(0.063) | 0.854(0.054) | 0.828(0.054) | 0.789(0.046) | 0.789(0.046) | 0.789(0.046) |
| | Nwalign | 0.807(0.048) | 0.852(0.039) | 0.792(0.076) | 0.789(0.076) | 0.773(0.080) | 0.773(0.080) |

## 3.5 Posterior analysis of the peptides

Once the list of the designed candidate peptides was obtained, we performed an additional analysis of the peptides before their synthesis. The first step consisted in the prediction of the secondary structure of the peptides using the PEP-FOLD 2.0 tool [82] (http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::PEP-FOLD). Those peptides that exhibited an alpha-helix secondary structure were selected due to they are likely to be antibacterial peptides. Moreover, it is recommended to select those peptides that do not present an homology higher than 50% with the antimicrobial peptides reported in the APD and CAMP database. This last step is to allow the possibility that the new peptide could be patented in case that the experimental results are satisfactory. Therefore, the selected peptides to be synthesized were those that fulfilled the previous requirements.

## 3.6 Experimental results

Multiple experimental tests were performed with peptides designed in this thesis and with analogues of these peptides. Due to limitations in the available resources the experimental tests were performed for approximately 20 candidate peptides. The synthesis of these peptides was performed via solid phase peptide synthesis (SPPS) [100] using the tea-bag procedure reported in [101], in accordance with standard Fmoc chemistry and with a 0.63 substituted rink amide 4MBHA resin and Fmoc amino acids [102]. These peptides were cleaved with a mix of trifluoroacetic acid (TFA)/triisopropylsilan (TIS)/ethanedithiol/$H_2O$ (92.5/2.5/2.5:2.5) for 2 hours and then retrieved by precipitation with cold diethyl ether [103]. Then, the peptides were desalted by gel exclusion chromatography using G-10 columns (Amersham, USA). Finally, the peptides were purified by Reverse Phase-High Performance Liquid Chromatography (RP-HPLC) in a Vydac C-18 preparative column, using as mobile phase the mixture of: (A) $H_2O$ + 0.01% TFA and (B) acetonitrile + 0.01% TFA.

Antibacterial activity of the peptides was carried out as described in [104]. To test the bactericidal activity of the peptides, samples of approximately $4.6x10^8$ CFU (colony-forming unit) per ml of *E. Coli* O157: H7, MRSA and *P. aeruginosa* were incubated in 96-well microplates containing lysogeny broth (LB) for gram-negative bacteria or Mueller-Hinton (MH) for gram-positive bacteria, and peptide solutions at nal concentration between a range of 0.5 and 100 $\mu$M of peptide. Ofloxacin was used as growth inhibition control. Bacterial cells were incubated at 37C with constant agitation for 1 hour and monitored by enumerating viable cells using growth kinetics. Bacterial growth was monitored by turbidimetry quantifying optical density (OD) at $\lambda$= 595 nm for 8 hours.

The results were highly satisfactory, the majority of the peptides were active against bacteria, and two of them (GIBIM-P6 and GIBIM-P5F8W) possess a competitive antibacterial activity, i.e, peptide GIBIM-P6 shows a minimal inhibitory concentration (MIC) of 5.0, 10 and 10 $\mu M$ for *Escherichia coli* O157:H7, methicillin-resistant *Staphylococcus aureus*, and *Pseudomona aeruginosa*, respectively. The peptide GIBIM-P5F8W shows a MIC of 0.5, 5.0 and 10 $\mu M$ for *Escherichia coli* O157:H7, methicillin-resistant *Staphylococcus aureus*, and *Pseudomona aeruginosa*, respectively. Moreover, this peptide at 25 $\mu M$ provoked membrane permeabilization of these bacteria as determined by the Sytox Green uptake assay. Table 3.7 shows the $MIC_{99}$ values for the designed peptides, while Table 3.8 shows the $MIC_{90}$ values for existing antimicrobial agents against *Escherichia coli*, *Staphylococcus aureus*, and *Pseudomona aeruginosa*. It can be observed that two of our peptides are highly competitive considering that their $MIC_{99}$ values are lower or equal to $MIC_{90}$ values of commercial antimicrobial agents against non-resistant strains. Moreover, it is expected that the $MIC_{99}$ values of the antimicrobial agents will be higher than their $MIC_{90}$ values, i.e., it is expected that a higher concentration of the antimicrobial agent would be required to inhibit the growth of 99% of bacteria, instead of the 90% of bacteria.

Table 3.7: $\text{MIC}_{99}$ values of our candidate peptides. Data given by Jennifer Cruz and Yuly Prada, students of the GIBIM Lab.

| Peptide | *E. coli* | MRSA | *P. aeruginosa* |
|---|---|---|---|
| GIBIM-P5F8W | 0.5 | 5 | 10 |
| GIBIM-P6 | 5 | 10 | 10 |
| GIBIM-JC1 | 10 | 25 | 10 |
| GIBIM-JC2 | 50 | 25 | >100 |
| GIBIM-JC3 | 50 | 25 | >100 |
| GIBIM-JC4 | 50 | 50 | 100 |
| GIBIM-JC5 | 50 | 75 | >100 |
| GIBIM-JC6 | >100 | 50 | >100 |
| GIBIM-JC7 | 100 | 75 | 75 |
| GIBIM-JC8 | 100 | 75 | >100 |
| GIBIM-JC9 | 75 | >100 | 100 |
| GIBIM-JC10 | 100 | 100 | 10 |
| GIBIM-JC11 | >100 | 50 | 25 |
| GIBIM-JC12 | 100 | 100 | 10 |
| GIBIM-JC13 | 50 | 75 | 25 |
| GIBIM-JC14 | 50 | 75 | 25 |
| GIBIM-JC15 | 75 | 100 | 50 |
| GIBIM-JC16 | 75 | 100 | 50 |
| GIBIM-YP1 | 50 | 25 | Not tested |

The cytotoxicity of the peptides in mammalian cells is obtained by evaluating their hemolytic activity on sheep erythrocytes. One mL of erythrocytes from defibrinated sheep blood were subjected to centrifugation (1000 x g, 10 min) and subsequently washed with Hanks + Glucose. The erythrocytes were resuspended in this medium at a density of 2x107 cells/mL, aliquots of 100 $\mu$L of this suspension were transferred to eppendorf tubes and incubated with peptide at the desired concentrations (37 C, 4 h). Erythrocytes were subsequently centrifuged (13,000 x g, 5 min) and aliquots of the supernatant (80 $\mu$L) were transferred to a 96 well plate to measure hemoglobin released at 550 nm in a plate reader. All peptides showed a low toxicity, with a percentage of less than 40% hemolysis at a concentration of 50 $\mu$M. Additionally, the cytotoxicity of GIBIM-P6 and GIBIM-P5F8W peptides was assessed in lung tissue cell line A549 to determine its selectivity in a range of 25 to 100 mM. Both peptides exhibited a low cytotoxicity on A549 cells.

The experimental tests were performed by the PhD student Jennifer Cruz and the master student Yuly Prada, both members of the GIBIM Lab from the School of Chemistry at Universidad Industrial de Santander.

## 3.7 Conclusions

The algorithm DEPRAMPs 1.0 might be useful in multiple applications that require the design of peptides that comply with the established ranges for charge, isoelectric point, hydrophobicity and instability index. Moreover, the input sequences that are used in this algorithm might lead to the desire function of the designed peptides. For example, if we want to design antibacterial peptides, then the input sequences should be those that have been reported as antibacterial.

Table 3.8: MIC$_{90}$ values of existing antimicrobial agents. Data taken from [105]

| Agent | *E. coli* | *S. aureus* | *P. aeruginosa* |
|---|---|---|---|
| Ofloxacin[a] | 0.12 | 16[b] | 128 |
| Omiganan | 32 | 16 | 256 |
| Bacitracin | – | 400 | – |
| Erythromicyn | – | >8 | – |
| Fusidic acid | – | 0.25 | – |
| Gentamicin | 2 | 0.5 | 8 |
| Levofloxacin | – | >8 | – |
| Mupirocin | – | ≤4 | – |
| Neomycin | 4 | >16 | 128 |
| Oxacillin | – | >2 | – |
| TAO | <1.2 | 20 | <1.2 |
| Vancomycin | – | 1 | |
| Ceftazidime | >16 | – | >16 |
| Ceftriaxone | >32 | – | >32 |
| Ciprofloxacin | >8 | – | >8 |
| Imipenem | 0.25 | – | 16 |
| Polymyxin B | 0.5 | – | 0.5 |

[a] MIC values taken from http://antibiotics.toku-e.com/antimicrobial_868_1.html
[b] MIC value against MRSA.

It is possible to perform classification of antimicrobial and antibacterial peptides using only the information given by the string kernels. However, the performance achieved with these classifiers is not significantly better than the performance obtained when peptide descriptors are used.

Although the peptides that have been designed with this strategy have shown satisfactory results, there are some improvements that could be performed:

- Due to the subsampling performed in the antibacterial peptides set, it is possible that we are losing important information of patterns responsible for the antibacterial activity in peptides. Therefore, it is recommended to work with the entire dataset of antibacterial peptides.

- In this strategy, the peptides that possess antibacterial activity in conjunction with other biological activity were not used. These peptides might be very useful in the pharmaceutical industry, since they could become new antibiotics and they could become a new medicine for other diseases. In this sense, it is highly recommended to use these peptides in the design of new antibacterial peptides.

- The time required to design new potential antibacterial peptides might be reduce if the classifiers are involved during the design process. With this strategy, multiple peptides designed by the genetic algorithm are rejected in the following steps, and multiple simulations and trials have to be performed in order to obtain a large list of new potential antibacterial peptides.

- The toxicity part should be included in the design of new antibacterial peptides, due to it is important that the peptides possess a high antibacterial activity with a low toxicity for human beings. In this sense, the synthesis of toxic peptides could be reduced, saving time and money in the discovery of new potential antibiotics.

# Chapter 4

# Creation of the final classifier of antibacterial peptides

In the fight against multi-drug resistant bacteria, the antibacterial peptides have appeared as a promising source of new antibiotics. These peptides are a subset of the antimicrobial peptides (AMPs). To screen for new antibacterial peptides, computational tools are being used to avoid screening peptides that will not show activity against bacteria or will be toxic for humans. The purpose of an antibacterial peptides classifier is to determine if a peptide has activity against bacteria.

## 4.1 Creation of the input dataset

In this part of the thesis, the positive dataset (further denoted as A) includes all the peptides, from the APD [36] and the CAMP [47] databases, that report antibacterial activity, i.e., those peptides that are active against bacteria and other microorganisms. The total number of antibacterial peptides was 2,339. The negative dataset includes 460 peptides, from the mentioned databases, which do not report antibacterial activity (further denoted as B) and 4,889 peptides obtained by Wang *et. al* [73] from non-secretory proteins, and they are used as non-antimicrobial peptides.

The selection of the non-antimicrobial peptides involves two steps. The first one is to delete those peptides with a length larger than 100 amino acids, due to the synthesis difficulties of long peptides. Then, those peptides that had a sequence identity higher than 50% in a cluster analysis done with cd-hit [106,107] were eliminated. After these selection steps 4,889 non-active non-antibacterial peptides are left (further denoted as set C). Figure 4.1 shows a graphical representation of the dataset used to create the antibacterial peptides classifier.

## 4.2 Selection of the machine learning algorithm

This step consists in the selection of a learning algorithm that could achieve very good performance, in classification processes, given by the accuracy of the predictions and the computational complexity [108]. In general, a classification algorithm is used to perform a binary or a multi class classification. The difference between these two processes consists in the possible value for the predicted feature $y$. If $y$ takes values from $\{1, 2, ..., N\}$, where $N$ is the number of possible classes and if $N > 2$ it would correspond to a multi class classification. But if $N = 2$, the process is a binary classification [92]. Commonly, the classification processes of antimicro-

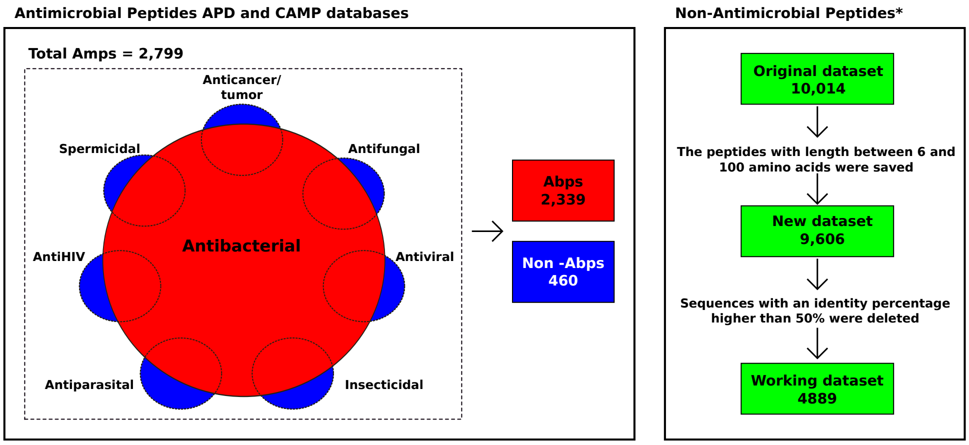

Antimicrobial Peptides APD and CAMP databases
Total Amps = 2,799
Anticancer/ tumor
Spermicidal
Antifungal
Antibacterial
AntiHIV
Antiviral
Antiparasital
Insecticidal
Abps 2,339
Non -Abps 460
Non-Antimicrobial Peptides*
Original dataset 10,014
The peptides with length between 6 and 100 amino acids were saved
New dataset 9,606
Sequences with an identity percentage higher than 50% were deleted
Working dataset 4889

Figure 4.1: Graphical representation of the dataset to create the antibacterial peptides classifier.

bial and antibacterial peptides have been done using binary classification.

However, there are some works that have applied different algorithms in the classification or prediction of antimicrobial peptides, such as: Fourier transformation to discover new candidate antimicrobial peptides [109], fuzzy logic for the prediction of antimicrobial activity [85], hidden Markov models in the prediction of candidate short cationic amphipatic antibacterial peptides [110], among others.

In this thesis, three types of antibacterial peptides classifiers were proposed, each one using a different configuration of negative datasets: antibacterial versus non-antibacterial (A vs B), antibacterial versus non-antimicrobial (A vs C), and antibacterial versus both non-antibacterial and non-antimicrobial (A vs B+C). Figure  4.2 shows the data configurations used in each one of these classifiers.

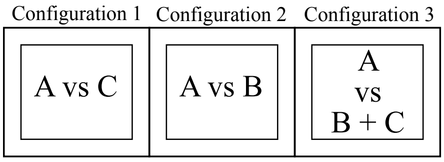


Figure 4.2: Proposed classifiers with their corresponding data configurations. A: antibacterial peptides, B: non-antibacterial peptides, C: non-antimicrobial peptides.

Four different classifiers were used: linear discriminant analysis (LDA) [111], support vector machines with linear kernel (SVM-linear) [111, 112], logistic classifier (LC) [111, 113] and k-nearest neighbors (kNN) [111]. The performance measures were calculated using a stratified 10-fold cross-validation loop, and the test set included samples of the three sets (A, B, and C). An inner 10-fold cross-validation was used to optimize the meta-parameters of the classifiers SVM-linear (C parameter), LC ($\lambda$ parameter) and kNN (k parameter).

The comparison of the classifiers might be considered as unfair for those classifiers that use only one of the negative subtypes in the training process, because the test data include samples of both negative subtypes. However, we consider this comparison as fair since our final goal is to use the classifier in the design of new antibacterial peptides.

## 4.3 Selection of peptide features

The selection of the peptide features that are used in a classification process affects the performance of the antimicrobial or antibacterial peptides classifier. Different strategies to deal with the selection of the best set of peptide features have been developed and can be grouped in two different approaches: automatic or manual selection.

The automatic selection approach allows the selection of the best set of features considering their performance in the mathematical model of the classifier. Some algorithms that have been used to perform this task are correlation-based methods, statistical criteria, methods based on information theory, genetic algorithms, simulated annealing, sequential feature forward selection, sequential backward feature elimination, among others [56]. Some algorithms that have been used to select peptide features in the rational design of AMPs are: principal component regression [70], partial least-squares regression [70], fractional factorial design [69] and incremental feature selection [73].

The manual selection comprehends those sets of descriptors selected by prior knowledge, i.e. they do not apply any algorithm to select them. Some studies, in the rational design of AMPs, that have used this approach can be found in [58, 61, 67, 68, 85, 110, 114].

In this section, the peptides were represented by the descriptors used in the previous chapter (See section 3.4.4), and we performed a feature set forward selection process. Instead of adding a single feature to the final set of selected features, we added a feature set in each step of the procedure. The learning algorithm used was the k-nearest neighbors for which the number of neighbors was optimized using an internal 10-fold cross-validation loop. An outer 10-fold cross-validation loop was used to obtain the performance measures to select the best feature set in each step; see Figure 4.3. The double cross-validation loop allows the evaluation of the classifiers performance using samples that were not used in the optimization process of the classifiers. Examples of studies that have used a double cross-validation loop can be found in [115, 116].

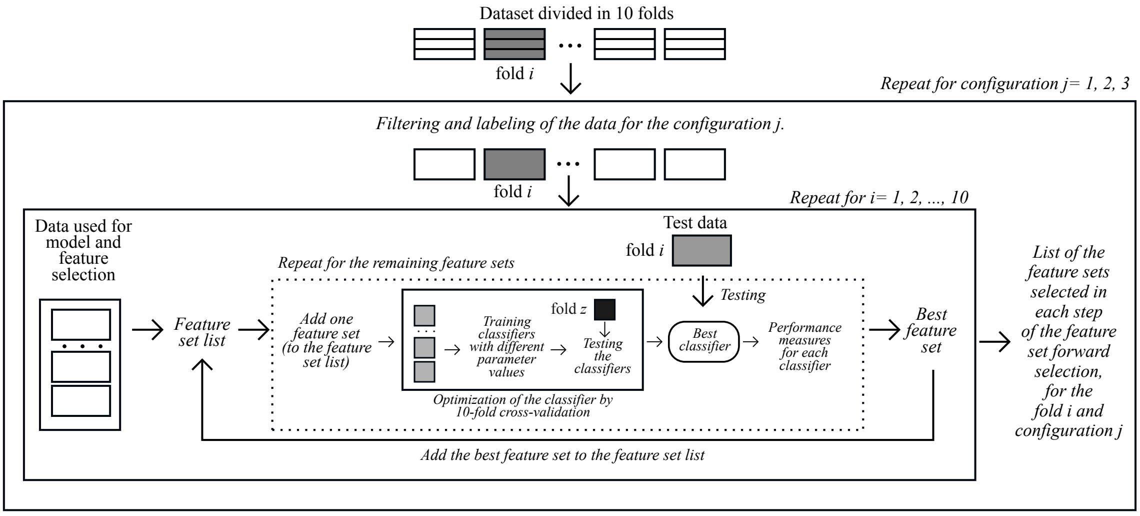

Dataset divided in 10 folds
fold *i*
*Repeat for configuration j= 1, 2, 3*
*Filtering and labeling of the data for the configuration j.*
fold *i*
*Repeat for i= 1, 2, ..., 10*
Data used for model and feature selection
Test data
fold *i*
*Repeat for the remaining feature sets*
*Feature set list*
*Add one feature set (to the feature set list)*
*Training classifiers with different parameter values*
fold *z*
*Testing the classifiers*
*Testing*
*Best classifier*
*Performance measures for each classifier*
*Optimization of the classifier by 10-fold cross-validation*
*Best feature set*
*Add the best feature set to the feature set list*
*List of the feature sets selected in each step of the feature set forward selection, for the fold i and configuration j*

Figure 4.3: Proposed methodology to perform the feature set forward selection.

## 4.4 Evaluation measures

For a fair comparison between the proposed configurations of classifiers, we created the classifiers inside the same 10-fold cross-validation loop, i.e. the samples assigned to the folds are synchronized. The performance measures used as optimization criteria were the area under the ROC curve (AUC), the partial area under the ROC curve which is integrated from 0% to 5% false positive (partial AUC, pAUC), and the average precision (AP). Figure 4.4 shows the graphical representation of the evaluation measures.

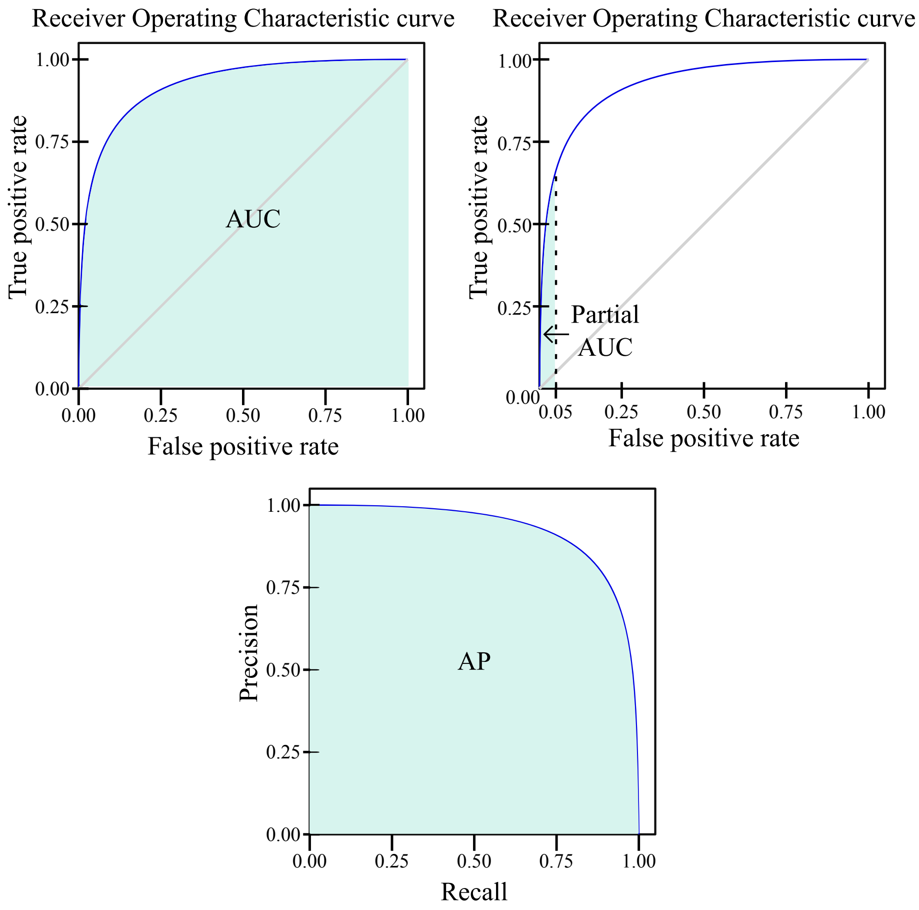


Figure 4.4: Graphical representation of the evaluation measures.

The optimization criterion plays an important role in model and feature selection processes. For example, the AUC measure, which comprehends the area of the entire ROC curve is not useful to differentiate between two classifiers that behave differently in one specific region of specificity. In a similar way, the AP, which comprehends the area under the precision-recall curve, does not allow a differentiation between two classifiers that achieve different precision values in different regions of sensitivity. In this sense, the pAUC is most useful when only certain regions of the ROC space are of particular interest. In our case, a small false positive rate (i.e., high specificity region) is preferred, due to the costs (in time and money) associated with the synthesis of peptides that do not possess antibacterial activity. Figure 4.5 presents an example where the selection of the best classifier is affected depending on the evaluation measure used. If the AUC value is used, then the best classifier is the blue one. On the contrary, if the pAUC is used, then the best classifier is the red one.

Moreover, the pAUC values might not be as good as desirable, since in multiple applications

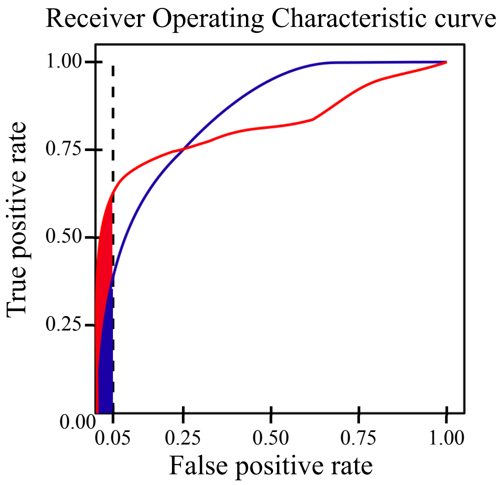


Figure 4.5: Two ROC curves are plotted. The pAUC value of the red curve is higher than the pAUC value of the blue one. However, if the AUC value is observed the situation is the opposite, the AUC value of the blue one is higher than the red one.

it is difficult to create a classifier that possess a low false positive rate and a high true positive rate.

## 4.5 Cascade models as a strategy to improve the performance.

Because our intent is to find the most promising antibacterial peptide, one may wonder if we should not construct classifiers that are even more focused on only finding the most positive peptides. For that we created a cascade classifier in which the first base classifier of the cascades evaluates the samples, and if the probability of being positive is higher than 0.5, then, samples are evaluated in the second base classifier. The final probability will be the one given by the second base classifier. Figure 4.6 shows the three different configurations that we explored.

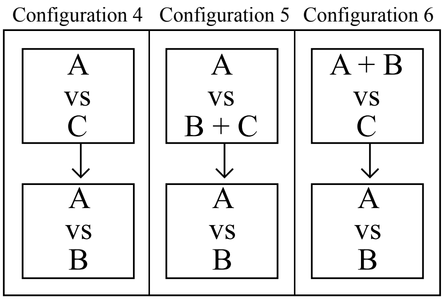


Figure 4.6: Proposed cascade classifiers with their corresponding datasets. A: antibacterial peptides, B: non-antibacterial peptides, C: non-antimicrobial peptides.

## 4.6 Results and Discussion

Figure 4.7 shows the t-SNE and PCA (principal component analysis) maps [117] for the samples of different peptide sets A, B and C, based on the proposed feature sets. These mappings suggest that there is not a clear distinction between the positive class and both negative classes.

As previous classifiers are able to achieve good results this suggests that distances in the original space are not well preserved in the 2D maps.

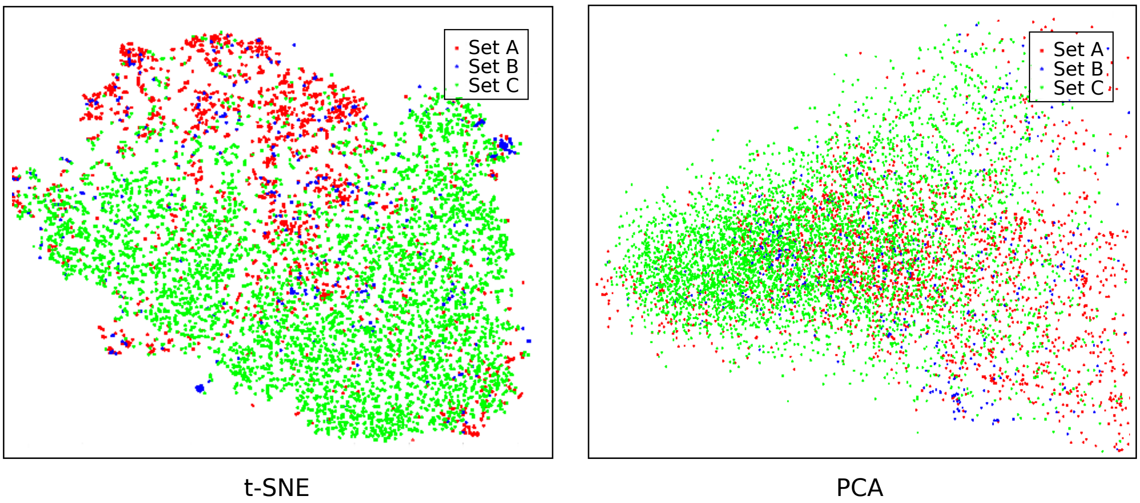


Figure 4.7: t-SNE and PCA maps of the dataset used to create the antibacterial peptides classifier, using the 1631 features of table 3.4. A: antibacterial peptides, B: non-antibacterial peptides and C: non-antimicrobial peptides.

### 4.6.1 A non-linear model might be suitable to classify antibacterial peptides than linear models.

Table 4.1 shows the pAUC, AP and AUC values for all the proposed classifiers. Bold values are the best significant values, according with ANOVA and Tukey tests.

The best performance values are obtained by the nonlinear kNN. For the pAUC there are no significant differences between configuration 2 and 3 classifiers. On the contrary, for the AP and AUC values, there are no significant differences between configuration 1 and 3 classifiers (p-value $> 0.05$). These results suggest that for the reduction of the false positives rate it is necessary to use the non-antibacterial peptides, i.e. set B, in the training process of a nonlinear classifier.

In contrast, AUC values suggest to use only the non-antimicrobial peptides (set C) in the negative set. According to this measure, there are no significant differences between the linear classifiers (SVC linear and LC), the nonlinear kNN for configuration 1 classifier, and the nonlinear kNN for configuration 3 classifier that uses both negative subtypes. Therefore, looking at AUC values might lead to wrong conclusions and care must be taken when using this performance measure.

Table 4.1: pAUC (for p=0.05), AP and AUC values, and between parenthesis the standard deviation over the 10 folds, for the classifiers trained according to configurations 1, 2 and 3 created using all available features. A: antibacterial peptides, B: non-antibacterial peptides and C: non-antimicrobial peptides.

| Measure | Classifier | LDA | SVC linear | LC | kNN |
|---|---|---|---|---|---|
| pAUC | Conf. 1: A vs C | 0.001(0.000) | 0.021(0.004) | 0.016(0.003) | 0.023(0.003) |
| | Conf. 2: A vs B | 0.002(0.000) | 0.002(0.001) | 0.001(0.000) | **0.027(0.002)** |
| | Conf. 3: A vs B+C | 0.020(0.003) | 0.020(0.003) | 0.013(0.002) | **0.030(0.003)** |
| AP | Conf. 1: A vs C | 0.312(0.006) | 0.841(0.037) | 0.807(0.024) | **0.878(0.017)** |
| | Conf. 2: A vs B | 0.290(0.024) | 0.364(0.023) | 0.341(0.011) | 0.808(0.023) |
| | Conf. 3: A vs B+C | 0.312(0.004) | 0.831(0.023) | 0.758(0.016) | **0.885(0.018)** |
| AUC | Conf. 1: A vs C | 0.493(0.004) | **0.922(0.018)** | **0.925(0.010)** | **0.943(0.007)** |
| | Conf. 2: A vs B | 0.595(0.017) | 0.568(0.032) | 0.555(0.027) | 0.833(0.021) |
| | Conf. 3: A vs B+C | 0.494(0.003) | 0.914(0.012) | 0.891(0.007) | **0.920(0.015)** |

### 4.6.2 The selection of the features sets is affected by both the optimization criterion and the configuration of the negative dataset.

Figure 4.8 shows the results of the feature set forward selection for the proposed configurations using different optimization criteria. Each row corresponds to a different feature set, and each column represents at which step, of the feature set forward selection, the feature set is being selected. The color in each box corresponds to the number of times that a feature set was chosen in each step of the selection process, inside a 10-fold cross-validation loop.

The selection of the feature sets is affected by both the configuration of the negative class, and the optimization criterion. When only set C is used as the negative class, results are much different from when set B or sets B+C are used. Apparently, feature set $f_{10}$ (which contains charge values) is very useful for the discrimination of set C from A, but when the more challenging set B is added, this does not hold anymore. These results are supported by multiple studies that claim that a positive net charge is essential for the initial interaction with negatively charged cell membranes [24, 76]. From a biological point of view, it is expected that peptides from set C do not possess a positive net charge and therefore, these features are preferred in the classification process. However, peptides from set B might possess a positive net charge that helps in the biological activity against other microorganisms, and consequently, these features are not longer preferred.

The preferred feature set for the classifiers that use the set B is the set $f_{13}$ that corresponds to the dipeptide composition. This feature set does not encode, explicitly, any physicochemical information, but just information on the amino acid content and the order of the amino acids in the peptide sequences. Moreover, when the pAUC is used as the optimization criterion, this feature set is highly selected for all the three classifiers. These results suggest that antibacterial peptides can be recognized from their peptide sequences or sequence fragments. This is supported by Bodapati *et. al* [118] who showed that the C-terminal region is required for specificity and dictates the antimicrobial profile, whereas, the N-terminal sequence is necessary for activity.

When sets B+C are used, and the optimization criterion is changed from pAUC to AP or AUC, the preferred feature set change from $f_{13}$ to $f_{12}$. These results show that the amino acid content of the peptides ($f_{12}$) is very useful in the discrimination of antibacterial peptides. However, if the goal is to reduce the number of false positives, the order of the amino acids

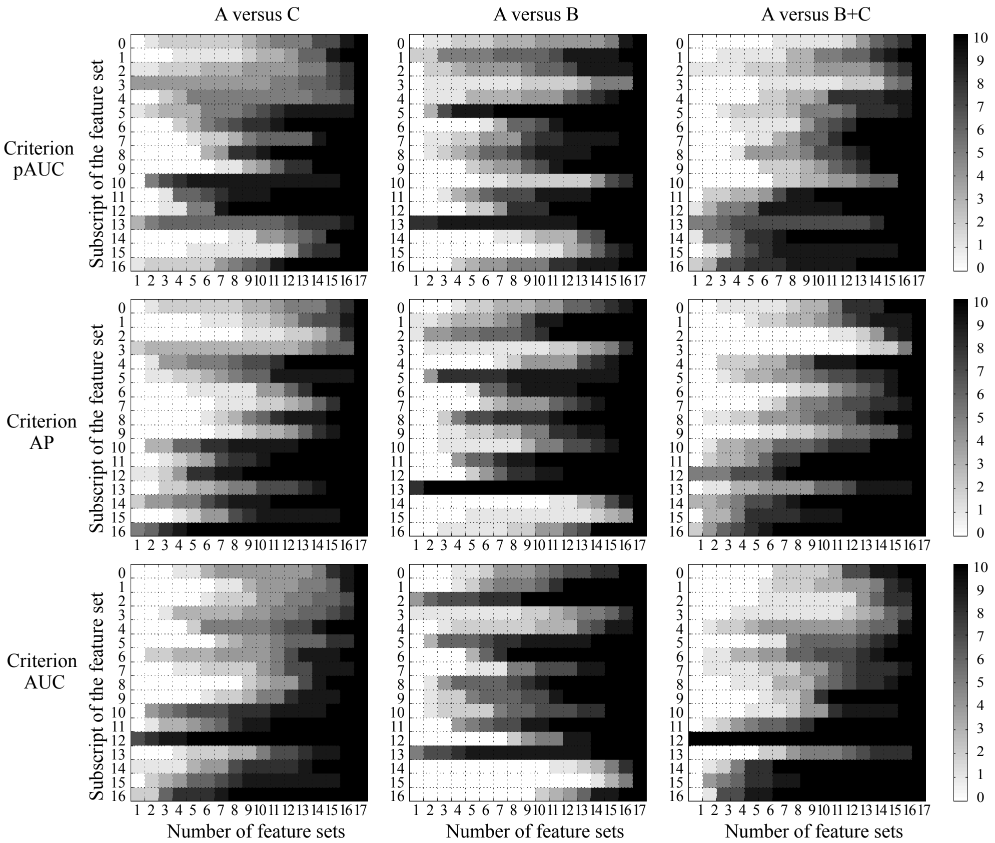


Figure 4.8: Results for the feature set forward selection process for the single classifiers. Each box contains the number of times that a feature set was chosen in each step of the selection process, inside a 10-fold cross-validation loop. A: antibacterial peptides, B: non-antibacterial peptides, C: non-antimicrobial peptides.

inside the peptide sequences might lead to better results. Additionally, if the AUC measure is used as the optimization criterion, not only the selected feature sets are affected, but also the recommended configuration of the classifier (i.e., AUC favors Configuration 1: A versus C).

Therefore, the model and feature selection processes should be performed with the optimization criterion that might lead to the desired results. In our case, the pAUC criterion, to select the classifier (with its corresponding features) that achieves a good performance with a low false positive rate.

### 4.6.3 Both negative subtypes should be used in the creation of an antibacterial peptides classifier.

A comparison between the three configurations of classifiers was performed using the feature sets that were most often selected for the classifiers trained, according to configurations 1 and 3 using the pAUC as the optimization criterion. For configuration 1 the best performance is obtained when the feature set $f_{13}$ or $f_3$ is used. For configuration 3 those are feature sets $f_{13}$ and $f_{14}$. Table 4.2 shows the performance values for the three configurations when different feature sets are used ($f_3$, $f_3 + f_{13}$, $f_{13}$, and $f_{13} + f_{14}$). We performed an ANOVA and a Tukey test for each performance measure, and bold values are the best significant values among the 12 classifiers. For the AP and AUC measures, there is no significant difference between configuration 1 and configuration 3. For the pAUC measure, configuration 1 is significantly worse than the other two configurations. Hence, configuration 3 obtained competitive results across the three performance measures, and therefore, the use of both negative subtypes is recommended in the creation of an antibacterial peptides classifier.

Table 4.2: pAUC (p=0.05), AUC and AP values obtained for each classifier strategies when using different feature sets

| Measure | Classifier | $f_3$ | $f_3 + f_{13}$ | $f_{13}$ | $f_{13} + f_{14}$ |
|---|---|---|---|---|---|
| pAUC | Conf. 1: A vs C | 0.024(0.003) | 0.024(0.004) | 0.023(0.003) | 0.023(0.003) |
| | Conf. 2: A vs B | 0.023(0.002) | 0.024(0.008) | **0.027(0.002)** | **0.026(0.002)** |
| | Conf. 3: A vs B+C | **0.026(0.003)** | **0.029(0.003)** | **0.029(0.002)** | **0.030(0.002)** |
| AP | Conf. 1: A vs C | **0.922(0.010)** | **0.930(0.011)** | **0.914(0.010)** | **0.928(0.009)** |
| | Conf. 2: A vs B | 0.784(0.023) | 0.794(0.082) | 0.805(0.032) | 0.798(0.031) |
| | Conf. 3: A vs B+C | **0.892(0.021)** | **0.897(0.018)** | 0.881(0.021) | **0.898(0.022)** |
| AUC | Conf. 1: A vs C | **0.857(0.020)** | **0.868(0.023)** | **0.851(0.019)** | **0.865(0.018)** |
| | Conf. 2: A vs B | 0.747(0.024) | 0.746(0.143) | **0.787(0.029)** | **0.780(0.027)** |
| | Conf. 3: A vs B+C | **0.847(0.025)** | **0.863(0.021)** | **0.894(0.021)** | **0.866(0.021)** |

### 4.6.4 Single classifiers outperform cascade classifiers.

Table 4.3 shows the results of the performances obtained with the proposed single and cascades classifiers using all feature sets. In general, the cascade classifiers obtained competitive results, although there is always a single classifier that outperforms all of the proposed cascades, for each of the three performance measures. From this, we conclude that the strategy of a cascade model is not preferred.

Table 4.3: Performance values obtained for single and cascade classifiers when the pAUC is used as the optimization criterion and all features sets are used.

| Classifier | pAUC(std) | AP(std) | AUC(std) |
|---|---|---|---|
| Conf. 1: A vs C | 0.023(0.003) | **0.878(0.017)** | **0.943(0.007)** |
| Conf. 2: A vs B | **0.027(0.002)** | 0.808(0.023) | 0.833(0.021) |
| Conf. 3: A vs B+C | **0.030(0.003)** | **0.885(0.018)** | 0.920(0.015) |
| Conf. 4: A vs C → A vs B | 0.023(0.003) | 0.823(0.017) | 0.877(0.014) |
| Conf. 5: A vs B+C → A vs B | 0.027(0.002) | 0.795(0.021) | 0.819(0.019) |
| Conf. 6: A+B vs C → A vs B | 0.014(0.002) | 0.777(0.017) | 0.872(0.011) |

Moreover, Figure 4.9 shows the best ROC curve obtained for each one of the proposed configurations. It can be observed that the pAUC is higher for configuration 3 classifier. However, if the AUC value is considered, the higher value is obtained with the configuration 1 classifier. Moreover, the ROC curves show that cascade classifiers and the configuration 2 classifier do not obtain a high true positive rate and the AUC value is not competitive in comparison with the AUC values obtained for configuration 1 and 3 classifiers. Finally, the ROC curve of configuration 2 is wider than the ROC curve of configuration 1 in the region delimited by the red line, hence, the pAUC value is higher for configuration 2 classifier.

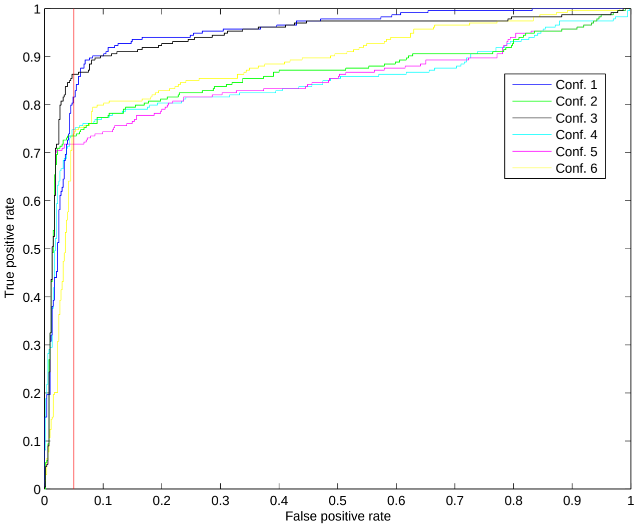


Figure 4.9: ROC curves of the 6 configurations. The red line indicates a value of 0.05 for the false positive rate.

### 4.6.5 The logistic classifier is preferred for the design of new antibacterial peptides

An analysis of the results obtained with the kNN classifier showed that the best value for the parameter $k$ was always set to 1 regarding the configuration of the negative dataset or the optimization criteria used. In this sense, this model might not be very good at generalization and another model should be used.

Moreover, in the design of new antibacterial peptides, it is very useful to obtain the probability that a peptide possess antibacterial activity and the probability that a peptide is a non-toxic peptide. Hence, the logistic classifier is a good option considering that LC provides estimates of the posterior probability of class membership while SVM is purely discriminative and it is necessary to map the SVM scores toward probabilities.

Therefore, a feature set forward selection process was performed for the configuration 3 classifier using LC. Figure  4.10 shows the obtained results. The feature sets that were highly selected were $f_3$ (composition, transition, distribution descriptors), $f_{13}$ (dipeptide composition) and $f_5$ (quasi-sequence-order descriptors). As in the case of the kNN, the feature sets that were highly selected with LC, were feature sets that are related to the amino acid composition and the order of the amino acids in the peptide sequences. Different models were built for the three configurations using different feature sets. Table  4.4 shows the pAUC, AP and AUC values for the different classifiers. The results of ANOVA and Tukey test show that there are not significant differences between configuration 3 classifiers using different feature sets, and any of these models could be selected as the final classifier. Finally, the feature selection process allowed to improve the performance of the logistic classifier.

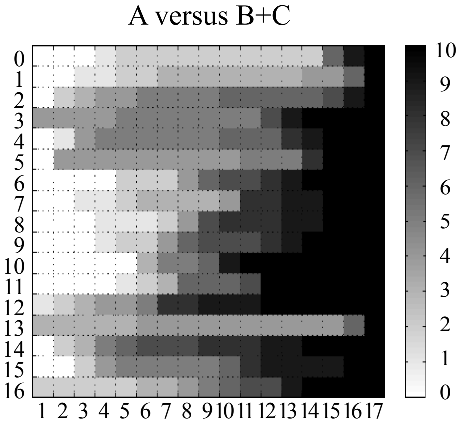


Figure 4.10: Results for the feature set forward selection process for configuration 3 classifier. Each box contains the number of times that a feature set was chosen in each step of the selection process, inside a 10-fold cross-validation loop. A: antibacterial peptides, B: non-antibacterial peptides, C: non-antimicrobial peptides.

Table 4.4: pAUC, AP and AUC values obtained for each single classifier when the pAUC, with p=0.05, is used as the optimization criterion and different feature sets are used

| | Classifier | $f_3$ | $f_{13}$ | $f_3 + f_5$ | $f_3 + f_{13}$ | $f_5 + f_{13}$ |
|---|---|---|---|---|---|---|
| pAUC | Conf. 1 | 0.016(0.003) | 0.015(0.003) | 0.015(0.003) | 0.014(0.003) | 0.016(0.003) |
| | Conf. 2 | 0.004(0.001) | 0.001(0.000) | 0.003(0.001) | 0.001(0.001) | 0.001(0.001) |
| | Conf. 3 | **0.022(0.004)** | **0.022(0.003)** | **0.023(0.004)** | **0.022(0.003)** | **0.022(0.003)** |
| AP | Conf. 1 | 0.821(0.026) | 0.818(0.021) | 0.827(0.025) | 0.816(0.022) | 0.822(0.024) |
| | Conf. 2 | 0.469(0.039) | 0.339(0.006) | 0.428(0.029) | 0.343(0.005) | 0.341(0.008) |
| | Conf. 3 | **0.851(0.025)** | **0.855(0.020)** | **0.865(0.027)** | **0.862(0.019)** | **0.854(0.022)** |
| AUC | Conf. 1 | **0.919(0.012)** | **0.918(0.011)** | **0.926(0.011)** | **0.922(0.011)** | **0.921(0.010)** |
| | Conf. 2 | 0.660(0.032) | 0.522(0.023) | 0.625(0.026) | 0.530(0.013) | 0.526(0.023) |
| | Conf. 3 | **0.925(0.012)** | **0.926(0.013)** | **0.934(0.011)** | **0.930(0.012)** | **0.926(0.011)** |

### 4.6.6 Final classifier

The classifiers obtained using pAUC as the optimization criterion and configuration 3 were applied to the sets A, B and C, and their probabilities were plotted. Figure 4.11 shows the boxplots obtained with each model. According with the results, we decided to work with the model obtained using the feature sets $f_3 + f_{13}$, because it is the model that allows the best differentiation between antibacterial peptides and non-antibacterial/non-antimicrobial peptides.

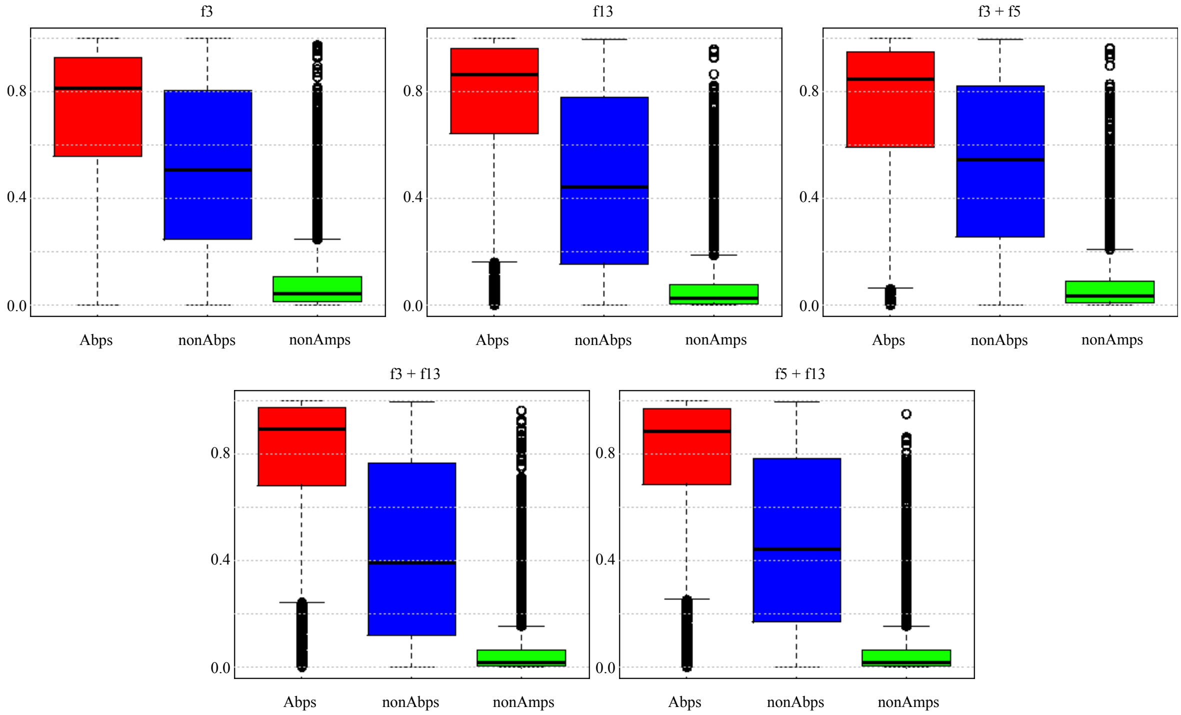


Figure 4.11: Boxplots of the estimated probability of a peptide to possess antibacterial activity, for configuration 3 classifiers using different feature sets.

## 4.7 Conclusions

The discovery of new antimicrobial peptides that possess antibacterial activity could lead to the creation of new antibiotics and might help in the fight against antibacterial resistance. We showed that when building an antibacterial peptides classifier, it is helpful to include both the non-antibacterial and the non-antimicrobial peptides as negative examples (instead of only one category).

The optimization criterion has an influence in both, the configuration of the classifier and the feature selection process. Based on our results, we recommend the use of the pAUC in model and feature selection, for those applications where a low false positive rate is preferred. This measure allows the selection of the classifier that possess a high true positive rate in a region of specificity. In contrast, the AUC and AP are not useful to discriminate between classifiers that possess different behaviors in distinct regions of specificity or sensitivity, respectively. Thereby, we recommend using these two measures after the selection processes to have an understanding of the final classifier performance.

We did find that features related to the sequence description of the peptide were preferred to be used by the classifier than features that have a biophysical meaning. These results suggest that the order of the amino acids, inside the peptide sequence, plays an important role in the discrimination between antibacterial and non-antibacterial peptides. For peptides with the same amino acid content and different peptide sequences, multiple biophysical features have the same value. This might be the case for features such as charge, weight, isoelectric point, among others. Thereby, sequence description features might lead to better performance of classifiers.

Cascade classifiers were not able to improve on the best single. This situation might be caused by the low performance of the second base classifier, due to the few number of train samples for this classifier. The final performance of these cascades is given by a combination of the performances of both base classifiers. From that we conclude that, the addition of the second base classifier, which is the same for all cascade configurations, reduces the performance achieved by the first one.

In the design of new antibacterial peptides using classification algorithms, it is important to consider learning algorithms that use most of the training data and that allow an estimation of the posterior probability. A binary classifier with a hard label might not be very useful in the design and improvement of antibacterial peptides, due to the impossibility to determine the peptide sequence with the higher probability of having antibacterial activity.

# Chapter 5

# Creation of the classifier of toxic peptides

There are multiple antimicrobial peptides that are active against multidrug-resistant bacteria. However, their toxicity is one of the main limitations for these molecules to become effective antibiotics. Hence, computational approaches that can predict the toxicity of peptides are highly demanding. These methods allow the design of peptides with low toxicity while retaining the biological activities, saving time and money in the creation of new therapeutics [28].

In the literature, we can find a number of user-friendly webservers that can be used in the design of non-toxic peptides. Here some examples:

- ClanTox is a classifier of animal toxins [119].
- BTXpred is used to predict bacterial toxins [120]
- NTXpred predicts neurotoxins [121]
- ToxinPred predicts toxic peptides [28]
- DBETH predicts bacterial toxins [122], among others

The datasets used to create these tools are different for each one of them. Hence, it is not possible to perform a fair comparison between them. Moreover, the datasets comprehends peptides with a length higher than 100 amino acids, with exception of the tool ToxinPred which uses peptides with less than 35 amino acids. Therefore, we decided to create a toxic peptides classifier that comprehends peptides with a length between 7 and 100 amino acids, and that evaluates a peptide and gives its probability of being toxic.

## 5.1 Materials and Methods

### 5.1.1 Dataset

In the literature, there are not many works that deals with the classification of toxic or non-toxic peptides. Therefore, there is not a benchmark dataset to work with. In the work of Gupta *et. al.* [28], a toxicity classifier was created using the information available in multiple databases and the results were satisfactory. However, the datasets used in this study comprehend peptides of length less than 35.

In this thesis we are working with peptides with a maximum length of 100 amino acids. Therefore, we used a similar strategy than the one used by Gupta *et. al.* [28], to create a dataset that comprehends toxic and non-toxic peptides with length between 7 and 100 amino acids.

The peptides of our dataset were obtained from the database SwissProt [123]. Only those peptides that are reviewed, do not possess unusual amino acids (i.e. X, Z, B and U) and their length is between 7 and 100 were considered. Moreover, the positive class or toxic peptides comprehends those peptides that have the keyword toxin (KW0800), while the peptides of the negative class or non-toxic peptides do not contain the keywords toxin (KW0800) and allergen(KW).

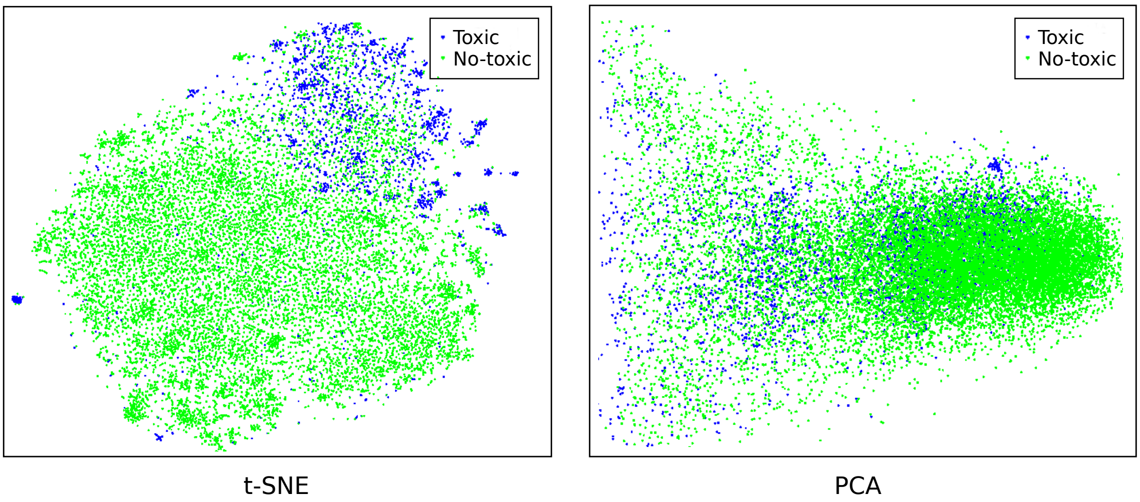
The resulting number of toxic peptides was 3,973 and the number of non-toxic peptides was 50,406. In order to reduce the ratio of imbalance between the two classes, we remove the sequences that were highly homologues using the program CD-Hit [107] with C=0.7. This step allow us to reduce the size of the negative class to 16,484 peptides in a not random way. Hence, it is expected to obtain a better performance for samples of the minority class [88].

### 5.1.2 Creation of the toxic peptides classifier

To create the desired classifier we used the same feature sets that were used in the creation of the antibacterial peptides classifier (See chapter 4). Additionally, we also performed a comparison between 4 learning algorithms and a feature set forward selection process. Finally, the classifier that will be used in the rational design of new peptides is the one that allows the best differentiation between the desired group of peptides, i.e. non-toxic and non-hemolytic antibacterial peptides, and the rest.

## 5.2 Results and Discussion

Figure 5.1 shows the t-SNE and PCA maps [117] for the toxic and non-toxic peptides. In the 2D maps, there is not a good separation between the positive and negative class. However, since linear classifiers obtained good performance results when all the features are used, it is likely that a clear differentiation exists in a higher dimensional space.



t-SNE

PCA

Figure 5.1: t-SNE and PCA maps of the dataset used to create the toxic peptides classifier, using the 1631 features of table 3.4.

### 5.2.1 The logistic classifier is preferred to create the toxic peptides classifier

As a first step, we performed a comparison between four learning algorithms (LDA, NMC, LC, and kNN) using all the available feature sets. The meta-parameters of the classifiers LC ($\lambda$) and kNN ($k$) were optimized using an inner 10-fold cross-validation loop, and the pAUC as the optimization criterion. This criterion was choosen because our main interest is to increase the number of true negatives (i.e., reduce the number of false positives).

Table 5.1 shows the performance measures obtained with the four classifiers. Bold values are the best significant values (p_value < 0.05).

Table 5.1: Comparison of learning algorithms

| Classifier | pAUC | AUC | AP |
|---|---|---|---|
| LDA | 0.001(0.000) | 0.486(0.006) | 0.192(0.001) |
| NMC | 0.004(0.001) | 0.351(0.017) | 0.698(0.012) |
| LC | 0.031(0.002) | 0.951(0.009) | 0.855(0.017) |
| kNN | **0.044(0.001)** | **0.960(0.008)** | **0.982(0.006)** |

Although the non-linear kNN algorithm obtained the best performance, this model might not be very useful due to the value of the parameter $k$ was always set to 1 by the optimization process. This situation is not desirable since this model might not be good at generalization. Therefore, the logistic classifier is preferred for the creation of the toxicity peptides classifier, and a feature set forward selection process was performed, using the pAUC as the optimization criterion. Figure 5.2 shows the results obtained. The feature sets that were highly selected were $f_{13}$, $f_2$, $f_{14}$ and $f_{16}$. Different models were built using combinations of these feature sets. Table 5.2 presents the pAUC, AP and AUC values for the different classifiers. The results of ANOVA and Tukey test show that there are not significant differences between these models, and any of them could be selected as the final classifier.

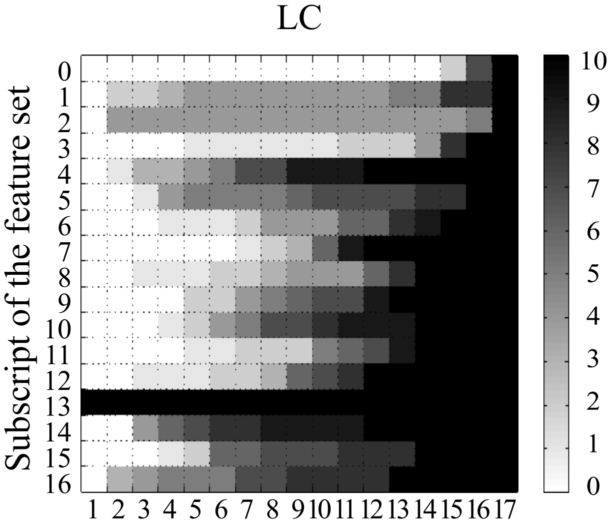


Figure 5.2: Results of the feature set selection process for the toxicity classifier using LC and pAUC as the optimization criterion.

Table 5.2: pAUC, AP and AUC values obtained for the logistic classifier when the pAUC, with p=0.05, is used as the optimization criterion and different feature sets are used

| | $f_{13} + f_2 + f_{14}$ | $f_{13} + f_{14} + f_{16}$ | $f_{13} + f_{16} + f_2$ | $f_{13} + f_2 + f_{14} + f_{16}$ |
|---|---|---|---|---|
| pAUC | 0.034(0.002) | 0.033(0.001) | 0.034(0.002) | 0.034(0.002) |
| AP | 0.883(0.015) | 0.877(0.014) | 0.884(0.017) | 0.889(0.015) |
| AUC | 0.959(0.007) | 0.958(0.008) | 0.959(0.008) | 0.960(0.007) |

### 5.2.2 The logistic classifier allows the differentiation of non-toxic and non-hemolytic peptides

Since the main objective is to design peptides that are likely to possess a high antibacterial activity and a low probability of being toxic, we divided the antibacterial peptides that were used in the creation of the antibacterial peptides classifier, in 4 groups that were created according to their toxicity or hemolicity:

1. **Group 1 - Abps non-toxic and non-hemolytic:** Those antibacterial peptides that belong to the APD or CAMP database and are found at SwissProt as non-hemolytic and non-toxic (188 peptides).

2. **Group 2 - Abps toxic and hemolytic:** Those antibacterial peptides that belong to the APD or CAMP database and are found at SwissProt as toxic and hemolytic (29 peptides).

3. **Group 3 - Abps toxic:** Those antibacterial peptides that belong to the APD or CAMP database and are found at SwissProt as toxic (14 peptides).

4. **Group 4 - Abps hemolytic:** Those antibacterial peptides that belong to the APD or CAMP database and are found at SwissProt as hemolytic (86 peptides).

The peptides of these 4 groups were evaluated in the proposed classifiers and their probability of being toxic were saved and plotted in boxplots (Figure 5.3). Although there are not significant differences between the models, the classifier created with the feature sets $f_{13}$, $f_2$, $f_{14}$, and $f_{16}$ allows a good differentiation between the peptides of our interest (non-toxic and non-hemolytic antibacterial peptides) and the toxic and/or hemolytic peptides. Therefore, the final classifier is built using these feature sets.

## 5.3 Conclusions

Predicting the probability that a peptide would be toxic is an important step in the design of novel potential antibiotics. In this sense, computational methods like logistic classifier are important tools that can be used to predict posterior probabilities with competitive results. Our final classifier was created with the logistic classifier and the values for the pAUC (with p=0.05), AP and AUC were 0.034, 0.889 and 0.960, respectively.

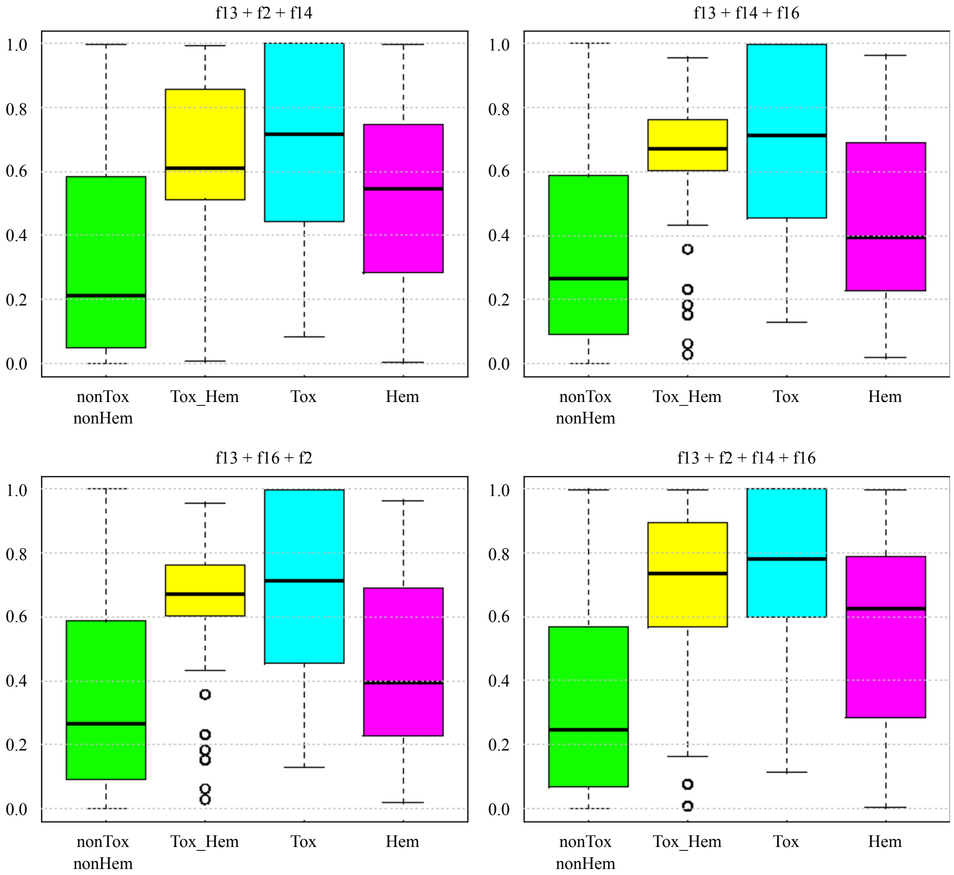


Figure 5.3: Boxplots of the estimated probability that a peptide would be toxic, using classifiers created with different feature sets

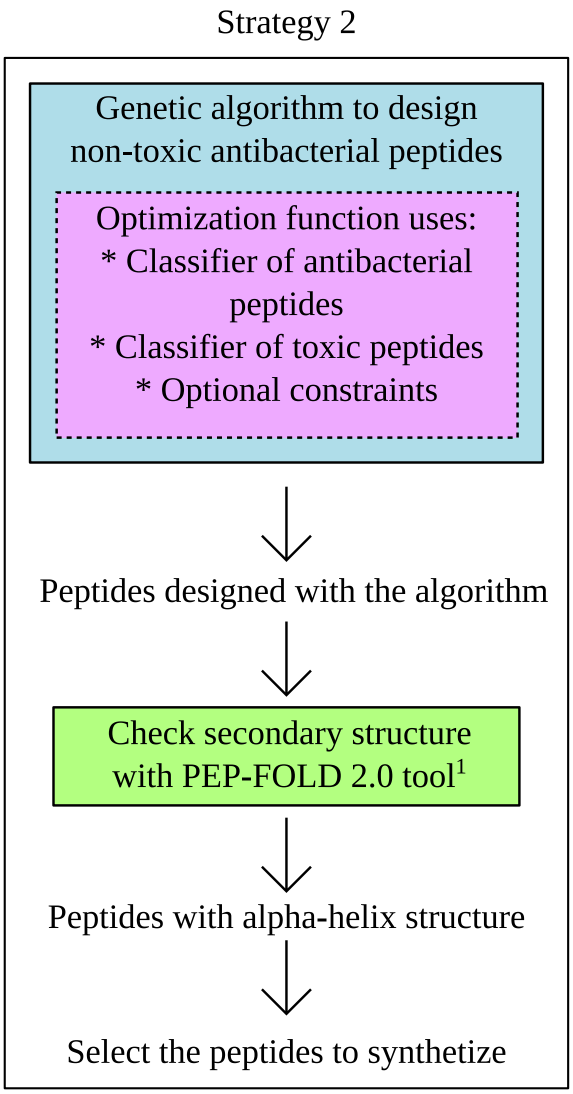
The features that were selected as the most informative in the context of toxicity are mainly those features based in the amino acid composition and the order of the amino acids inside the peptide sequence. Additionally, the features obtained with the Geary autocorrelation allowed a better discrimination between non-toxic and non-hemolytic antibacterial peptides. In this sense, physicochemical properties like atomic mass, atomic Van der Waals volume, atomic electronegativities and atomic polarizabilities of the amino acids present in a peptide, should be analyzed in order to determine a possible relationship between them and the toxicity and hemolicity of the peptides.

The final classifier allows the differentiation between the desired peptides (i.e., non-toxic and non-hemolytic antibacterial peptides) and the toxic and/or hemolytic peptides. Therefore, it is used in the rational design of new antibacterial peptides in the next chapter.

# Chapter 6

# Strategy 2 to design potential non-toxic antibacterial peptides

In this strategy, the genetic algorithm involves the decision of the classifiers in the design process of the peptides. Moreover, a toxic peptides classifier in now involved in the design of new peptides. Once the new potential non-toxic antibacterial peptides have been designed, the next step comprehends the analysis of their secondary structure by using the PEP-FOLD 2.0 tool [82]. The idea is to synthesize those peptides that possess a predicted alpha-helix structure. Figure 6.1 shows the proposed strategy.


Strategy 2
Genetic algorithm to design non-toxic antibacterial peptides
Optimization function uses:
* Classifier of antibacterial peptides
* Classifier of toxic peptides
* Optional constraints
Peptides designed with the algorithm
Check secondary structure with PEP-FOLD 2.0 tool[1]
Peptides with alpha-helix structure
Select the peptides to synthetize

[1] http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::PEP-FOLD

Figure 6.1: Workflow of the proposed strategy 2.

## 6.1 Genetic algorithm PRODEABPs 1.0

In this strategy, we designed and developed a genetic algorithm (PRODEABPs 1.0) that allows the generation of a desired number of peptides for which the predicted probabilities of being toxic and antibacterial are under and above the thresholds established by the user, respectively.

Moreover, this algorithm designs peptides, of a given length, that satisfy the established ranges for each one of the physicochemical descriptors selected in this thesis: charge, hydrophobicity, isoelectric point, and instability index. The user can modify the default ranges. However, we recommend to use the ranges that were established in the genetic algorithm DEPRAMPs 1.0.

## 6.2 Overview of the designed genetic algorithm

The optimization problem to solve is given by the Equation 6.1,

$$\begin{aligned} &Maximize\ fitness(p)\\ &\quad subject\ to:\\ &\quad h_1(p): prob_{abp}(p) > desProb_{abp}\\ &\quad h_2(p): prob_{tox}(p) < desProb_{tox}\\ &\quad h_3(p): x1_{min} < x_1(p) < x1_{max}\\ &\quad h_4(p): x2_{min} < x_2(p) < x2_{max}\\ &\quad h_5(p): x3_{min} < x_3(p) < x3_{max}\\ &\quad h_6(p): \qquad\quad x_4(p) < x4_{max} \end{aligned} \tag{6.1}$$

where, $p$ is the amino acid sequence of a peptide, $fitness(p)$ is the fitness of the peptide $p$ and is given by the Equation 6.2, $prob_{abp}(p)$ is the probability that the peptide $p$ has antibacterial activity, and $prob_{tox}(p)$ is the probability that the peptide $p$ is toxic. The desired thresholds for the antibacterial probability and the toxicity probability are given by $desProb_{abp}$ and $desProb_{tox}$, respectively. $x_1(p), x_2(p), x_3(p)$, and $x_4(p)$ are the charge, isoelectric point, hydrophobicity and instability index of the peptide $p$, respectively. The range of possible values for each of these physicochemical properties is given by the user, the minimum allowed value is $x_{i_{min}}$ and the maximum is $x_{i_{max}}$ where the value of $i$ indicates the desired property.

$$fitness(p) = \left(\frac{1}{1+e^{(-prob_{abp}(p)+0.9)}} - 0.5\right) * 100 \ + \ \left(\frac{1}{1+e^{(-prob_{non-tox}(p)+0.9)}} - 0.5\right) * 200 \tag{6.2}$$

In the Equation 6.2, the term $prob_{non-tox}(p)$ is the probability that the peptide $p$ is a non-toxic peptide, and is given by $1 - prob_{tox}(p)$. The design of this function was based in the assumption that a high antibacterial probability value and a low toxicity probability value are desired. Hence, a shifted sigmoid function was used for each probability in order to assign positive values to probabilities that are higher than 0.9. However, the designed function works for any desired probability value due to its value increases if the probability value increases. Additionally, in the fitness function the contribution of the non-toxic probability is multiplied by twice the factor of the antibacterial part, due to the difficulty to design peptides that has a low toxicity probability. With this strategy, the peptides that are non-toxic are preferred over those that have a high antibacterial probability. When a peptide has both probabilities in the desired values, the peptide becomes a candidate peptide and it is preferred over those peptides that only posses one probability in the desired value. Moreover, the $prob_{abp}(p)$ and $prob_{tox}(p)$ values are given by the final classifiers of the chapter 4 and 5, respectively. Finally, the penalization function to deal with the constraints is included in the selection of the parents in order to lead the solution to the desired intervals. Figure 6.2 shows the plot of the fitness function used in the genetic algorithm.

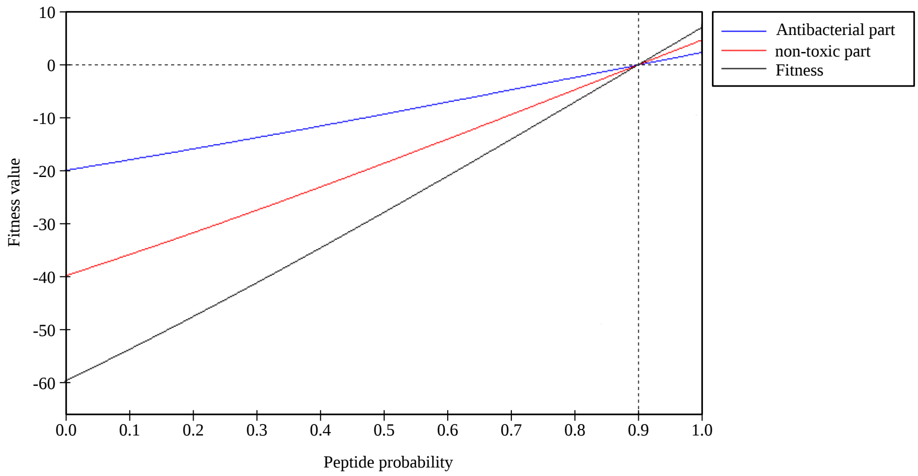


Figure 6.2: Graphic of the fitness function used in the genetic algorithm. The contribution of the non-toxic probability is twice the contribution of the antibacterial probability since it is more difficult to find a peptide with a high probability of not being toxic, than to find a peptide with a high probability of being antibacterial.

The workflow of this genetic algorithm is the same as the one used in the genetic algorithm DEPRAMPs 1.0 (See Figure 6.3). However, there are important differences in the fitness function, the selection of the parents in the creation of the offspring, and the selection of the candidate peptides.

### 6.2.1 Selection of the parents in the creation of the offspring

Two parents are selected to create one child, hence, the number of selected parents is twice the number of desired children. The selection of one parent is as follows:

1. $Parent_a$ and $Parent_b$ are selected, by roulette selection, based on their fitness values.
2. The penalization for each parent is obtained. This value is given by the Equation 6.3.

$$
\begin{gathered}
p_{i_{low}}(p) = \begin{cases} 0 & if \quad (x_{i_{min}} - x_i(p)) < 0 \\ \frac{x_{i_{min}} - x_i(p)}{x_{i_{min}}} & otherwise \end{cases} \\
p_{i_{up}}(p) = \begin{cases} 0 & if \quad (x_i(p) - x_{i_{max}}) < 0 \\ \frac{x_i(p) - x_{i_{max}}}{x_{i_{max}}} & otherwise \end{cases} \\
p_4(p) = \begin{cases} 0 & if \quad (x_4(p) - x_{4_{max}}) < 0 \\ \frac{x_4(p) - x_{4_{max}}}{x_{4_{max}}} & otherwise \end{cases} \\
penalization(p) = (\sum_{i=1}^{i=3} (p_{i_{low}}(p)^2 + p_{i_{up}}(p)^2) + p_4(p)^2) * -1
\end{gathered}
\tag{6.3}
$$

where $x_1(p)$, $x_2(p)$, $x_3(p)$ and $x_4(p)$ are the charge, isoelectric point, hydrophobicity and instability index of the peptide $p$, respectively.

3. The best parent is selected based on these criteria:
   - If both parents have penalization equal to zero, then the best parent is the one with the highest fitness value.
   - If only one parent has penalization equal to zero, then this is the best parent.

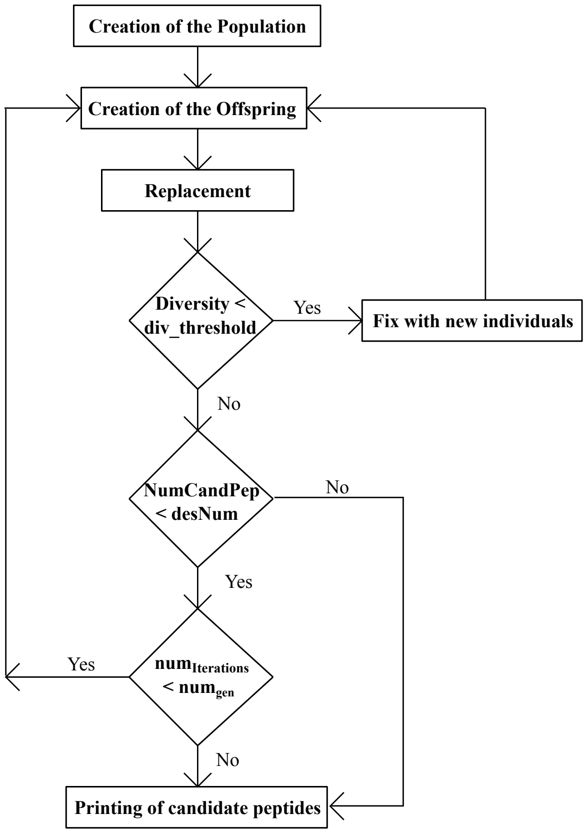

Creation of the Population
Creation of the Offspring
Replacement
Diversity < div_threshold
Yes
Fix with new individuals
No
NumCandPep < desNum
No
Yes
Yes
num$_{Iterations}$ < num$_{gen}$
No
Printing of candidate peptides

Figure 6.3: Workflow of the genetic algorithm PRODEABPs 1.0.

- If none has penalization equal to zero, then, the best parent is the one with the smaller penalization value.

Once all the parents have been selected, the children are created in the same way that they are created in the genetic algorithm DEPRAMPs 1.0.

### 6.2.2 Selection of candidate peptides

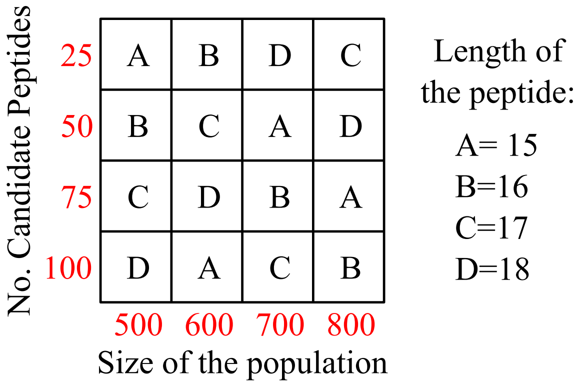
In this genetic algorithm, a candidate peptide is a peptide that complies with the desired probabilities of antibacterial activity and toxicity, i.e. a peptide that complies $h_1$ and $h_2$ constraints. Moreover, if the parameter of hard constraints ($hardCons$) is set to $yes$, then a candidate peptide should also satisfy the constraints $h_3, h_4, h_5$, and $h_6$.

## 6.3 Simulations performed

In order to determine a suitable factor between number of candidate peptides and size of the population, we performed simulations with the number of candidate peptides set to 25, and the size of the population set to 1, 2, ...,10 times the number of candidate peptides. Five simulations were performed for each value. For these simulations, the number of generations was set to 5,000 and the parameter $hardCons$ was set to $yes$. The desired antibacterial probability ($desProb_{abp}$) was set to 0.99 and the desired toxicity probability ($desProb_{tox}$) was set to 0.01.

Additionally, we performed multiple simulations of the genetic algorithm using different values for the parameters: size of the peptides ($desLength$), number of candidate peptides ($desNum$) and size of the population ($sizePop$). For this purpose, we used a Latin Square (see

Figure 6.4) with these parameters. In total, 16 different simulations were designed. Each of the simulations was run 6 times, 3 times with the parameter $hardCons = yes$ i.e., the candidate peptides must fulfill all the constraints, and 3 times with the parameter $hardCons = no$ (the candidate peptides fulfill constraints $h_1$ and $h_2$).



|  | 500 | 600 | 700 | 800 |
|---|---|---|---|---|
| 25 | A | B | D | C |
| 50 | B | C | A | D |
| 75 | C | D | B | A |
| 100 | D | A | C | B |

No. Candidate Peptides
Size of the population
Length of the peptide:
A= 15
B=16
C=17
D=18

Figure 6.4: Latin square of the performed simulations.

## 6.4 Results

### 6.4.1 The size of the population should be about four times the number of candidate peptides

Figure 6.5 shows the average time and number of generations required to design 25 candidate peptides when different sizes of the population are used. For these simulations the parameter $hardCons$ was set to $yes$ due to it is expected that the required time to perform a simulation will be higher when all the constraints must be fulfilled. The simulations that used 25 and 50 peptides in the population were not satisfactory. They could not reach the desired number of candidate peptides in 2000 generations.

In general, an increment on the size of the population might lead to a reduction in the number of generations required to design the candidate peptides. However, this increment might lead to an increment on the computational time of the simulation, due to the fitness evaluation of the peptides. Therefore, according with our results, the size of the population should be set to about four times the number of candidate peptides in order to reduce both, the computational time and the number of generations. Moreover, it can be observed that an increment in the size of the population does not imply a reduction of time or number of generations. This situation might be caused by the random nature of the crossover and mutation processes, or by the initial population of the genetic algorithm.

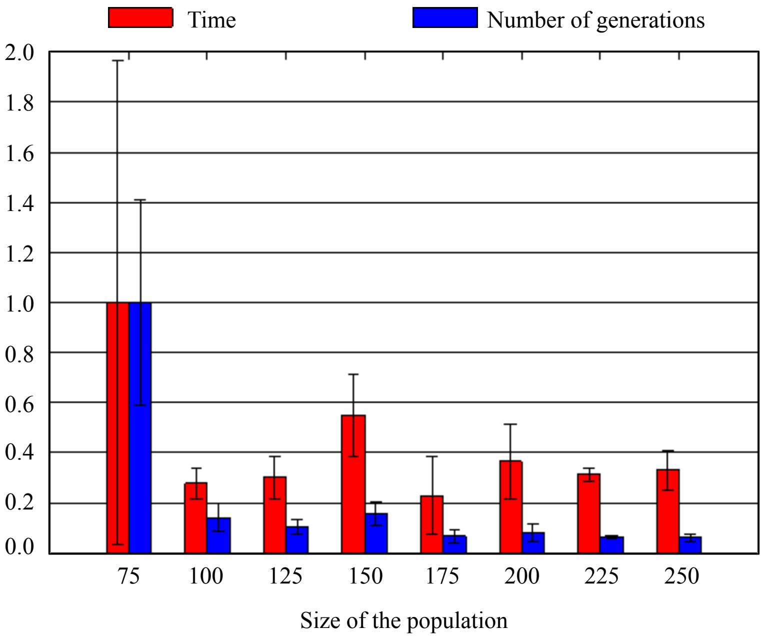


Figure 6.5: Normalized average time and number of generations required to design 25 peptides using different sizes of the population. Time=1 ≡ 6363.983 seconds, and number of generations=1 ≡ 404 generations.

### 6.4.2 The initial population influences in the performance of the genetic algorithm

Figure 6.6 shows the colormaps for the time and number of generations required to obtain the desired number of candidate peptides using different sizes of the population. The maps on the left are obtained when the parameter *hardCons* is set to *no*. The colormaps of the right when the parameter *hardCons* is set to *yes*. The values inside each color box represents the value and the standard deviation of the time for the colormaps in the top of the figure, and the value and the standard deviation of the number of generations for the colormaps in the bottom of the figure.

As mentioned before, it can be observed that an increment in the size of the population does not imply a reduction in time or number of generations. Additionally, there are some cases where the standard deviation is extremely high. Therefore, it is likely that the initial population has a great influence in the performance of the genetic algorithm. If the initial population has a high diversity with some candidate peptides, the simulation will be faster than when the initial population does not contain any candidate peptide. However, due to the diversity of the population is controlled in each generation, the algorithm is able to obtain the desired results with the fixing strategy. In these fixing procedures new peptides are created, and if the number of generations is large enough, the genetic algorithm will be able to achieve the desired number of candidate peptides.

On the other hand, depending on the quality of the initial population, the number of generations seems to be less when the size of the population increases. Having a higher number of peptides increases the chances of finding different candidate peptides in each generation. However, it is important to consider that a larger population also implies more computational time due to the calculation of the fitness values. Therefore, we recommend that the size of the population would be around four times the number of candidate peptides.

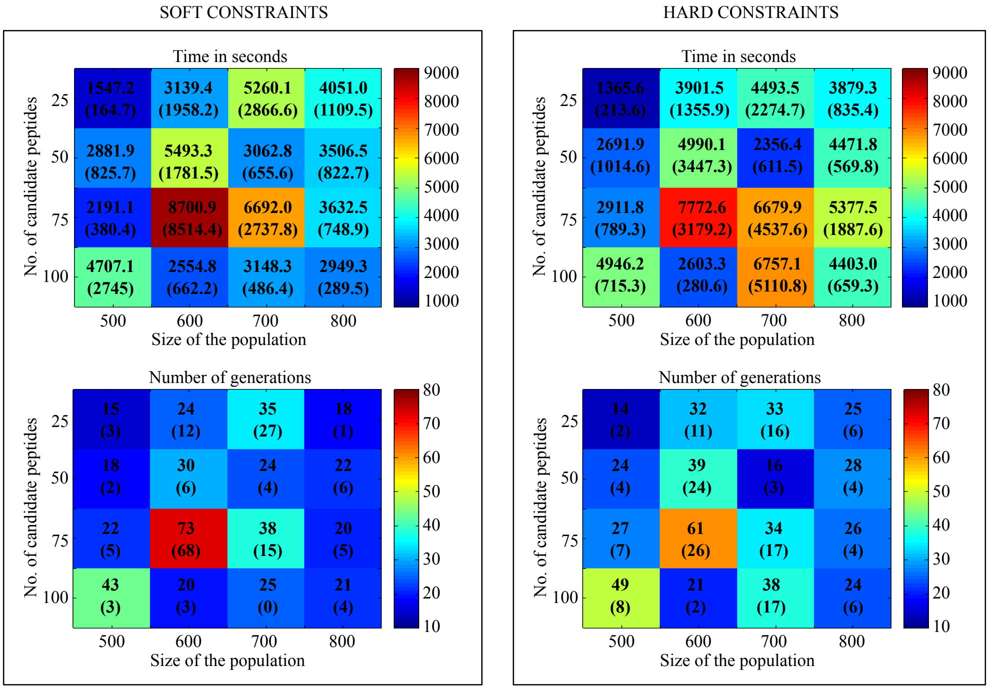


Figure 6.6: Time and number of generations required to obtain the desired number of candidate peptides when different sizes of the population are used. The values inside each color box represents the standard deviation. The simulations were performed in a computer with 4 GB of RAM memory, and an Intel® Core$^{TM}$ i5-480M processor with 2 cores and 2.67GHz of frequency.

### 6.4.3 The designed peptides comply with the established constraints

Two groups of candidate peptides were created. The first group named *soft constraints* that contains all the candidate peptides that were obtained during the simulations using the parameter $hardCons = no$. The second group comprehends those peptides that were designed using the parameter $hardCons = yes$. This last group is named *hard constraints*. The physicochemical properties of the designed peptides were calculated and their corresponding boxplots are shown in Figure 6.7.

All the designed peptides, from both groups, comply with the restrictions for both antibacterial and toxicity probability set by the user. Additionally, all the peptides of the group *hard constraints* also comply with the constraints for the physicochemical properties. Similarly, most of the peptides from the group *soft constraints* fulfill these constraints too, because the selection of the parents (in each generation) involves the evaluation of these constraints.

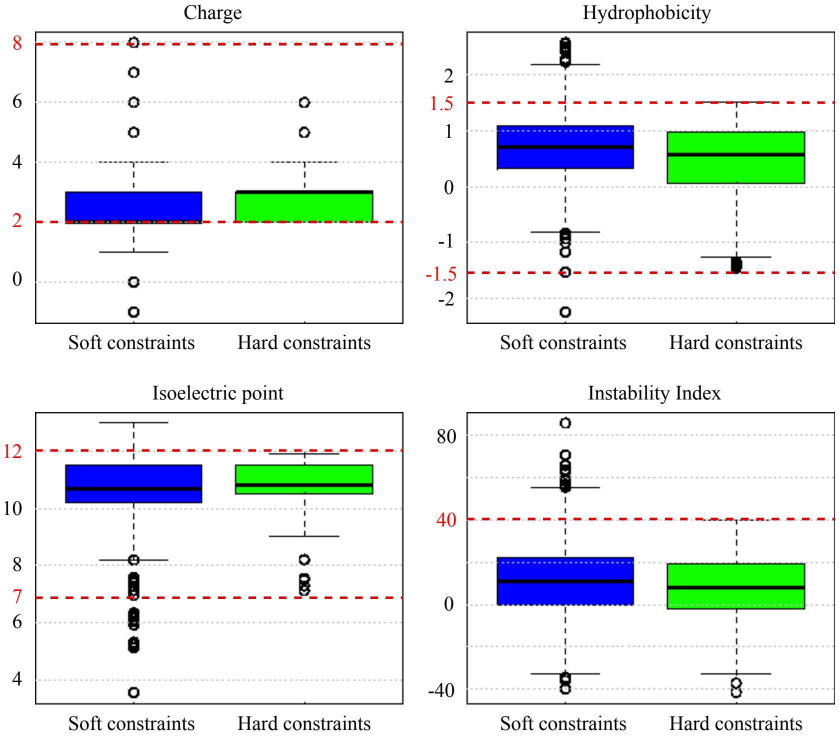


Figure 6.7: Physicochemical properties of the designed peptides.

## 6.5 Posterior analysis of the peptides

As in the strategy 1, the peptides that have been designed with the genetic algorithm PRODEABPs 1.0, should be analyzed with the PEP-FOLD 2.0 tool [82] (http://mobyle.rpbs.univ-paris-diderot.fr/cgi-bin/portal.py#forms::PEP-FOLD) in order to determine their secondary structure. The peptides that display a predicted alpha-helix secondary structure should be selected for the next step. Figure 6.8 shows the primary and secondary structure of two peptides designed with the genetic algorithm PRODEABPs 1.0., and that posses an alpha-helix secondary structure according with PEP-FOLD 2.0 tool.

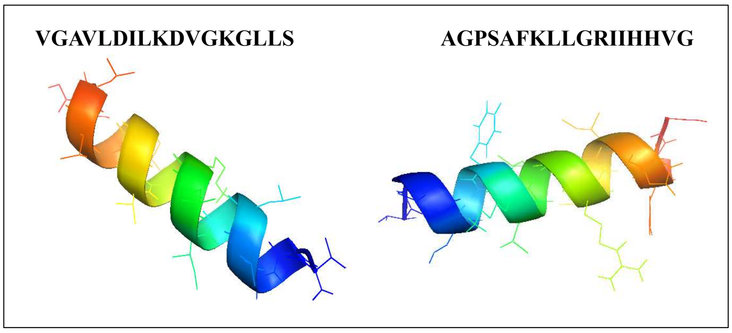


Figure 6.8: Primary and secondary structure of two peptides designed with the genetic algorithm PRODEABPs 1.0.

Later, the remaining peptides should be analyzed in their primary structure, i.e., multiple conditions about the amino acid content of the peptides can be established in the selection of the peptides. For example, the number of cysteines (C) in the peptide could be set to zero due to the difficulty that they represent in the synthesis process of the peptides. Therefore, it is recommended to consider the properties that might affect the antimicrobial activity and the toxicity of the peptides that were presented in Chapter 2 (See section 2.2.5). Finally, it is recommended to select those peptides that do not present an homology higher than 50% with the antimicrobial peptides reported in the APD and CAMP database. This step is required in case that the designed peptide will be patented.

## 6.6 Conclusions

The proposed genetic algorithm allows the design of new potential non-toxic antibacterial peptides. The speed of the algorithm depends on how good is the initial population. However, if the number of generations is large enough, the genetic algorithm will be able to obtain the desired results due to the fixing procedure that allows the insertion of new peptides in the population. This diversity control is a fundamental part in the development of the genetic algorithm. A good diversity in the population helps to find the desired number of candidate peptides.

The fact that the initial population leads to different results is not a disadvantage for us, since the main goal of the algorithm is to help in the creation of a large library of potential non-toxic antibacterial peptides. In this sense, we are not interested on finding a global minimum, the purpose is to find multiple local minimums. These minimums corresponds to those peptides that comply with the desired constraints, and therefore, peptides that are likely to be effective in experimental tests.

The design of new potential non-toxic antibacterial peptides is more efficient using the strategy 2 than the strategy 1, because the genetic algorithm gives as results those peptides that are already classified as antibacterial peptides. Moreover, this strategy involves the decision of a toxic peptides classifier which helps in the selection of peptides that not only would be effective against bacteria, but also peptides that would not be toxic for human beings.

The main differences between the two strategies are:

- The optimization function used in the genetic algorithms.
- The selection of the parents in the strategy 2 involves a competition between peptides, where the best ones, in terms of penalization values, are selected.
- The genetic algorithm of the strategy 2 allows to design peptides that are not forced to fulfill the constraints related with the physicochemical properties, although, the suggested ranges are considered in the design of the peptides.
- The classifiers used in the strategy 2 were created with a more complete dataset. Therefore, if strategy 1 is going to be used, we recommend to use the dataset used in the strategy 2. Moreover, we also recommend to use the toxic peptides classifier in the strategy 1.

After the generation of multiple candidate non-toxic antibacterial peptides, the selection of those that are going to be synthesized is a difficult step, since all of the designed peptides are equally likely to be effective. Therefore, additional filters are required and further studies have to be developed in order to improve the prediction of new potential antibacterial peptides. Important aspects such as the stability of the peptides and the mechanism of action could enrich the genetic algorithm designed in this thesis.

# Chapter 7

# Conclusions and future works

## 7.1 Summary

This thesis helps to overcome one of the limitations in the rational design of antibacterial peptides since it allows the design of new peptides that are likely to be non-toxic and to possess antibacterial activity. Among the main contributions of this thesis we can find the proposed methodology, the designed peptides that have shown a competitive antibacterial activity and a low cytotoxicity, the softwares DEPRAMPs 1.0 and PRODEABPs 1.0, and journal articles.

The methodology proposed in this thesis allows the design of peptides using the probabilities for antibacterial activity and toxicity of a peptide at the same time. Since the design of the classifiers was focused on the reduction of the false positive rate, it is likely that this thesis helps to the reduction of time and costs associated with the synthesis of peptides that do not possess the desired properties. Moreover, the results associated with this thesis are of high importance for the research groups CEMOS and GIBIM. The bioinformatics research line was strengthened with the development of this thesis and the support given to undergrad and master thesis in this area. Moreover, thesis of undergrad, master and other Ph.D. students used peptides designed in this thesis. Additionally, the proposed methodology is in the process of being patented since some pharmaceutical companies have shown interest in the work that has been performed in this thesis. Furthermore, two peptides designed in this thesis are in the process of being patented and it is expected that they could become potential new antibiotics in a near future (around 10 years), and in a long term might save thousands of lives.

On the other hand, although the antimicrobial peptides have been studied in the last decades, in our personal opinion, the area of the rational design of antibacterial peptides is in its first stages of development. We believed that the design of new effective antibacterial peptides requires the integration of multiple aspects such as the toxicity, stability, mechanism of action, and selectivity against gram-positive or gram-negative bacteria, among others. Additionally, one of the main limitations in this field is the inability to describe their mechanism of action in physical-chemical terms and the lack of explicit, molecular, structure-function relationships. Therefore, we believed that future works in the rational design of antibacterial peptides are required urgently in order to find a potential solution to the threat of multidrug-resistant bacteria.

Finally, we encourage the scientific community to increase the research in the search of new potential antibiotics. Scientific research should not be focused in the return of the investment but on the real needs of human society. In this sense, we believed that this thesis is an important step in the search of effective antibacterial peptides that could become new antibiotics in the future.

## 7.2 Products of this thesis

### 7.2.1 Journal Articles

**Published**

- Paola Rondón-Villarreal, Daniel A. Sierra, Rodrigo Torres. Machine Learning in the Rational Design of Antimicrobial Peptides. *Current Computer Aided-Drug Design*, ISSN 1573-4099, Vol 10, Number 3, 2014, Pag 183.

- Daniel Osorio, Paola Rondón-Villarreal, Rodrigo Torres. Peptides: A Package for Data Mining of Antimicrobial Peptides. *The R Journal*, ISSN 2073-4859, Vol 7, Number 1, 2015, Pag 4.

- Paola Rondón-Villarreal, Daniel A. Sierra, Rodrigo Torres. Classification of antimicrobial peptides by using the p-spectrum kernel and support vector machines. *Advances in Computational Biology. Advances in Intelligent Systems and Computing*, ISSN 2194-5365, Vol 232, Pag 155.

- Daniel Osorio, Paola Rondón-Villarreal, Rodrigo Torres. Stability analysis of antimicrobial peptides in solvation conditions by molecular dynamics. *Advances in Computational Biology. Advances in Intelligent Systems and Computing*, ISSN 2194-5365, Vol 232, 2014, Pag 127.

- Jennifer Ruiz, Jhon Calderon, Paola Rondón-Villarreal, Rodrigo Torres. Analysis of structure and hemolytic activity relationships of antimicrobial peptides (AMPs). *Advances in Computational Biology. Advances in Intelligent Systems and Computing*, ISSN 2194-5365, Vol 232, 2014, Pag 253.

- Yuly Andrea Prada, Fanny Guzmán, Paola Rondón-Villarreal, Patricia Escobar, Claudia Ortiz, Daniel Sierra, Rodrigo Torres, Enrique Mejía. A new synthetic peptide with antibacterial potential *in vitro* against *Escherichia coli* O157:H7 and Methicilin resistant *Staphylococcus aureus* (MRSA). *Probiotics and Antimicrobial Proteins*, First online: 15 June 2016.

**In revision process**

- Paola Rondón-Villarreal, Marcel J.T. Reinders, Francy Camacho, Rodrigo Torres, Daniel A. Sierra, David M.J. Tax. Implications of the Optimization Criterion When Creating an Antibacterial Peptide Classifier.

- Jenniffer Cruz, Paola Rondón-Villarreal, Claudia Ortiz, Fanny Guzmán, Claudio Alvarez, Roberto Fernández-Lafuente, Luis Rivas, María Ángeles Ábengozar, Fernando Albericio, Mauricio Urquiza, Daniel A. Sierra, Rodrigo G. Torres. Rational Design and Evaluation of Novel Antimicrobial Peptides Bioactives Against *Escherichia coli* O157:H7, methicillin-resistant *Staphylococcus aureus* and *Pseudomonas aeruginosa.*

### 7.2.2 Participation in academic and scientific events

- Finalist in the Falling Walls Lab Berlin. Presentation title: Breaking the wall of antibiotic resistance. Berlin, Germany 2015.

- Oral Presentations at the 2$^{nd}$ Colombian Congress on Computational Biology and Bioinformatics CCBCOL. Manizales, Colombia 2013:

  - Classification of antimicrobial peptides by using the p-spectrum kernel and support vector machines. *Advances in Computational Biology.* Presented by Paola Rondón-Villarreal.
  - Stability analysis of antimicrobial peptides in solvation conditions by molecular dynamics. *Advances in Computational Biology.* Presented by Daniel Osorio.
  - Analysis of structure and hemolytic activity relationships of antimicrobial peptides (AMPs). Presented by Jennifer Ruiz.

- Poster at the 2015 IEEE Thirty Fifth Central American and Panama Convention (CONCAPAN XXXV). C. Lastre-Domínguez, P. Rondón-Villarreal, D.A. Sierra, Classification of Peptides Using Ensembles by Applying Different Strategies that Deal with Imbalanced Data and Combination Rules. Tegucigalpa, Honduras 2015. Presented by Carlos Lastre.

### 7.2.3 Support to final degree projects

- Master thesis. *Clasificación de Péptidos a partir de diferentes métodos y estrategias de ensamble de clasificadores en condición desbalanceada.* Author: Carlos Mauricio Lastre Domínguez. Advisor: Daniel Alfonso Sierra Bueno. Co-advisor: Nydia Paola Rondón Villarreal.

- Undergrad thesis. *Análisis del potencial de disrupción de membranas y estabilidad de péptidos catiónicos antimicrobianos por simulaciones de dinámica molecular.* Author: Daniel Camilo Osorio Hurtado. Advisor: Rodrigo Gonzalo Torres Sáez. Co-advisor: Nydia Paola Rondón Villarreal.

- Undergrad thesis. *Sistema de clasificación de péptidos antibacterianos utilizando máquinas de soporte vectorial.* Author: Francy Liliana Camacho Urrea. Advisor: Lola Xiomara Bautista. Co-advisors: Nydia Paola Rondón Villarreal, Rodrigo Gonzalo Torres Sáez, Daniel Alfonso Sierra Bueno.

- Peptides that were designed in this thesis have been used in the PhD thesis of Jennifer Cruz, master thesis of Yuly Andrea Prada, Marlon Yesid Cáceres, and Andrés Mauricio Castañeda. These students are members of the research group GIBIM at the Chemistry School - Universidad Industrial de Santander.

### 7.2.4 Potential non-toxic antibacterial peptides

- Peptides GIBIM-P6 and GIBIM-P5F8W that are in the process of being patented.

### 7.2.5 Softwares

- Software DEPRAMPs 1.0 to design peptides with specific physicochemical properties.

- Software PRODEABPs 1.0 to design non-toxic antibacterial peptides.

## 7.3 Future works

Once a library of potential non-toxic antibacterial peptides has been created, an important step is the selection of the best peptides to synthesize. This selection process requires additional information of the peptides in order to determine which ones are the best option. Therefore, future works should be done in order to improve the design of potential new antibacterial peptides. These are some of the improvements that could be performed to the proposed genetic algorithm:

- To include the prediction of the secondary structure of the peptides in the design process. As a result, it would be possible to design peptides that will possess a predicted alpha-helix structure, reducing the time spent in the posterior analysis of the designed peptides.

- To include the probability that a peptide will be stable. In this sense, it is important to design peptides that will be stable and that would not be susceptible to degradation at a systemic level.

- To create classifiers that allow the differentiation of peptides that are active against gram-positive, gram-negative or both types of bacteria.

- To create classifiers that allow the prediction of the most likely mechanism of action for a peptide.

- To include more physicochemical properties related with the antibacterial activity, toxicity and stability of the peptides.

- To include molecular dynamics simulations in the posterior analysis of the best candidate peptides before the synthesis process.

# References

[1] Amábile-Cuevas, C. F. . Global Perspectives of Antibiotic Resistance. In Sosa, A. D. J. , Byarugaba, D. K. , Amábile-Cuevas, C. F. , Hsueh, P.-R. , Kariuki, S. , and Okeke, I. N. , editors, *Antimicrobial Resistance in Developing Countries*, chapter 1, pages 3– 13. Springer New York, New York, 2010.

[2] Infectious Diseases Society of America. The 10 x '20 Initiative: pursuing a global commitment to develop 10 new antibacterial drugs by 2020. *Clinical infectious diseases*, 50(8):1081–1083, 2010.

[3] Projan, S. J. and Bradford, P. A. . Late stage antibacterial drugs in the clinical pipeline. *Current Opinion in Microbiology*, 10(5):441–446, 2007.

[4] HHS, U. D. o. H. , Services, H. , CDC, C. f. D. C. , and Prevention. Antibiotic resistance threats in the United States, 2013. Technical report, USA, 2013.

[5] WHO, W. H. O. . Antimicrobial Resistance Global Report on Surveillance. Technical report, France, 2014.

[6] ECDC, E. C. f. D. P. and Control. Antimicrobial resistance surveillance in Europe 2014. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net). Technical report, Stockholm, 2015.

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

[8] Bettiol, E. , Wetherington, J. D. , Schmitt, N. , and Harbarth, S. . Challenges and Solutions for Clinical Development of New Antibacterial Agents: Results of a Survey among Pharmaceutical Industry Professionals. *Antimicrobial agents and chemotherapy*, 59(7):3695–3699, 2015.

[9] Roca, I. , Akova, M. , Baquero, F. , Carlet, J. , Cavaleri, M. , Coenen, S. , Cohen, J. , Findlay, D. , Gyssens, I. , Heure, O. , Kahlmeter, G. , Kruse, H. , Laxminarayan, R. , Liébana, E. , López-Cerero, L. , MacGowan, A. , Martins, M. , Rodríguez-Baño, J. , Rolain, J.-M. , Segovia, C. , Sigauque, B. , Taconelli, E. , Wellington, E. , and Vila, J. . The global threat of antimicrobial resistance: Science for intervention. *New Microbes and New Infections*, 6:22–29, 2015.

[10] Pharmaceutical Research and Manufacturers of America. *2015 Biopharmaceutical Research Industry Profile*. PhRMA, Washington, DC, 2015.

[11] Projan, S. J. . Why is big Pharma getting out of antibacterial drug discovery? *Current Opinion in Microbiology*, 6(5):427–430, 2003.

[12] Poupard, J. A. . Is The Pharmaceutical Industry Responding to the Challenge of Increasing Bacterial Resistance ? *Clinical Microbiology Newsletter*, 28(2):13–15, 2005.

[13] Williams, K. J. and Bax, R. P. . Challenges in developing new antibacterial drugs. *Current Opinion in Investigational Drugs*, 10(2):157–163, 2009.

[14] Brown, E. D. and Wright, G. D. . Antibacterial drug discovery in the resistance era. *Nature*, 529:336–343, 2016.

[15] Boucher, H. W. , Talbot, G. H. , Benjamin Jr, D. K. , Bradley, J. , Guidos, R. J. , Jones, R. N. , Murray, B. E. , Bonomo, R. A. , and Gilbert, D. . 10 x '20 Progress-Development of New Drugs Active Against Gram-Negative Bacilli: An Update From the Infectious Diseases Society of America. *Clinical Infectious Diseases*, 56(12):1685–1694, 2013.

[16] Giuliani, A. , Pirri, G. , and Nicoletto, S. F. . Antimicrobial peptides: an overview of a promising class of therapeutics. *Central European Journal of Biology*, 2(1):1–33, 2007.

[17] Rondón-Villarreal, P. , Sierra, D. A. , and Torres, R. . Machine Learning in the Rational Design of Antimicrobial Peptides. *Current Computer-Aided Drug Design*, 10:183–190, 2014.

[18] Rondón-Villarreal, P. , Sierra, D. A. , and Torres, R. . Classification of antimicrobial peptides by using the p -spectrum kernel and support vector machines. *Advances in Computational Biology. Advances in Intelligent Systems and Computing*, 232:155–161, 2014.

[19] Rondón-Villarreal, P. , Reinders, M. J. T. , Camacho, F. , Torres, R. , Sierra, D. A. , and Tax, D. M. . Implications of the Optimization Criterion when Creating an Antibacterial Peptide Classifier. *Paper in revision.*

[20] Wang, G. . Database-Guided Discovery of Potent Peptides to Combat HIV-1 or Superbugs. *Pharmaceuticals*, 6(6):728–758, 2013.

[21] Baltzer, S. A. and Brown, M. H. . Antimicrobial Peptides - Promising Alternatives to Conventional Antibiotics. *Journal of Molecular Microbiology and Biotechnology*, 20(4):228–235, 2011.

[22] Cherkasov, A. , Hilpert, K. , Jenssen, H. v. , Fjell, C. D. , Waldbrook, M. , Mullaly, S. C. , Volkmer, R. , and Hancock, R. E. W. . Use of Artificial Intelligence in the Design of Small Peptide Antibiotics Effective against a Broad Spectrum of Highly Antibiotic-Resistant Superbugs. *ACS Chemical Biology*, 4(1):65–74, 2009.

[23] Hancock, R. E. W. and Sahl, H.-G. . Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nature Biotechnology*, 24(12):1551–1557, 2006.

[24] Tan, J. , Huang, J. , Huang, Y. , and Chen, Y. . Effects of single amino acid substitution on the biophysical properties and biological activities of an amphipathic $\alpha$-helical antibacterial peptide against Gram-negative bacteria. *Molecules*, 19(8):10803–10817, 2014.

[25] Pasupuleti, M. , Schmidtchen, A. , and Malmsten, M. . Antimicrobial peptides: key components of the innate immune system. *Critical Reviews in Biotechnology*, 32(2):143–171, 2011.

[26] Takahashi, D. , Shukla, S. K. , Prakash, O. , and Zhang, G. . Structural determinants of host defense peptides for antimicrobial activity and target cell selectivity. *Biochimie*, 92(9):1236–1241, 2010.

[27] Okorochenkov, S. A. , Zheltukhina, G. A. , and Nebolsin, V. E. . Antimicrobial peptides: the mode of action and perspectives of practical application. *Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry*, 5:95–102, 2011.

[28] Gupta, S. , Kapoor, P. , Chaudhary, K. , Gautam, A. , Kumar, R. , Consortium, O. S. D. D. , and Raghava, G. P. S. . In silico approach for predicting toxicity of peptides and proteins. *PloS one*, 8(9):e73957–e73966, 2013.

[29] Kim, H. , Jang, J. H. , Kim, S. C. , and Cho, J. H. . De novo generation of short antimicrobial peptides with enhanced stability and cell specificity. *The Journal of antimicrobial chemotherapy*, 69(1):121–132, 2014.

[30] Nelson, D. L. and Cox, M. M. . *Lehninger Principles of Biochemistry Fourth Edition*. W. H. Freeman, 2004.

[31] Marr, A. K. , Gooderham, W. J. , and Hancock, R. E. . Antibacterial peptides for therapeutic use: obstacles and realistic outlook. *Current Opinion in Pharmacology*, 6:468–472, 2006.

[32] Tossi, A. and Sandri, L. . Molecular Diversity in Gene-Encoded, Cationic Antimicrobial Polypeptides. *Current Pharmaceutical Design*, 8(9):743–761, 2002.

[33] Wade, D. and Englund, J. . Synthetic antibiotic peptides database. *Protein and peptide letters*, 9(1):53–57, 2002.

[34] Whitmore, L. and Wallace, B. A. . The Peptaibol Database: a database for sequences and structures of naturally occurring peptaibols. *Nucleic Acids Research*, 32(Database issue):D593–D594, 2004.

[35] Wang, Z. and Wang, G. . APD: the Antimicrobial Peptide Database. *Nucleic Acids Research*, 32(Database issue):D590–D592, 2004.

[36] Wang, G. , Li, X. , and Wang, Z. . APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Research*, 37(Database issue):D933–D937, 2009.

[37] Sundararajan, V. S. , Gabere, M. N. , Pretorius, A. , Adam, S. , Christoffels, A. , Lehväslaiho, M. , Archer, J. A. C. , and Bajic, V. B. . DAMPD: a manually curated antimicrobial peptide database. *Nucleic Acids Research*, 40(Database issue):D1108–D1112, 2012.

[38] Gueguen, Y. , Garnier, J. , Robert, L. , Lefranc, M.-P. , Mougenot, I. , De Lorgeril, J. , Janech, M. , Gross, P. S. , Warr, G. W. , Cuthbertson, B. , Barracco, M. A. , Bulet, P. , Aumelas, A. , Yang, Y. , Bo, D. , Xiang, J. , Tassanakajon, A. , Piquemal, D. , and Bachère, E. . PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. *Developmental & Comparative Immunology*, 30(3):283–288, 2006.

[39] Wang, C. K. L. , Kaas, Q. , Chiche, L. , and Craik, D. J. . CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Research*, 36(Database issue):D206–D2010, 2007.

[40] Jong, A. , de, Heel, A. J. , van, Kok, J. , and Kuipers, O. P. . BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Research*, 38(Web Server issue):W647–W651, 2010.

[41] Fjell, C. D. , Hancock, R. E. W. , and Cherkasov, A. . AMPer: a database and an automated discovery tool for antimicrobial peptides. *Bioinformatics*, 23(9):1148–1155, 2007.

[42] Hammami, R. , Zouhir, A. , Hamida, J. B. , and Fliss, I. . BACTIBASE: a new web-accessible database for bacteriocin characterization. *BMC Microbiology*, 7:89–94, 2007.

[43] Hammami, R. , Zouhir, A. , Le Lay, C. , Hamida, J. B. , and Fliss, I. . BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiology*, 10(22):1–5, 2010.

[44] Seebah, S. , Suresh, A. , Zhuo, S. , Choong, Y. H. , Chua, H. , Chuon, D. , Beuerman, R. , and Verma, C. . Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Research*, 35(Database issue):D265–D268, 2007.

[45] Li, Y. and Chen, Z. . RAPD: a database of recombinantly-produced antimicrobial peptides. *FEMS Microbiology Letters*, 289(2):126–129, 2008.

[46] Hammami, R. , Hamida, J. B. , Vergoten, G. , and Fliss, I. . PhytAMP: a database dedicated to antimicrobial plant peptides. *Nucleic Acids Research*, 37(Database issue):D963–D968, 2009.

[47] Thomas, S. , Karnik, S. , Barai, R. S. , Jayaraman, V. K. , and Idicula-Thomas, S. . CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Research*, 38(Database issue):D774–D780, 2010.

[48] Piotto, S. P. , Sessa, L. , Concilio, S. , and Iannelli, P. . YADAMP: yet another database of antimicrobial peptides. *International Journal of Antimicrobial Agents*, 39(4):346–351, 2012.

[49] Novković, M. , Simunić, J. , Bojović, V. , Tossi, A. , and Juretić, D. . DADP: the Database of Anuran Defense Peptides. *Bioinformatics*, 28(10):1406–1407, 2012.

[50] Li, J. , Qu, X. , He, X. , Duan, L. , Wu, G. , Bi, D. , Deng, Z. , Liu, W. , and Ou, H.-Y. . ThioFinder : A Web-Based Tool for the Identification of Thiopeptide Gene Clusters in DNA Sequences. *PloS one*, 7(9):e45878–e45886, 2012.

[51] Wu, H. , Lu, H. , Huang, J. , Li, G. , and Huang, Q. . EnzyBase: a novel database for enzybiotic studies. *BMC Microbiology*, 12:54–58, 2012.

[52] Zhao, X. , Wu, H. , Lu, H. , Li, G. , and Huang, Q. . LAMP: A Database Linking Antimicrobial Peptides. *PloS one*, 8(6):e66557–e66562, 2013.

[53] Gogoladze, G. , Grigolava, M. , Vishnepolsky, B. , Chubinidze, M. , Duroux, P. , Lefranc, M.-P. , and Pirtskhalava, M. . DBAASP: database of antimicrobial activity and structure of peptides. *FEMS microbiology letters*, 357(1):63–68, 2014.

[54] Di Luca, M. , Maccari, G. , Maisetta, G. , and Batoni, G. . BaAMPs: the database of biofilm-active antimicrobial peptides. *Biofouling*, 31(2):193–199, 2015.

[55] Wang, G. . Improved Methods for Classification, Prediction, and Design of Antimicrobial Peptides. In *Computational Peptidology*, volume 1268, pages 43–66. 2015.

[56] Dudek, A. Z. , Arodz, T. , and Gálvez, J. . Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial Chemistry & High Throughput Screening*, 9(3):213–228, 2006.

[57] Cherkasov, A. . 'Inductive' Descriptors : 10 Successful Years in QSAR. *Current Computer-Aided Drug Design*, 1(1):21–42, 2005.

[58] Hilpert, K. , Elliott, M. R. , Volkmer-Engert, R. , Henklein, P. , Donini, O. , Zhou, Q. , Winkler, D. F. H. , and Hancock, R. E. W. . Sequence requirements and an optimization strategy for short antimicrobial peptides. *Chemistry & Biology*, 13(10):1101–1107, 2006.

[59] Jenssen, H. v. , Lejon, T. , Hilpert, K. , Fjell, C. D. , Cherkasov, A. , and Hancock, R. E. W. . Evaluating different descriptors for model design of antimicrobial peptides with enhanced activity toward P. aeruginosa. *Chemical Biology Drug Design*, 70(2):134–142, 2007.

[60] Jenssen, H. , Fjell, C. D. , Cherkasov, A. , and Hancock, R. E. W. . QSAR modeling and computer-aided design of antimicrobial peptides. *Journal of Peptide Science*, 14(1):110–114, 2008.

[61] Avram, S. , Duda-Seiman, D. , Borcan, F. , Radu, B. , Duda-Seiman, C. , and Mihailescu, D. . Evaluation of Antimicrobial Activity of New Mastoparan Derivatives Using QSAR and Computational Mutagenesis. *International Journal of Peptide Research and Therapeutics*, 17(1):7–17, 2011.

[62] Zhu, X. , Ma, Z. , Wang, J. , Chou, S. , and Shan, A. . Importance of Tryptophan in Transforming an Amphipathic Peptide into a Pseudomonas aeruginosa-Targeted Antimicrobial Peptide. *PloS one*, 9(12):e114605–e114623, 2014.

[63] Wu, X. , Wang, Z. , Li, X. , Fan, Y. , He, G. , Wan, Y. , Yu, C. , Tang, J. , Li, M. , Zhang, X. , Zhang, H. , Xiang, R. , Pan, Y. , Liu, Y. , Lu, L. , and Yang, L. . In vitro and in vivo activities of antimicrobial peptides developed using an amino acid-based activity prediction method. *Antimicrobial agents and chemotherapy*, 58(9):5342–5349, 2014.

[64] Maccari, G. , Di Luca, M. , Nifosí, R. , Cardarelli, F. , Signore, G. , Boccardi, C. , and Bifone, A. . Antimicrobial Peptides Design by Evolutionary Multiobjective Optimization. *PLoS Computational Biology*, 9(9):e1003212–e1003223, 2013.

[65] Fjell, C. D. , Jenssen, H. v. , Cheung, W. A. , Hancock, R. E. W. , and Cherkasov, A. . Optimization of antibacterial peptides by genetic algorithms and cheminformatics. *Chemical biology & drug design*, 77(1):48–56, 2011.

[66] Juretić, D. , Vukicević, D. , Ilić, N. , Antcheva, N. , and Tossi, A. . Computational design of highly selective antimicrobial peptides. *Journal of Chemical Information and Modeling*, 49(12):2873–2882, 2009.

[67] Avram, S. , Mihailescu, D. , Borcan, F. , and Milac, A.-L. . Prediction of improved antimicrobial mastoparan derivatives by 3D-QSAR-CoMSIA/CoMFA and computational mutagenesis. *Monatshefte für Chemie - Chemical Monthly*, 143:535–543, 2012.

[68] Torrent, M. , Andreu, D. , Nogués, V. M. , and Boix, E. . Connecting Peptide Physicochemical and Antimicrobial Properties by a Rational Prediction Model. *PLoS one*, 6(2):e16968–e16975, 2011.

[69] Taboureau, O. , Olsen, O. H. , Nielsen, J. D. , Raventos, D. , Mygind, P. H. , and Kristensen, H.-H. . Design of novispirin antimicrobial peptides by quantitative structure-activity relationship. *Chemical Biology & Drug Design*, 68(1):48–57, 2006.

[70] Fjell, C. D. , Jenssen, H. v. , Hilpert, K. , Cheung, W. A. , Panté, N. , Hancock, R. E. W. , and Cherkasov, A. . Identification of novel antibacterial peptides by chemoinformatics and machine learning. *Journal of Medicinal Chemistry*, 52(7):2006–2015, 2009.

[71] Xiao, X. , Wang, P. , Lin, W.-Z. , Jia, J.-H. , and Chou, K.-C. . iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical biochemistry*, 436(2):168–177, 2013.

[72] Lata, S. , Sharma, B. , and Raghava, G. . Analysis and prediction of antibacterial peptides. *BMC Bioinformatics*, 8:263–272, 2007.

[73] Wang, P. , Hu, L. , Liu, G. , Jiang, N. , Chen, X. , Xu, J. , Zheng, W. , Li, L. , Tan, M. , Chen, Z. , Song, H. , Cai, Y.-D. , and Chou, K.-C. . Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PloS one*, 6(4):e18476–e18484, 2011.

[74] Porto, W. F. , Fernandes, F. C. , and Franco, O. L. . An SVM Model Based on Physicochemical Properties to Predict Antimicrobial Activity from Protein Sequences with Cysteine Knot Motifs. *Lectures Notes in Computer Science*, 6268:59–62, 2010.

[75] Lata, S. , Mishra, N. K. , and Raghava, G. P. . AntiBP2: improved version of antibacterial peptide prediction. *BMC Bioinformatics*, 11(Suppl 1):S19–S25, 2010.

[76] Bahar, A. A. and Ren, D. . Antimicrobial peptides. *Pharmaceuticals*, 6(12):1543–1575, 2013.

[77] Maccari, G. , Luca, M. D. , and Nifosì, R. . In Silico Design of Antimicrobial Peptides. In Zhou, P. and Huang, J. , editors, *Computational Peptidology*, volume 1268 of *Methods in Molecular Biology*, pages 195–219. Springer New York, New York, NY, 2015.

[78] Boman, H. G. . Antibacterial peptides: basic facts and emerging concepts. *Journal of Internal Medicine*, 254(3):197–215, 2003.

[79] Gupta, S. , Kapoor, P. , Chaudhary, K. , Gautam, A. , Kumar, R. , and Raghava, G. P. S. . Peptide Toxicity Prediction. In *Computational Peptidology*, volume 1268, pages 143–157. 2015.

[80] Consortium, T. U. . UniProt: a hub for protein information. *Nucleic Acids Research*, 43(Database issue):D204–D212, 2015.

[81] Wimley, W. C. and Hristova, K. . Antimicrobial peptides: successes, challenges and unanswered questions. *The Journal of membrane biology*, 239(1-2):27–34, 2011.

[82] Shen, Y. , Maupetit, J. , Derreumaux, P. , and Tufféry, P. . Improved PEP-FOLD approach for peptide and miniprotein structure prediction. *Journal of Chemical Theory and Computation*, 10(10):4745–4758, 2014.

[83] Chen, W. and Luo, L. . Classification of antimicrobial peptide using diversity measure with quadratic discriminant analysis. *Journal of Microbiological Methods*, 78(1):94–96, 2009.

[84] Langham, A. A. , Khandelia, H. , Schuster, B. , Waring, A. J. , Lehrer, R. I. , and Kaznessis, Y. N. . Correlation between simulated physicochemical properties and hemolycity of protegrin-like antimicrobial peptides: Predicting experimental toxicity. *Peptides*, 29(7):1085–1093, 2008.

[85] Mikut, R. and Hilpert, K. . Interpretable Features for the Activity Prediction of Short Antimicrobial Peptides Using Fuzzy Logic. *International Journal of Peptide Research and Therapeutics*, 15(2):129–137, 2009.

[86] Khosravian, M. , Faramarzi, F. K. , Beigi, M. M. , Behbahani, M. , and Mohabatkar, H. . Predicting Antibacterial Peptides by the Concept of Chou's Pseudo-amino Acid Composition and Machine Learning Methods. *Protein & Peptide Letters*, 20(2):180–186, 2013.

[87] Joseph, S. , Karnik, S. , Nilawe, P. , Jayaraman, V. K. , and Idicula-Thomas, S. . ClassAMP: a prediction tool for classification of antimicrobial peptides. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 9(5):1535–1538, 2012.

[88] He, H. and Garcia, E. A. . Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[89] Estabrooks, A. , Jo, T. , and Japkowicz, N. . A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Computational Intelligence*, 20(1):18–36, 2004.

[90] Galar, M. , Fernández, A. , Barrenechea, E. , Bustince, H. , and Herrera, F. . A review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE transactions on systems, man, and cybernetics. Part C, Applications and Reviews*, 42(4):463–484, 2012.

[91] Cristianini, N. , Shawe-Taylor, J. , and Saunders, C. . Kernel Methods: A Paradigm for Pattern Analysis. In Camps-Valls, G. , Rojo-Álvarez, J. L. , and Martínez-Ramón, M. , editors, *Kernel Methods in Bioengineering, Signal and Image Processing*, pages 1–40. Idea Group Publishing, London, 2007.

[92] Shawe-Taylor, J. and Cristianini, N. . *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, 2004.

[93] Leslie, C. S. , Eskin, E. , Cohen, A. , Weston, J. , and Noble, W. S. . Mismatch string kernels for discriminative protein classification. *Bioinformatics*, 20(4):467–476, 2004.

[94] Needleman, S. B. and Wunsch, C. D. . A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[95] Durbin, R. , Eddy, S. , Krogh, A. , and Mitchison, G. . *Biological Sequence Analysis Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, United Kingdom, 1998.

[96] Smith, T. F. and Waterman, M. S. . Identification of Common Molecular Subsequences. *Journal of Molecular Biology*, 147(1):195–197, 1981.

[97] Cao, D.-S. , Xu, Q.-S. , and Liang, Y.-Z. . propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, 29(7):960–962, 2013.

[98] Berg, B. A. , van den, Reinders, M. J. , Roubos, J. A. , and Ridder, D. , de. SPiCE: a web-based tool for sequence-based protein classification and exploration. *BMC Bioinformatics*, 15:93–102, 2014.

[99] Ruan, J. , Wang, K. , Yang, J. , Kurgan, L. A. , and Cios, K. . Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. *Artificial intelligence in medicine*, 35(1-2):19–35, 2005.

[100] Matsuzaki, K. . Why and how are peptide-lipid interactions utilized for self-defense? Magainins and tachyplesins as archetypes. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1462:1–10, 1999.

[101] Houghten, R. A. . General method for the rapid solid-phase synthesis of large numbers of peptides: specificity of antigen-antibody interaction at the level of individual amino acids. *Proceedings of the National Academy of Sciences of the United States of America*, 82(15):5131–5135, 1985.

[102] Fields, G. B. and Noble, R. L. . Solid phase peptide synthesis utilizing 9-fluorenylmethoxycarbonyl amino acids. *International Journal of Peptide and Protein Research*, 35(3):161–214, 1990.

[103] Jofré, C. , Guzmán, F. , Cárdenas, C. , Albericio, F. , and Marshall, S. H. . A natural peptide and its variants derived from the processing of infectious pancreatic necrosis virus (IPNV) displaying enhanced antimicrobial activity: A novel alternative for the control of bacterial diseases. *Peptides*, 32:852–858, 2011.

[104] Haney, E. F. , Lau, F. , and Vogel, H. J. . Solution structures and model membrane interactions of lactoferrampin, an antimicrobial peptide derived from bovine lactoferrin. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1768(10):2355–2364, 2007.

[105] Fritsche, T. R. , Rhomberg, P. R. , Sader, H. S. , and Jones, R. N. . Antimicrobial activity of omiganan pentahydrochloride tested against contemporary bacterial pathogens commonly responsible for catheter-associated infections. *Journal of Antimicrobial Chemotherapy*, 61:1092–1098, 2008.

[106] Li, W. and Godzik, A. . Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[107] Huang, Y. , Niu, B. , Gao, Y. , Fu, L. , and Li, W. . CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010.

[108] Ali, S. and Smith, K. A. . On learning algorithm selection for classification. *Applied Soft Computing*, 6(2):119–138, 2006.

[109] Nagarajan, V. , Kaushik, N. , Murali, B. , Zhang, C. , Lakhera, S. , Elasri, M. O. , and Deng, Y. . A Fourier Transformation based Method to Mine Peptide Space for Antimicrobial Activity. *BMC Bioinformatics*, 7(Suppl 2):S2–S9, 2006.

[110] Polanco, C. and Samaniego, J. L. . Detection of selective cationic amphipatic antibacterial peptides by Hidden Markov models. *Acta Biochimica Polonica*, 56(1):167–176, 2009.

[111] Webb, A. R. and Copsey, K. D. . *Statistical Pattern Recognition.* Wiley, Chichester, UK, third edition, 2011.

[112] Burges, C. J. . A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[113] Hastie, T. , Tibshirani, R. , and Friedman, J. . *The Elements of Statistical Learning.* Springer, New York, NY, USA, springer series in statistics edition, 2001.

[114] Jiang, Z. , Vasil, A. I. , Gera, L. , Vasil, M. L. , and Hodges, R. S. . Rational Design of alpha-Helical Antimicrobial Peptides to Target Gram-negative Pathogens, Acinetobacter baumannii and Pseudomonas aeruginosa: Utilization of Charge, 'Specificity Determinants', Total Hydrophobicity, Hydrophobe Type and Location as Design Parameters to Improve the Therapeutic Ratio . *Chemical Biology & Drug Design*, 77(4):225–240, 2011.

[115] Wessels, L. F. A. , Reinders, M. J. T. , Hart, A. A. M. , Veenman, C. J. , Dai, H. , He, Y. D. , and Veer, L. J. , van't. A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics*, 21(19):3755–3762, 2005.

[116] Szymaska, E. , Saccenti, E. , Smilde, A. K. , and Westerhuis, J. A. . Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*, 8(Suppl 1):S3–S16, 2012.

[117] Maaten, L. V. , der and Hinton, G. . Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[118] Bodapati, K. C. , Soudy, R. , Etayash, H. , Stiles, M. , and Kaur, K. . Design, synthesis and evaluation of antimicrobial activity of N-terminal modified Leucocin A analogues. *Bioorganic & medicinal chemistry*, 21(13):3715–3722, 2013.

[119] Naamati, G. , Askenazi, M. , and Linial, M. . ClanTox: a classifier of short animal toxins. *Nucleic acids research*, 37(Web Server issue):W363 – W368, 2009.

[120] Saha, S. and Raghava, G. P. S. . BTXpred: Prediction of bacterial toxins. *In silico Biology*, 7(4-5):405–412, 2007.

[121] Saha, S. and Raghava, G. P. S. . Prediction of neurotoxins based on their function and source. *In silico Biology*, 7(4-5):369–387, 2007.

[122] Chakraborty, A. , Ghosh, S. , Chowdhary, G. , Maulik, U. , and Chakrabarti, S. . DBETH: A Database of Bacterial Exotoxins for Human. *Nucleic Acids Research*, 40(Database issue):D615–D620, 2012.

[123] Bairoch, A. and Apweiler, R. . The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1):45–48, 2000.

# Bibliography

Ali, S. and Smith, K. A. (2006). On learning algorithm selection for classification. *Applied Soft Computing*, 6(2):119–138.

Amábile-Cuevas, C. F. (2010). Global Perspectives of Antibiotic Resistance. In Sosa, A. D. J., Byarugaba, D. K., Amábile-Cuevas, C. F., Hsueh, P.-R., Kariuki, S., and Okeke, I. N., editors, *Antimicrobial Resistance in Developing Countries*, chapter 1, pages 3– 13. Springer New York, New York.

Avram, S., Duda-Seiman, D., Borcan, F., Radu, B., Duda-Seiman, C., and Mihailescu, D. (2011). Evaluation of Antimicrobial Activity of New Mastoparan Derivatives Using QSAR and Computational Mutagenesis. *International Journal of Peptide Research and Therapeutics*, 17(1):7–17.

Avram, S., Mihailescu, D., Borcan, F., and Milac, A.-L. (2012). Prediction of improved antimicrobial mastoparan derivatives by 3D-QSAR-CoMSIA/CoMFA and computational mutagenesis. *Monatshefte für Chemie - Chemical Monthly*, 143:535–543.

Bahar, A. A. and Ren, D. (2013). Antimicrobial peptides. *Pharmaceuticals*, 6(12):1543–1575.

Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, 28(1):45–48.

Baltzer, S. A. and Brown, M. H. (2011). Antimicrobial Peptides - Promising Alternatives to Conventional Antibiotics. *Journal of Molecular Microbiology and Biotechnology*, 20(4):228–235.

Bettiol, E., Wetherington, J. D., Schmitt, N., and Harbarth, S. (2015). Challenges and Solutions for Clinical Development of New Antibacterial Agents: Results of a Survey among Pharmaceutical Industry Professionals. *Antimicrobial agents and chemotherapy*, 59(7):3695–3699.

Bodapati, K. C., Soudy, R., Etayash, H., Stiles, M., and Kaur, K. (2013). Design, synthesis and evaluation of antimicrobial activity of N-terminal modified Leucocin A analogues. *Bioorganic & medicinal chemistry*, 21(13):3715–3722.

Boman, H. G. (2003). Antibacterial peptides: basic facts and emerging concepts. *Journal of Internal Medicine*, 254(3):197–215.

Boucher, H. W., Talbot, G. H., Benjamin Jr, D. K., Bradley, J., Guidos, R. J., Jones, R. N., Murray, B. E., Bonomo, R. A., and Gilbert, D. (2013). 10 x '20 Progress-Development of New Drugs Active Against Gram-Negative Bacilli: An Update From the Infectious Diseases Society of America. *Clinical Infectious Diseases*, 56(12):1685–1694.

Brown, E. D. and Wright, G. D. (2016). Antibacterial drug discovery in the resistance era. *Nature*, 529:336–343.

Burges, C. J. (1998). A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167.

Cao, D.-S., Xu, Q.-S., and Liang, Y.-Z. (2013). Propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics*, 29(7):960–962.

Chakraborty, A., Ghosh, S., Chowdhary, G., Maulik, U., and Chakrabarti, S. (2012). DBETH: A Database of Bacterial Exotoxins for Human.*Nucleic Acids Research*, 40(Database issue):D615–D620.

Chen, W. and Luo, L. (2009). Classification of antimicrobial peptide using diversity measure with quadratic discriminant analysis. *Journal of Microbiological Methods*, 78(1):94–96.

Cherkasov, A. (2005). 'Inductive' Descriptors : 10 Successful Years in QSAR. *Current Computer-Aided Drug Design*, 1(1):21–42.

Cherkasov, A., Hilpert, K., Jenssen, H. v., Fjell, C. D., Waldbrook, M., Mullaly, S. C., Volkmer, R., and Hancock, R. E. W. (2009). Use of Artificial Intelligence in the Design of Small Peptide Antibiotics Effective against a Broad Spectrum of Highly Antibiotic-Resistant Superbugs. *ACS Chemical Biology*, 4(1):65–74.

Consortium, T. U. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, 43(Database issue):D204–D212.

Cristianini, N., Shawe-Taylor, J., and Saunders, C. (2007). Kernel Methods: A Paradigm for Pattern Analysis. In Camps-Valls, G., Rojo-Álvarez, J. L., and Martínez-Ramón, M., editors, *Kernel Methods in Bioengineering, Signal and Image Processing*, pages 1–40. Idea Group Publishing, London.

de Jong, A., van Heel, A. J., Kok, J., and Kuipers, O. P. (2010). BAGEL2: mining for bacteriocins in genomic data. *Nucleic Acids Research*, 38(Web Server issue):W647–W651.

der Maaten, L. V. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

Di Luca, M., Maccari, G., Maisetta, G., and Batoni, G. (2015). BaAMPs: the database of biofilm-active antimicrobial peptides. *Biofouling*, 31(2):193–199.

Dudek, A. Z., Arodz, T., and Gálvez, J. (2006). Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Combinatorial Chemistry & High Throughput Screening*, 9(3):213–228.

Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998).*Biological Sequence Analysis Probabilistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge, United Kingdom.

ECDC, E. C. f. D. P. and Control (2015). Antimicrobial resistance surveillance in Europe 2014. Annual Report of the European Antimicrobial Resistance Surveillance Network (EARS-Net).

Technical report, Stockholm.

Estabrooks, A., Jo, T., and Japkowicz, N. (2004). A Multiple Resampling Method for Learning from Imbalanced Data Sets. *Computational Intelligence*, 20(1):18–36.

Fields, G. B. and Noble, R. L. (1990). Solid phase peptide synthesis utilizing 9-fluorenylmethoxycarbonyl amino acids. *International Journal of Peptide and Protein Research*, 35(3):161–214.

Fjell, C. D., Hancock, R. E. W., and Cherkasov, A. (2007). AMPer: a database and an automated discovery tool for antimicrobial peptides.*Bioinformatics*, 23(9):1148–1155.

Fjell, C. D., Jenssen, H. v., Cheung, W. A., Hancock, R. E. W., and Cherkasov, A. (2011). Optimization of antibacterial peptides by genetic algorithms and cheminformatics. *Chemical biology & drug design*, 77(1):48–56.

Fjell, C. D., Jenssen, H. v., Hilpert, K., Cheung, W. A., Panté, N., Hancock, R. E. W., and Cherkasov, A. (2009). Identification of novel antibacterial peptides by chemoinformatics and machine learning.*Journal of Medicinal Chemistry*, 52(7):2006–2015.

Fritsche, T. R., Rhomberg, P. R., Sader, H. S., and Jones, R. N. (2008). Antimicrobial activity of omiganan pentahydrochloride tested against contemporary bacterial pathogens commonly responsible for catheter-associated infections. *Journal of Antimicrobial Chemotherapy*, 61:1092–1098.

Galar, M., Fernández, A., Barrenechea, E., Bustince, H., and Herrera, F. (2012). A review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. *IEEE transactions on systems, man, and cybernetics. Part C, Applications and Reviews*, 42(4):463–484.

Giuliani, A., Pirri, G., and Nicoletto, S. F. (2007). Antimicrobial peptides: an overview of a promising class of therapeutics.*Central European Journal of Biology*, 2(1):1–33.

Gogoladze, G., Grigolava, M., Vishnepolsky, B., Chubinidze, M., Duroux, P., Lefranc, M.-P., and Pirtskhalava, M. (2014). DBAASP: database of antimicrobial activity and structure of peptides. *FEMS microbiology letters*, 357(1):63–68.

Gueguen, Y., Garnier, J., Robert, L., Lefranc, M.-P., Mougenot, I., De Lorgeril, J., Janech, M., Gross, P. S., Warr, G. W., Cuthbertson, B., Barracco, M. A., Bulet, P., Aumelas, A., Yang, Y., Bo, D., Xiang, J., Tassanakajon, A., Piquemal, D., and Bachère, E. (2006). PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature.*Developmental & Comparative Immunology*, 30(3):283–288.

Gupta, S., Kapoor, P., Chaudhary, K., Gautam, A., Kumar, R., Consortium, O. S. D. D., and Raghava, G. P. S. (2013). In silico approach for predicting toxicity of peptides and proteins. *PloS one*, 8(9):e73957–e73966.

Gupta, S., Kapoor, P., Chaudhary, K., Gautam, A., Kumar, R., and Raghava, G. P. S. (2015). Peptide Toxicity Prediction. In *Computational Peptidology*, volume 1268, pages 143–157.

Hammami, R., Hamida, J. B., Vergoten, G., and Fliss, I. (2009). PhytAMP: a database dedicated to antimicrobial plant peptides. *Nucleic Acids Research*, 37(Database issue):D963–D968.

Hammami, R., Zouhir, A., Hamida, J. B., and Fliss, I. (2007). BACTIBASE: a new web-accessible database for bacteriocin characterization. *BMC Microbiology*, 7:89–94.

Hammami, R., Zouhir, A., Le Lay, C., Hamida, J. B., and Fliss, I. (2010). BACTIBASE second release: a database and tool platform for bacteriocin characterization. *BMC Microbiology*, 10(22):1–5.

Hancock, R. E. W. and Sahl, H.-G. (2006). Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nature Biotechnology*, 24(12):1551–1557.

Haney, E. F., Lau, F., and Vogel, H. J. (2007). Solution structures and model membrane interactions of lactoferrampin, an antimicrobial peptide derived from bovine lactoferrin. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1768(10):2355–2364.

Hastie, T., Tibshirani, R., and Friedman, J. (2001).*The Elements of Statistical Learning.* Springer, New York, NY, USA, springer series in statistics edition.

He, H. and Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284.

HHS, U. D. o. H., Services, H., CDC, C. f. D. C., and Prevention (2013). Antibiotic resistance threats in the United States, 2013. Technical report, USA.

Hilpert, K., Elliott, M. R., Volkmer-Engert, R., Henklein, P., Donini, O., Zhou, Q., Winkler, D. F. H., and Hancock, R. E. W. (2006). Sequence requirements and an optimization strategy for short antimicrobial peptides. *Chemistry & Biology*, 13(10):1101–1107.

Houghten, R. A. (1985). General method for the rapid solid-phase synthesis of large numbers of peptides: specificity of antigen-antibody interaction at the level of individual amino acids. *Proceedings of the National Academy of Sciences of the United States of America*, 82(15):5131–5135.

Huang, Y., Niu, B., Gao, Y., Fu, L., and Li, W. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682.

Infectious Diseases Society of America (2010). The 10 x '20 Initiative: pursuing a global commitment to develop 10 new antibacterial drugs by 2020. *Clinical infectious diseases*, 50(8):1081–1083.

Jenssen, H., Fjell, C. D., Cherkasov, A., and Hancock, R. E. W. (2008). QSAR modeling and computer-aided design of antimicrobial peptides. *Journal of Peptide Science*, 14(1):110–114.

Jenssen, H. v., Lejon, T., Hilpert, K., Fjell, C. D., Cherkasov, A., and Hancock, R. E. W. (2007). Evaluating different descriptors for model design of antimicrobial peptides with enhanced activity toward P. aeruginosa. *Chemical Biology Drug Design*, 70(2):134–142.

Jiang, Z., Vasil, A. I., Gera, L., Vasil, M. L., and Hodges, R. S. (2011). Rational Design of alpha-Helical Antimicrobial Peptides to Target Gram-negative Pathogens, Acinetobacter baumannii and Pseudomonas aeruginosa: Utilization of Charge, 'Specificity Determinants', Total Hydrophobicity, Hydrophobe Type and Location as Design Parameters to Improve the Therapeutic Ratio. *Chemical Biology & Drug Design*, 77(4):225–240.

Jofré, C., Guzmán, F., Cárdenas, C., Albericio, F., and Marshall, S. H. (2011). A natural peptide and its variants derived from the processing of infectious pancreatic necrosis virus (IPNV) displaying enhanced antimicrobial activity: A novel alternative for the control of bacterial diseases. *Peptides*, 32:852–858.

Joseph, S., Karnik, S., Nilawe, P., Jayaraman, V. K., and Idicula-Thomas, S. (2012). ClassAMP: a prediction tool for classification of antimicrobial peptides. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 9(5):1535–1538.

Juretić, D., Vukicević, D., Ilić, N., Antcheva, N., and Tossi, A. (2009). Computational design of highly selective antimicrobial peptides. *Journal of Chemical Information and Modeling*, 49(12):2873–2882.

Khosravian, M., Faramarzi, F. K., Beigi, M. M., Behbahani, M., and Mohabatkar, H. (2013). Predicting Antibacterial Peptides by the Concept of Chou's Pseudo-amino Acid Composition and Machine Learning Methods.*Protein & Peptide Letters*, 20(2):180–186.

Kim, H., Jang, J. H., Kim, S. C., and Cho, J. H. (2014). De novo generation of short antimicrobial peptides with enhanced stability and cell specificity.*The Journal of antimicrobial chemotherapy*, 69(1):121–132.

Langham, A. A., Khandelia, H., Schuster, B., Waring, A. J., Lehrer, R. I., and Kaznessis, Y. N. (2008). Correlation between simulated physicochemical properties and hemolycity of protegrin-like antimicrobial peptides: Predicting experimental toxicity. *Peptides*, 29(7):1085–1093.

Lata, S., Mishra, N. K., and Raghava, G. P. (2010). AntiBP2: improved version of antibacterial peptide prediction. *BMC Bioinformatics*, 11(Suppl 1):S19–S25.

Lata, S., Sharma, B., and Raghava, G. (2007). Analysis and prediction of antibacterial peptides.*BMC Bioinformatics*, 8:263–272.

Leslie, C. S., Eskin, E., Cohen, A., Weston, J., and Noble, W. S. (2004). Mismatch string kernels for discriminative protein classification.*Bioinformatics*, 20(4):467–476.

Li, J., Qu, X., He, X., Duan, L., Wu, G., Bi, D., Deng, Z., Liu, W., and Ou, H.-Y. (2012). ThioFinder : A Web-Based Tool for the Identification of Thiopeptide Gene Clusters in DNA Sequences. *PloS one*, 7(9):e45878–e45886.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.*Bioinformatics*, 22(13):1658–1659.

Li, Y. and Chen, Z. (2008). RAPD: a database of recombinantly-produced antimicrobial peptides. *FEMS Microbiology Letters*, 289(2):126–129.

Maccari, G., Di Luca, M., Nifosí, R., Cardarelli, F., Signore, G., Boccardi, C., and Bifone, A. (2013). Antimicrobial Peptides Design by Evolutionary Multiobjective Optimization.*PLoS Computational Biology*, 9(9):e1003212–e1003223.

Maccari, G., Luca, M. D., and Nifosì, R. (2015). In Silico Design of Antimicrobial Peptides. In Zhou, P. and Huang, J., editors, *Computational Peptidology*, volume 1268 of *Methods in*

*Molecular Biology*, pages 195–219. Springer New York, New York, NY.

Marr, A. K., Gooderham, W. J., and Hancock, R. E. (2006). Antibacterial peptides for therapeutic use: obstacles and realistic outlook.*Current Opinion in Pharmacology*, 6:468–472.

Matsuzaki, K. (1999). Why and how are peptide-lipid interactions utilized for self-defense? Magainins and tachyplesins as archetypes.*Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1462:1–10.

Mikut, R. and Hilpert, K. (2009). Interpretable Features for the Activity Prediction of Short Antimicrobial Peptides Using Fuzzy Logic. *International Journal of Peptide Research and Therapeutics*, 15(2):129–137.

Naamati, G., Askenazi, M., and Linial, M. (2009). ClanTox: a classifier of short animal toxins. *Nucleic acids research*, 37(Web Server issue):W363 – W368.

Nagarajan, V., Kaushik, N., Murali, B., Zhang, C., Lakhera, S., Elasri, M. O., and Deng, Y. (2006). A Fourier Transformation based Method to Mine Peptide Space for Antimicrobial Activity. *BMC Bioinformatics*, 7(Suppl 2):S2–S9.

Needleman, S. B. and Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, 48(3):443–453.

Nelson, D. L. and Cox, M. M. (2004).*Lehninger Principles of Biochemistry Fourth Edition.* W. H. Freeman.

Novković, M., Simunić, J., Bojović, V., Tossi, A., and Juretić, D. (2012). DADP: the Database of Anuran Defense Peptides.*Bioinformatics*, 28(10):1406–1407.

Okorochenkov, S. A., Zheltukhina, G. A., and Nebolsin, V. E. (2011). Antimicrobial peptides: the mode of action and perspectives of practical application.*Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry*, 5:95–102.

Pasupuleti, M., Schmidtchen, A., and Malmsten, M. (2011). Antimicrobial peptides: key components of the innate immune system. *Critical Reviews in Biotechnology*, 32(2):143–171.

Pharmaceutical Research and Manufacturers of America (2015). *2015 Biopharmaceutical Research Industry Profile.* PhRMA, Washington, DC.

Piotto, S. P., Sessa, L., Concilio, S., and Iannelli, P. (2012). YADAMP: yet another database of antimicrobial peptides. *International Journal of Antimicrobial Agents*, 39(4):346–351.

Polanco, C. and Samaniego, J. L. (2009). Detection of selective cationic amphipatic antibacterial peptides by Hidden Markov models *Acta Biochimica Polonica*, 56(1):167–176.

Porto, W. F., Fernandes, F. C., and Franco, O. L. (2010). An SVM Model Based on Physicochemical Properties to Predict Antimicrobial Activity from Protein Sequences with Cysteine Knot Motifs. *Lectures Notes in Computer Science*, 6268:59–62.

Poupard, J. A. (2005). Is The Pharmaceutical Industry Responding to the Challenge of Increasing Bacterial Resistance? *Clinical Microbiology Newsletter*, 28(2):13–15.

Projan, S. J. (2003). Why is big Pharma getting out of antibacterial drug discovery? *Current Opinion in Microbiology*, 6(5):427–430.

Projan, S. J. and Bradford, P. A. (2007). Late stage antibacterial drugs in the clinical pipeline. *Current Opinion in Microbiology*, 10(5):441–446.

Roca, I., Akova, M., Baquero, F., Carlet, J., Cavaleri, M., Coenen, S., Cohen, J., Findlay, D., Gyssens, I., Heure, O., Kahlmeter, G., Kruse, H., Laxminarayan, R., Liébana, E., López-Cerero, L., MacGowan, A., Martins, M., Rodríguez-Baño, J., Rolain, J.-M., Segovia, C., Sigauque, B., Taconelli, E., Wellington, E., and Vila, J. (2015). The global threat of antimicrobial resistance: Science for intervention. *New Microbes and New Infections*, 6:22–29.

Rondón-Villarreal, P., Reinders, M. J. T., Camacho, F., Torres, R., Sierra, D. A., and Tax, D. M. Implications of the Optimization Criterion when Creating an Antibacterial Peptide Classifier. *Paper in revision.*

Rondón-Villarreal, P., Sierra, D. A., and Torres, R. (2014a). Classification of antimicrobial peptides by using the p -spectrum kernel and support vector machines. *Advances in Computational Biology. Advances in Intelligent Systems and Computing*, 232:155–161.

Rondón-Villarreal, P., Sierra, D. A., and Torres, R. (2014b). Machine Learning in the Rational Design of Antimicrobial Peptides. *Current Computer-Aided Drug Design*, 10:183–190.

Ruan, J., Wang, K., Yang, J., Kurgan, L. A., and Cios, K. (2005). Highly accurate and consistent method for prediction of helix and strand content from primary protein sequences. *Artificial intelligence in medicine*, 35(1-2):19–35.

Saha, S. and Raghava, G. P. S. (2007a). BTXpred: Prediction of bacterial toxins.*In silico Biology*, 7(4-5):405–412.

Saha, S. and Raghava, G. P. S. (2007b). Prediction of neurotoxins based on their function and source. *In silico Biology*, 7(4-5):369–387.

Seebah, S., Suresh, A., Zhuo, S., Choong, Y. H., Chua, H., Chuon, D., Beuerman, R., and Verma, C. (2007). Defensins knowledgebase: a manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Research*, 35(Database issue):D265–D268.

Shawe-Taylor, J. and Cristianini, N. (2004).*Kernel Methods for Pattern Analysis.* Cambridge University Press, New York.

Shen, Y., Maupetit, J., Derreumaux, P., and Tufféry, P. (2014). Improved PEP-FOLD approach for peptide and miniprotein structure prediction.*Journal of Chemical Theory and Computation*, 10(10):4745–4758.

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

Sundararajan, V. S., Gabere, M. N., Pretorius, A., Adam, S., Christoffels, A., Lehväslaiho, M., Archer, J. A. C., and Bajic, V. B. (2012). DAMPD: a manually curated antimicrobial peptide database.*Nucleic Acids Research*, 40(Database issue):D1108–D1112.

Szymaska, E., Saccenti, E., Smilde, A. K., and Westerhuis, J. A. (2012). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. *Metabolomics*, 8(Suppl 1):S3–S16.

Taboureau, O., Olsen, O. H., Nielsen, J. D., Raventos, D., Mygind, P. H., and Kristensen, H.-H. (2006). Design of novispirin antimicrobial peptides by quantitative structure-activity relationship. *Chemical Biology & Drug Design*, 68(1):48–57.

Takahashi, D., Shukla, S. K., Prakash, O., and Zhang, G. (2010). Structural determinants of host defense peptides for antimicrobial activity and target cell selectivity.*Biochimie*, 92(9):1236–1241.

Tan, J., Huang, J., Huang, Y., and Chen, Y. (2014). Effects of single amino acid substitution on the biophysical properties and biological activities of an amphipathic $\alpha$-helical antibacterial peptide against Gram-negative bacteria. *Molecules*, 19(8):10803–10817.

Thomas, S., Karnik, S., Barai, R. S., Jayaraman, V. K., and Idicula-Thomas, S. (2010). CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Research*, 38(Database issue):D774–D780.

Torrent, M., Andreu, D., Nogués, V. M., and Boix, E. (2011). Connecting Peptide Physicochemical and Antimicrobial Properties by a Rational Prediction Model.*PLoS one*, 6(2):e16968–e16975.

Tossi, A. and Sandri, L. (2002). Molecular Diversity in Gene-Encoded, Cationic Antimicrobial Polypeptides.*Current Pharmaceutical Design*, 8(9):743–761.

van den Berg, B. A., Reinders, M. J., Roubos, J. A., and de Ridder, D. (2014). SPiCE: a web-based tool for sequence-based protein classification and exploration.*BMC Bioinformatics*, 15:93–102.

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

Wade, D. and Englund, J. (2002). Synthetic antibiotic peptides database. *Protein and peptide letters*, 9(1):53–57.

Wang, C. K. L., Kaas, Q., Chiche, L., and Craik, D. J. (2007). CyBase: a database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Research*, 36(Database issue):D206–D2010.

Wang, G. (2013).Database-Guided Discovery of Potent Peptides to Combat HIV-1 or Superbugs.*Pharmaceuticals*, 6(6):728–758.

Wang, G. (2015). Improved Methods for Classification, Prediction, and Design of Antimicrobial Peptides. In *Computational Peptidology*, volume 1268, pages 43–66.

Wang, G., Li, X., and Wang, Z. (2009). APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Research*, 37(Database issue):D933–D937.

Wang, P., Hu, L., Liu, G., Jiang, N., Chen, X., Xu, J., Zheng, W., Li, L., Tan, M., Chen, Z., Song, H., Cai, Y.-D., and Chou, K.-C. (2011). Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PloS one*, 6(4):e18476–e18484.

Wang, Z. and Wang, G. (2004). APD: the Antimicrobial Peptide Database.*Nucleic Acids Research*, 32(Database issue):D590–D592.

Webb, A. R. and Copsey, K. D. (2011). *Statistical Pattern Recognition.* Wiley, Chichester, UK, third edition.

Wessels, L. F. A., Reinders, M. J. T., Hart, A. A. M., Veenman, C. J., Dai, H., He, Y. D., and van't Veer, L. J. (2005). A protocol for building and evaluating predictors of disease state based on microarray data.*Bioinformatics*, 21(19):3755–3762.

Whitmore, L. and Wallace, B. A. (2004). The Peptaibol Database: a database for sequences and structures of naturally occurring peptaibols.*Nucleic Acids Research*, 32(Database issue):D593–D594.

WHO, W. H. O. (2014). Antimicrobial Resistance Global Report on Surveillance. Technical report, France.

Williams, K. J. and Bax, R. P. (2009). Challenges in developing new antibacterial drugs. *Current Opinion in Investigational Drugs*, 10(2):157–163.

Wimley, W. C. and Hristova, K. (2011). Antimicrobial peptides: successes, challenges and unanswered questions. *The Journal of membrane biology*, 239(1-2):27–34.

Wu, H., Lu, H., Huang, J., Li, G., and Huang, Q. (2012). EnzyBase: a novel database for enzybiotic studies. *BMC Microbiology*, 12:54–58.

Wu, X., Wang, Z., Li, X., Fan, Y., He, G., Wan, Y., Yu, C., Tang, J., Li, M., Zhang, X., Zhang, H., Xiang, R., Pan, Y., Liu, Y., Lu, L., and Yang, L. (2014). In vitro and in vivo activities of antimicrobial peptides developed using an amino acid-based activity prediction method. *Antimicrobial agents and chemotherapy*, 58(9):5342–5349.

Xiao, X., Wang, P., Lin, W.-Z., Jia, J.-H., and Chou, K.-C. (2013). iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical biochemistry*, 436(2):168–177.

Zhao, X., Wu, H., Lu, H., Li, G., and Huang, Q. (2013). LAMP: A Database Linking Antimicrobial Peptides. *PloS one*, 8(6):e66557–e66562.

Zhu, X., Ma, Z., Wang, J., Chou, S., and Shan, A. (2014). Importance of Tryptophan in Transforming an Amphipathic Peptide into a Pseudomonas aeruginosa-Targeted Antimicrobial Peptide. *PloS one*, 9(12):e114605–e114623.