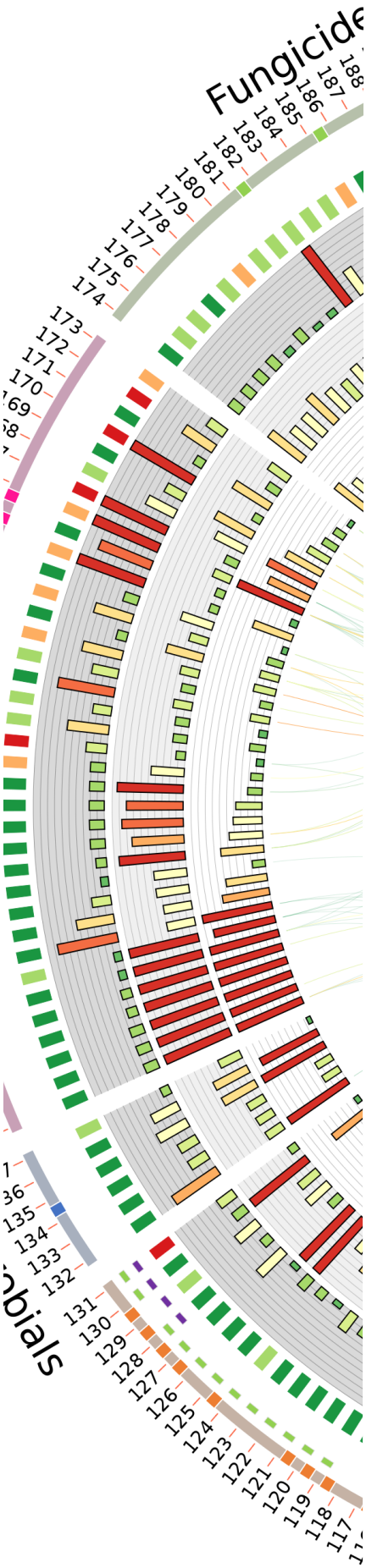
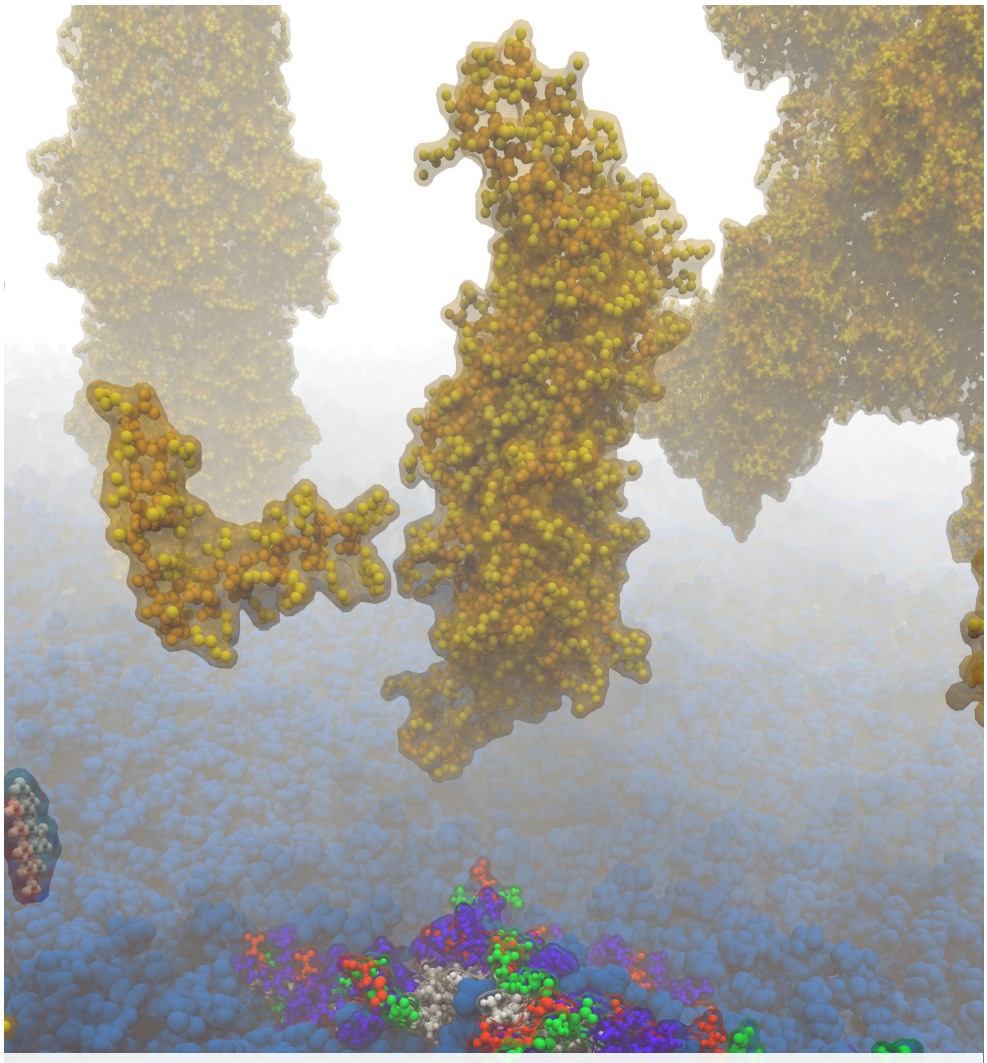
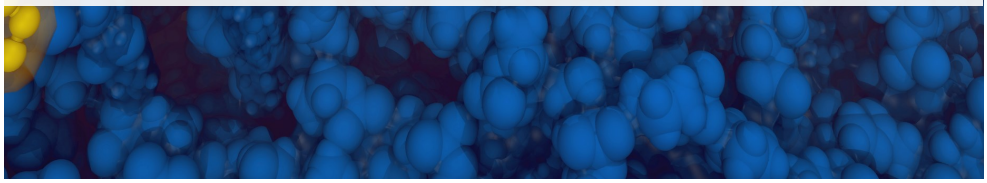




Screening for Common Antimicrobial Sequence Motifs and Testing Their Activity Against Model of the Inner Membrane of Escherichia coli

Svetomir Hitov
Supervisor: Peter Fojan

2017



Screening for Common Antimicrobial Sequence Motifs and Testing Their Activity Against Model of the Inner Membrane of *Escherichia coli*

Svetomir Hitov

June 10, 2017

AALBORG UNIVERSITY

Department of Physics
and Nanotechnology
Skjernvej 4a, 9220 Allborg Ø
Telephone: (0045) 96 35 97 31
Fax: (0045) 98 13 63 93

Title: Screening for Common Antimicrobial Sequence Motifs and Testing Their Activity Against Model of the Inner Membrane of *Escherichia coli*

Author: Svetomir Hitov

Supervisor: Peter Fojan

Number of pages: 67

Date: June 10, 2017

Abstract

Antimicrobial Peptides (AMPs) are a group of peptides that are produced primarily by eukaryotic organisms to fight invading bacteria, viruses, fungi, etc. Except for the antimicrobial activity, these peptides show immunoregulatory, antibiofilm, and anticancer activities. AMPs are interesting research objects because they are molecules, which in the future can be a replacement for the failing conventional antibiotics. Furthermore, there are known cases in which antimicrobial peptides are part of the longer and bigger proteins. After cleavage of these large proteins, the actual AMPs are released in the environment. We can exploit this property to find cheap and easy ways to mass produce AMPs. In this work the space of AMPs is explored for the presence of such peptide sequences and their activity against the membrane of the Gram-negative bacteria *Escherichia coli* is researched.

Contents

Chapter 1

Introduction and Overview

Goals of the project

The current project has the aim of finding sequence motifs that have high occurrences in the currently known AMP sequences present in the annotated databases. These sequences will then be used to search for more putative antimicrobial sequences amongst sequences that are not annotated. The focus of the project will be AMP sequences that are reported to act on the membrane of the microbial species. The ones that are found most interesting will be analysed for their action on a model of a bacterial membrane.

1.1 Antimicrobial Peptides

AMPs are a class of peptides synthesised by a wide range of organisms. In the beginning, it was believed that AMP are limited to the vertebrates. With further studies, it was discovered that similar peptides are produced by nearly every group of the living beings. These peptides have different forms and modes of action but their common property is the destruction of invading or competitive microbial species

Advantages and disadvantages compared to conventional antibiotics Distinctive advantages of the AMPs are the ability to kill microbial species within a short time frame. Most of the AMPs are believed to employ more than one mechanism of action against the same target, which makes developing of resistance more difficult. A distinct disadvantage of the AMPs, from an industrial perspective, is the complicated and costly production of such peptide entities.

1.1.1 General Overview of Antimicrobial Peptides

As already stated, AMPs, also known as host defence peptides, are a group of peptides synthesised by invertebrates and vertebrates. Recently, plants, fungi and even bacteria have also been found to synthesise AMPs. These peptides, with a few exceptions, have modest direct antimicrobial activity against different kinds of Gram-positive and Gram-negative bacteria, viruses, fungi, and protozoa. Some AMPs have also shown activity against insects and some cancer cells [1]. Moreover, they have immunomodulatory properties that are connected to the innate immune system: anti-infective and anti-inflammatory activity, increasing chemokine production, adjuvant

and enhancing wound healing and angiogenesis, exert pro- and anti-apoptotic effect on different immune cell types [1, 2, 3].

1.1.2 Structure and Classes of AMPs

There are many different kinds of AMPs, including defensins, magainins, cecropins, cathelicidins, etc [3]. This diversity is based upon their modes of action, structures and bioactivities. Most of the AMPs have no distinctive structure in aqueous solution, but adopt an amphipathic conformation in the presence of biological membranes [2]. The different structures add to diversity in the modes of action. In general, the initial affinity between the AMP and structures of the microorganism is attributed to electrostatic attraction between anionic molecules on the target surface and cationic residues of the AMP. In spite of the lack of precise definitions of the mechanisms by which AMPs work, there are two main arms of theories. The first arm consists of models connected with membrane disruption and the second, models associated with intracellular targets.

1.1.3 Targets of AMPs and Modes of Action

Bacterial Membrane

Membrane-disruptive models include barrel-stave, toroidal-pore, aggregate, detergent, carpet and sinking raft models. In the heart of these models is the idea that the AMPs form kind of a pore or disrupt the cell membrane of the target cell. This leads to three possible outcomes: formation of a transient channel, micellization or dissolution of the membrane, or translocation of the peptide across the membrane [4]. This can lead to leaking of cell constituents out of the cell or disruption of the intrinsic properties of the cell membrane and the cell wall. Finally, these changes result in the dead of the cell.

Barrel-stave Model The barrel-stave model was first proposed by Baumann and Mueller [5], Boheim [6], Ehrenstein and Lecar [7] for the mode of action of Alamethicin and its analogues members of the peptaibol family of AMPs. The model was additionally refined by Laver [8]. This model predicts that short, rod-like peptides with α-helical structure self-assemble on the lipid surface. In the presence of an electrical field, the rod-like peptides insert themselves in the membrane and line a cylindrical, electrolyte-filled pore. The name of the model comes from the similarity of the pore that is formed like a barrel with the peptides aligned at the edge of the pore, like staves. The alamethicin pores can contain between 3 and 11 peptides and have on average inner diameter of 1.8 nm and outer diameter of 4.0 nm [9, 10, 11]. The model predicts that the peptides should be amphiphilic with the hydrophobic parts facing the interior of the membrane and hydrophilic parts lining the interior of the channel. There is experimental evidence that alamethicin, the cyclic decameric cationic peptide gramicidin S [4], pardaxin and the α-5 segment of *Bacillus thuringiensis* δ-endotoxin [12] follow this mechanism.

Toroidal-pore Model The toroidal-pore model, to a lesser extent known as wormhole model, predicts that antimicrobial peptide helices insert themselves in the membrane, forcing the lipids to bend and open a whole in the membrane. In

this model, the monolayers bend continuously and there is no contact between the peptides and the interior of the membrane. The pore is lined by both lipid head groups and then a layer of peptides, which is perpendicular to the membrane [4]. This model differs from the barrel-stave model as the peptides are always associated with the lipid head groups. This shields them from eventual electrostatic repulsion between the positive charges of the peptides when they are inserted in the membrane as in the barrel-stave model. This kind of pores is believed to be formed by protegrins, melittin, magainins, MSI-78 (a synthetic analogue of magainin-2) [13, 14, 15, 11], LL-37 [16]. Magainin pores, in comparison to alamethicin (barrel-stave model), are larger and have more variable pore size. The inner diameter is from 3.0 nm to 5.0 nm and the outer diameter ranges from 7.0 nm to 8.4 nm. Each pore is formed by 4 to 7 molecules and around 90 lipid molecules [13, 14, 15, 11]. Variations of this model are the disordered-toroidal pore model [17], "huge toroidal pore" [18, 19, 20] and the chaotic pore model [21, 22].

Aggregate Model The aggregate Model was proposed by Wu et al. [23]. In this model the peptides insert themselves in the membrane, forming micelle-like aggregates with the lipids. This leads to the formation of channels with different sizes and shapes. The model bears some resemblance to the toroidal-pore model but here the peptides don't adopt any particular orientation. It can explain membrane permeabilisation and membrane translocation for several peptides, e.g. polyphemusin [4].

Carpet Model The idea of the model is that the cationic peptides cover the surface of the bilayer attracted by the negative charges of the lipid heads. Upon reaching a threshold concentration, a general disturbance of the bilayer by the peptides in a detergent-like manner leads to the formation of micelles [24, 25, 16]. The peptides are oriented parallel to the membrane surface until they start disrupting the membrane curvature, thus leading to the formation of putative toroidal-like pores and micellization, and finally to the creation of holes in the membrane. Peptides believed to act in this manner are Dermaseptin S, cecropins [26, 27], mellitin, caerin 1.1, ovispirin, Trichogin GA IV, LL-37 [11, 12].

Detergent Model The detergent model is a variation of the carpet model [2]. It explains the action of AMPs by describing them like detergent molecules. These detergent-like molecules have specific interactions with the lipid membrane, especially after reaching Critical Micelle Concentration (CMC). For the amphiphilic peptides, we cannot talk about CMC in the complete sense of the term, but it is known that they can exist as oligomers that can have properties different than the monomeric molecules. The model postulates that depending on the peptide and lipid composition of the membrane the interaction between the two might lead to the disintegration of the membrane.

Sinking raft model In this model, the peptides are supposed to form oligomers that sink in the interior of the membrane and then emerge on the inner side of the membrane. The model was devised to explain how some peptides traverse the bilayer without changing their orientation to perpendicular with respect to the membrane. The oligomers are believed to turn with their hydrophilic sides towards each other,

leaving their hydrophobic parts facing the membrane. In that conformation, they traverse the entire bilayer and appear on the inside of the cell membrane [28, 29, 30]. Proteins that may employ similar kind of actions are δ-lysin [30] and cecropin A [28, 22].

Intracellular Targets

Apart from their activity against the membrane, it is believed that some AMPs have an effect on intracellular processes. Once inside the cell, these peptides have a wide range of different targets. Interfering with the constituents of the cell might lead to flocculation of intracellular contents, alteration of cytoplasmic membrane septum formation, inhibition of cell-wall synthesis, binding to nucleic acids, inhibition of protein synthesis and enzymatic activity [11, 2] It is important to note that even when the peptides are not directly acting on the membrane they still need to cross it in order to reach their targets [4]. Some authors even propose the idea of "multitarget" mechanism [31] in which one highly cationic peptide can bind and interact with the membrane and several anionic molecules inside the target cell. Besides this, it is proposed that different peptides may have different modes of action depending on the peptide concentration, target species, tissue localisation and growth phase of the bacteria. This hypothesis is in correlation with the fact that it is difficult for the pathogens to develop cationic-peptide-resistance [32]. Furthermore, it has been suggested that AMP can modulate the activity of some autolysins [33] and host-derived phospholipases [34], activating these enzymes and leading to lipid damage. Such synergetic activity can have an important role in the innate immune response [11, 35].

1.1.4 Classes of Particulate Interest for This Work

Cecropins

Cecropins are a family of cationic α-helical AMPs initially isolated from the hemolymph of the *Hyalophora cecropia* moth [36, 37]. It was then found in other insects - Lepidopteran and Dipteran. Later, a mammalian cecropin - Cecropin P1 was isolated from porcine intestines [38]. The family consists of 30 to 39 amino acids long peptides with high positive charge. They are very potent antibacterial agents, both against Gram-positive and Gram-negative bacteria. The mammalian analogue Cecropin P is as potent against Gram-negative but has reduced activity against Gram-positive bacteria [27].

The cecropins are believed to act on the cellular membrane by the carpet model [27, 39] but there is also evidence that they affect intracellular processes and transcription in *E. coli* in subinhibitory concentrations [40, 2]. They are induced upon infection [39] and act as broad-spectrum antimicrobials against organisms with anionic membranes, including fungi [41]. Most of the cecropins demonstrate random-coil structure in aqueous solution but form an amphiphilic α-helix, with a hinge in the middle, when cell or model membrane is present [12, 20].

There are two interesting facts about them. The first is that all-D-cecropins, cecropins with inverted (retro) sequences and inverse-D-cecropins (retro-enantio) retain the antibiotic activity of the original molecule [42, 43, 44]. Secondly, their genes are subject to fewer mutations than the rest antimicrobial peptides [45].

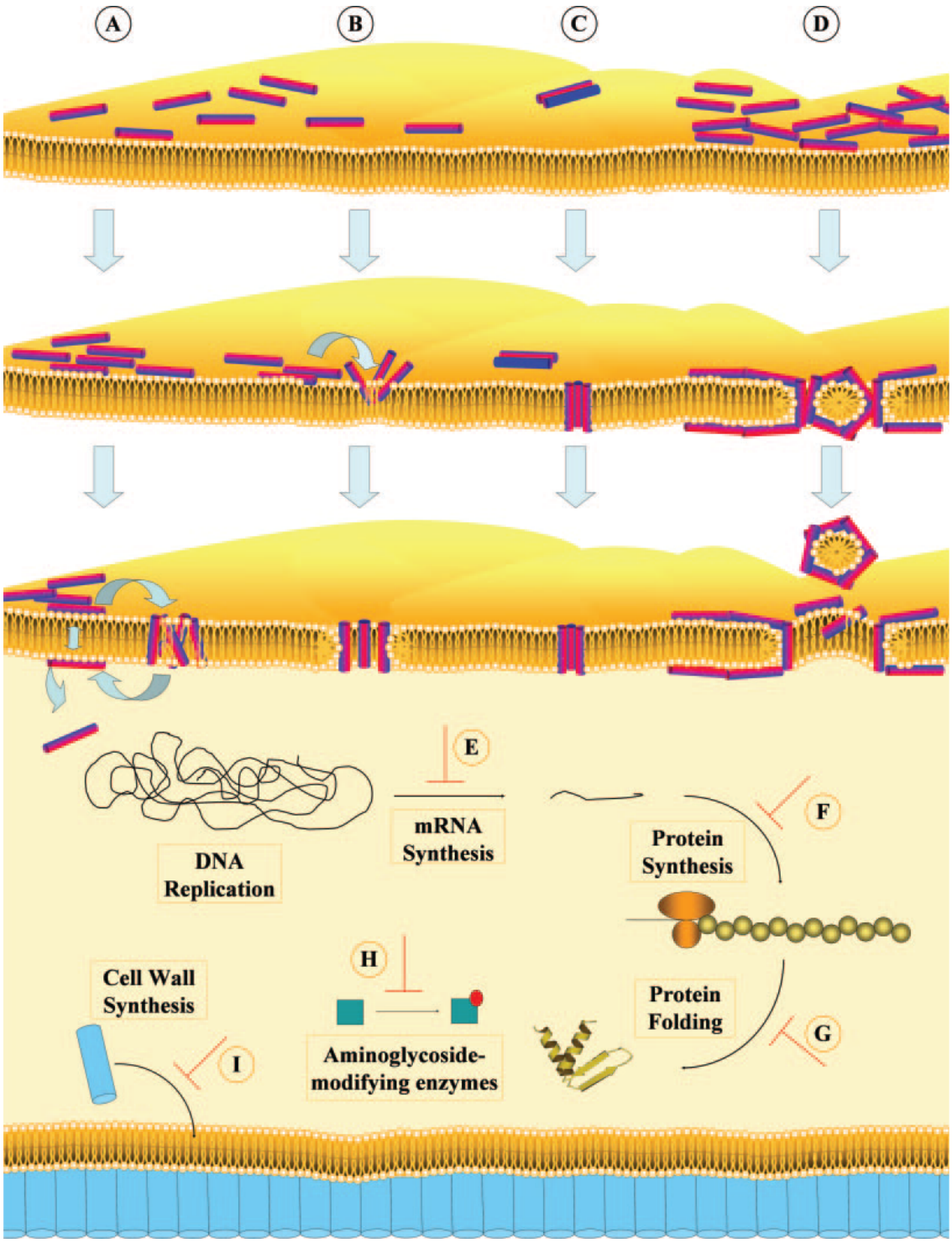


Figure 1.1: Mechanisms of bacterial action. Taken from Jenssen et al. [4]. A - aggregate model, B - toriodal-pore model, C barrel-stave model, D - carpet model, E-I - intracellular targets

Bacteriocins

Bacteriocins are another family of mostly cationic, membrane-permeabilising peptides, varying in length between 26 and 60 amino acids. They are primarily synthesised by bacterial species, but similar peptides are found in plant and animals, including humans, meaning that they are widely distributed in nature. Great interest for the food industry are the bacteriocins produced by lactic acid bacteria. These bacteria are widely used in food fermentation, especially in dairy products. This makes them potential target compounds for food preservatives and therapeutic agent for gastrointestinal infections. One example for such peptide is the Nisin [46, 47].

A group of interest for the current work is class IIa bacteriocins and in particular pediocin-like peptides. They have highly conservative hydrophilic and charged N-terminal having a disulphide bridge and a common YGNGV/L sequence. The C-terminal is more variable and amphiphilic in nature [48]. They have strong anti-*Listeria* activity and kill by permeabilizing the cell membrane [49, 50]. An important aspect of their properties, from a technological perspective, is their thermostability and the retention of activity at a wide range of pH values [48].

Cathelicidins

This family encompasses mammalian AMPs having a common, highly conserved proregion, called Cathelin, and a variable C-terminal antimicrobial domain [51, 52]. As most AMPs they have a positive charge, this is especially true for the C-terminal domain. They are present in most domesticated animals like pigs, goats, cattle, sheep and also in laboratory animals like mice [53]. Cathelicidins are stored in the cytoplasmic granules of neutrophil leukocytes and are released upon activation of the cells or they are secreted by epithelial tissue [54]. In order to release the mature peptide, a proteolytic cleavage separated the N-terminal cathelin domain from the C-terminal active peptide [4]. Some of them assume α-helical conformation, may or may not contain disulphide bonds and some of them are rich in proline, arginine and tryptophan [51]. In humans, they are some of the most important AMPs. Their absence leads to very severe negative effects for the host [55].

Cathelicidins are potent AMPs. The porcine-derived PR-39 is reported to effectively kill bacteria by stopping their Deoxyribonucleic acid (DNA) and protein synthesis [56, 20]. In broth microdilution assay, human-derived cathelicidin LL-37 shows considerable antimicrobial potency with Minimum Inhibitory Concentration (MIC) of less than $10\,\mu g\,mL^{-1}$ against a wide range of Gram-positive and Gram-negative bacteria, even in high salt concentration ($100\,mmol\,l^{-1}$ NaCl). Other clinically important bacterial species like methicillin-resistant *Staphylococcus aureus* and the fungi *Proteus mirabilis* and *Candida albicans* were resistant at high salt concentration but were susceptible to the antimicrobial activity in low-salt media. Cathelicidins are known to change their activity depending on the salt concentration [57, 58]. In *Escherichia coli*, LL-37 is reported to permeabilised both the inner and outer membrane, and to bind to Lipopolysaccharides (LPSs) cooperatively with high-affinity [59]. Moreover, *Candida albicans* hyphae have been treated with fragments of LL-37. As a result, they changed their appearance and the filamentous growth was mostly stopped. The peptide fragments changed the membrane permeability characteristics and are even believed to have activated reaction oxygen species production [60].

Apart from their antimicrobial activity, some of them promote wound healing, by inducing synthesis of syndecans [61]. PR-39 can alleviate myocardial damage after experimental ischemia in rodent models [52] and induce angiogenesis [62]. The human-derived LL-37 is also multifunctional. It stimulates chemotaxis [63] by acting as a receptor ligand [64]. LL-37 is also able to neutralise the action of LPSs [51] and reduce the levels of Tumor Necrosis Factor (TNF) in macrophages [65]. Furthermore, cathelicidins are known to induce the transcription and release of chemokines [66] and the release of histamine by mast cells [67]. Both of this processes supporting the recruitment of different cells of the immune system [68].

1.2 Peptide Databases

Nowadays, there are many sequence and structure databases. The most prominent protein database is UniProt [69] (`http://www.uniprot.org/`) for sequence and annotation information and RCSB PDB [70] (`http://www.rcsb.org`) for structural information. Both of these databases are used in the current project in order to extract information about the antimicrobial peptides. Table 1.1 gives and overview of the AMP databases and web resources that are available today. Unfortunately, most of these specialised databases are outdated and with limited support and functionality. For further review please refer to Torrent et al. [71], Aguilera-Mendoza et al. [72], Liu et al. [1] and `https://omictools.com/antimicrobial-peptide-data-category`.

Table 1.1: List of antimicrobial peptide databases and resources

Database	Summary	Size	Link	Reference
APD3	Natural AMPs with defined sequences and activities	2883	http://aps.unmc.edu/AP/	[73]
AVPdb	Dedicated resource of experimentally verified anti-viral peptides	2683	http://crdd.osdd.net/servers/avpdb	[74]
BaAMPs	Database dedicated to AMPs specifically tested against microbial biofilms	219	http://www.baamps.it/	[75]
BACTIBASE	Calculated or predicted physicochemical properties of bacteriocins produced by both Gram-positive and Gram-negative bacteria	230	http://bactibase.pfba-lab-tun.org/	[76]
Bagel	Web-based application identifing the Open Reading Frames of putative bacteriocins		http://bagel2.molgenrug.nl/	[77]
$CAMP_{R3}$	Contains information on the conserved sequence signatures captured as patterns and Hidden Markov Modelss (HMMs)	8164	http://www.camp3.bicnirrh.res.in/	[78]
Cybase	Dedicated to cyclic peptides	948[1]	http://www.cybase.org.au	[79]

[1]This is the total number of peptides. Not all of them are AMPs

DADP	Database of anuran defense peptides	2571	http://split4.pmfst.hr/dadp/	[80]
DAMPD	Manually curated database of known and putative AMPs. Replacement for the old ANTIMIC [81] database	1232	http://apps.sanbi.ac.za/dampd/	[82]
DBAASP	Antimicrobial activity and structure of peptides	10125	https://dbaasp.org/home	[83]
Defensin knowledgebase	Manually curated database and information source devoted to the defensin family	363	http://defensins.bii.a-star.edu.sg/	[84]
DRAMP	Manually curated database harbouring diverse annotations of AMPs	17508	http://dramp.cpu-bioinfor.org/	[85]
Hemolytik	Manually curated database of experimentally validated Hemolytic and Non-hemolytic peptides	≈ 5000	http://crdd.osdd.net/raghava/hemolytik	[86]
HIPdb	Manually curated database of experimentally validated HIV inhibitory peptides	981	crdd.osdd.net/servers/hipdb	[87]
LAMP	Manually curated database of natural and synthetic AMPs	5547	http://biotechlab.fudan.edu.cn/database/lamp/	[88]
MilkAMP	Database of antimicrobial dairy peptides	371	http://milkampdb.org/	[89]

Peptaibols	Database for the sequences of peptaibols (fungal origin)	317	http://peptaibol.cryst.bbk.ac.uk	[90]
PhytAMP	AMP sequeces from plant origin	271	http://phytamp.hammamilab.org/	[91]
SATPdb	Database of structurally annotated therapeutic peptides		http://crdd.osdd.net/raghava/satpdb/	[92]
Thiobase	Sulfur-rich, highly modified heterocyclic peptide antibiotics	≈ 100	db-mml.sjtu.edu.cn/THIOBASE/	[93]
YADAMP	Short α-helical peptides interacting with cell membranes	2525	http://www.yadamp.unisa.it	[94]

1.3 Molecular Dynamics of AMPs

1.3.1 Overview of Molecular Dynamics of Biomolecules

Molecular Dynamics (MD) simulations are proven and tested method for gathering information about atomic and molecular interactions. They are giving an unprecedented level of detail about the structure of molecular species. And all of this without ever walking in the laboratory or using any expensive materials and reagents.

Despite these advantages, MD have certain drawbacks. One of them is the required computational infrastructure. In order to do complex simulations, having a meaningful time span, researchers require fast and powerful computers. Furthermore, running and analysing complex computer simulations require careful consideration and verification that the results are not unphysical artefacts. This requires experienced staff and can very well be beyond the capabilities of every research lab or experimental scientist [95].

All-Atom Molecular dynamics

In all-atom MD simulations, the approach is to represent each and every atom with an interaction site. This interaction sites or particles have certain mass and volume and their interactions with the rest of the system are governed by an empirical force field. These force fields are the model that governs the simulation. They are the set of potential energy functions that are used to approximate the molecular energy surface. For example, Class I additive potential energy functions have the form:

$$U(\mathbf{R}) = \sum_{bonds} K_b(b-b_0)^2 + \sum_{angles} K_\Theta(\Theta-\Theta_0)^2 + \sum_{dihedrals} K_\chi\left(1+\cos n\chi - \delta\right)$$
$$+ \sum_{impropers} K_i mp(\varphi-\varphi_0)^2 + \sum_{nonbond}\left(\epsilon_{ij}\left[\left(\frac{Rmin_{ij}}{r_{ij}}\right)^{12} - \left(\frac{Rmin_{ij}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\epsilon r_{ij}}\right) \tag{1.1}$$

These kind of equations are similar to those used in early force fields and are still in use in some force fields today. They express the relation of structure $\mathbf{R}$ to the potential energy U and are approximations to the potential energy landscape of the system.

Empirical force fields, based on these rules, are many. Some of the most widely used are CHARMM [96], AMBER [97], OPLS-AA [98], etc.

CHARMM Force Field CHARMM is an additive force and has one of the most extensive coverages of the chemical space. It supports proteins [96, 99], nucleic acids [100, 101, 102], lipids [103, 104, 105], carbohydrates [106, 107, 108, 109]. Furthermore, it has an extension to cover more of the chemical space, especially compounds common in medicinal chemistry called CHARMM General FF (CGenFF) [110]. With this force field researches were able to fold some small proteins from completely unfolded

to native state [111]. The force field has the following functional form:

$$U(\mathbf{R}) = \sum_{bonds} K_b(b - b_0)^2 + \sum_{angles} K_\Theta(\Theta - \Theta_0)^2 + \sum_{Urey-Bradley} K_{UB}(S - S_0)^2 + \sum_{dihedrals} K_\chi\left(1 + \cos n\chi - \delta\right) + \sum_{impropers} K_i mp(\varphi - \varphi_0)^2 + \sum_{nonbondedpairs} \left\{ \epsilon_{ij}^{min} \left[\left(\frac{Rmin_{ij}}{r_{ij}} \right)^{12} - \left(\frac{Rmin_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{\epsilon r_{ij}} \right\} + \sum_{residues} U_{CMAP}(\phi, \psi) \quad (1.2)$$

This potential energy function is based on fixed charges. All internal terms are taken to be harmonic with exception of dihedral angle term, which is a sinusoidal expression. In the Urey-Bradley term, S is the distance between atom A and C in configuration A-B-C and is used in special cases. In the improper dihedral angle term, ω is the (pseudo)-dihedral angle defined by A-B-C-D, when atoms A, B and D are bonded to a central atom C. Both of these terms, Urey-Bradley and improper dihedral, are used to optimise the fit of the energy function to vibrational spectra and out-of-plane motion. Non-bonded interactions are included between atoms with point charges q_i and q_j and the Lennard-Jones (LJ) potential has the 12-6 form. These interactions are calculated for all pairs of atoms within the user specified cut-off distance [112]. The latest implementation of the force field is the CHARMM36.

CHARMM36 Force Field As accurate force field parameters are essential for good MD simulations, all force field authors are trying constantly to improve them. The C36 version improves significantly the representation of the potential energy surface of proteins by two major corrections. First, the backbone Correction Map (CMAP) potential was refined against a range of data for dipeptides and experimental data on small peptides such as hairpins and helices. Second, the side-chain dihedral potential was optimised against quantum mechanical energies from dipeptides and NMR data from unfolded proteins. Further small improvements include revision of the LJ potential for aliphatic hydrogen, improved treatment of parameters for guanidium ions and new parameters for tryptophan.

As a result, the current version has several advantages. It corrects the propensity of the previous version (C22) to overstabilise helices. It also brings a significant improvement to torsion angles for both folded and unfolded proteins [99]. The final refinement of C36 is the C36m force field.

CHARMM36m Force Field The CHARMM36m brings improvement to the backbone CMAP potential. With this improvements, it is able to better simulate intrinsically disordered peptides and proteins. This refinement brings optimisation to the CMAP potential and improved modelling for guanidinium and carboxylate salt bridges [113].

Coarse-Grained Molecular Dynamics

Coarse-grained systems take several atoms of the original structure and replace them with a particle (bead) that represents their overall properties. Doing that reduces

considerably the degrees of freedom of the system. This speeds up computations tremendously because most of the algorithm used in MD are not scaling well. There are different Coarse-Grained (CG) force fields, with a variety of coarse-graining methodology, based on the question asked. For example, the UNRES [114] is a force field used to model interactions involving amino acid side chains. The OPEP force field [115] and the force field from Bereau and Deserno [116] are used for protein folding, protein structural and aggregation studies. Two other widely used biomolecular force fields are the PACE CG force field [117] and the MARTINI force field [118, 119, 120]. For further reading about the topic please refer to Barnoud and Monticelli [121].

MARTINI Force Field The MARTINI force field is maybe the most widely used and thoroughly tested of the CG models. It uses 4-to-1 mapping. This means that, in general, every CG bead represents 4 heavy atoms (4 atoms and the hydrogen atoms bonded to them). In the case of ring systems, the authors of the force field found this mapping is inadequate and used different finer mapping. For example, 2-to-1 mapping in the benzene molecule [118]. In the beginning, this force field was designed and parameterized to study lipids and lipid interaction but was later enhanced with parameters for proteins, as of version 2.1 [119]. Further improvements on the force field were implemented with version 2.2, with reparametrizing some of the amino acids [120]. In the same time, more major classes of molecules were added to the force field: carbohydrates [122], polymers [123, 124, 125], DNA [126], polyelectrolytes [127] and the work on RNA is in beta phase, according to the website of the group, developing the force field `http://cgmartini.nl/`.

There is a total of 18 beads in the model. The four main bead types, according to their polarity, are polar (P), non-polar (N), apolar (C) and charged (Q). Every one of these main types may have a subtype based on its hydrogen-bonding capabilities: donor (d), acceptor (a), donor-acceptor (da) or none (0) or the degree of polarity represented by a number from 1 (low polarity) to 5 (high polarity)

Non-bonded interactions The non-bonded interactions are described by a LJ potential with the form 12-6:

$$U_{LJ}(r) = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r}\right)^{12} - \left(\frac{\sigma_{ij}}{r}\right)^{6}\right] \tag{1.3}$$

The well-depth ranges from $\epsilon_{ij} = 5.6\,\mathrm{kJ\,mol^{-1}}$ for interactions between strongly polar groups and $\epsilon_{ij} = 2.0\,\mathrm{kJ\,mol^{-1}}$ for interactions between polar and apolar groups. The effective size of all particle types is governed by the LJ parameter $\sigma = 0.47\,\mathrm{nm}$. An exception to this parameters are the particles in a ring-like structure. For these particles $\sigma = 0.43\,\mathrm{nm}$ and ϵ_{ij} is scaled to 75% of the normal value.

In addition to the LJ potential, charged particles of type Q interact with shifted Coulombic potential energy function:

$$U_{el}(r) = \frac{q_i q_j}{4\pi\epsilon_0\epsilon_r r} \tag{1.4}$$

Here the relative dielectric constant is equal to $\epsilon = 15$ for non-polarisable water or $\epsilon = 2.5$. For further information about the different types of interaction and interaction matrix, the reader is referred to Marrink et al. [118], Marrink and Tieleman [128]

Bonded interactions Bonded interactions are described with standard energy functions common to all force fields. These include harmonic bond and angle potential, and multimodal dihedral potential. Weak harmonic potential governs the bonds:

$$V_{bond}(R) = \frac{1}{2} K_{bond}(R - R_{bond})^2 \tag{1.5}$$

The equilibrium distance is $R_{bond} = \sigma = 0.47\,\mathrm{nm}$ and a force constant is $K_{bond} = 1250\,\mathrm{kJ\,mol^{-1}\,nm^{-2}}$. Furthermore, the LJ potential is excluded between bonded beads.

The angles are represented by a weak harmonic potential of the cosine type:

$$V_{angle}(\Theta) = \frac{1}{2} K_{angle}\{\cos\Theta - \cos\Theta_0\}^2 \tag{1.6}$$

For aliphatic chains the model uses force constant $K_{angle} = 25\,\mathrm{kJ\,mol^{-1}}$ and equilibrium bond angle $\Theta_0 = 180°$. With *cis* double bond $K_{angle} = 45\,\mathrm{kJ\,mol^{-1}}$ and $\Theta_0 = 120°$. For additional non-bonded parameters, please refer again to Marrink et al. [118]. For Bonded and non-bonded parameters of the proteins refer to Monticelli et al. [119].

Water representation - Polarisable vs. non-polarisable In the standard MARTINI model, the mapping of the water molecules follows the 4-to-1 mapping, used for the rest of the molecules. In this case, 4 water molecules are represented by 1 CG bead. These beads have no electrostatic charge. This means that they are not feeling any electrostatic fields and do not experience any polarisation effects. In order to replicate the electrostatic interactions of all-atom water models a uniform relative dielectric constant with a value of $\epsilon = 15$ is used. Because of the implicit screening, brought by this approach there are artefacts when a polar or charged compound is partitioning into a low dielectric medium, like a lipid bilayer. In the MARTINI force field, two approaches for the introduction of polarizable water model are implemented. The widely used polarizable model of Yesylevskyy et al. [129] and the Big Multipole Water (BMW) model of Wu et al. [130]. The former model adds two charged particles (WP and WM) connected to the central water bead (W) and are constrained by the distance $l = 0.14\,\mathrm{nm}$ to the central bead. The two charged particles bear opposite charges $q = \pm 0.46\,e$. They have no LJ interaction with the rest of the particles and therefore can only take part in Coulomb interactions. The position of the particles compared to one another is governed by a harmonic angle potential with equilibrium angle $\Theta = 0°$ and force constant $K_\Theta = 4.2\,\mathrm{kJ\,mol^{-1}\,rad^{-2}}$. Furthermore, every two particles connected to the same central bead are not feeling each other's charges. The dipole momentum depends on the distance between the particles, which can range from 0 to $2l$. It appears that this model has taken precedence and is more widely used. Moreover, this model is the one that can be downloaded from the official website. For further information about this model please refer to Yesylevskyy et al. [129] The second approach describes the 4 all-atom water molecules with one central site - having a charge of $q = -2\,e$, and two additional sites with $q = 1\,e$. The two additional sites are constrained to $l = 0.12\,\mathrm{nm}$ from the central site (bead) at an angle of $\Theta = 120°$. As in the other model, only the central bead has non-Coulomb interaction with the rest of the system. This interaction is modelled after a modified Born-Mayer-Huggins potential. For further reference see [130]

Recently, a refined polarizable Martini water model was developed. This model is named refPOL and is designed specifically to be used with long-range electrostatic methods like Particle-Mesh Ewald (PME). In this new model, the structure of the polarizable water remains the same, but the charges of the PW and WM particles and the self-interaction between the W beads are reparametrized. This reparametrization brings slightly better results for the mass density and dielectric constant of the water model, especially at T =300 K [131]. With this new parameter sets and refined molecular dynamics parameter options for use with accelerated Graphics Processing Unit (GPU) elaborated by De Jong et al. [132] and further by Michalowsky et al. [131], a man can get an increase in performance and accuracy of the CG simulation.

Coarse grained back mapping

In order to increase the resolution of the system and get more information from the CG model, many researchers are employing inverse mapping. Other terms are backmapping or reverse transformation. This is reversing the CG structure back to all-atom structure and continuing the simulation for a short time with this all-atom topology. In other words, we are letting the system to evolve in CG and when we want more information about the interactions that are happening we are turning to the wealth of information an all-atom simulation can give us. This reversing to an all-atom model can give us great details, especially in protein-lipid and protein-protein interactions. There are two steps to the process. In the first step, an all-atom structure is devised from the CG structure. In the second step, the resulting all-atom structure is relaxed in order to get a structure ready for a simulation. Depending on which step is emphasised, there are two main approaches to the problem. The first is using a very reliable formation of initial structures from the CG. This is done using molecular fragments that come from databases and use statistical scores for the fragments. In most cases, this gives a near-optimal configuration but requires a large database containing enough atomistic fragments and correspondence between the fragments in the database and the structure. The second approach is based on using approximate conversion, based on geometrical considerations and rules or randomly placing the atoms in the vicinity of the corresponding CG beads. This approach is more versatile but requires a subsequent energy minimization and relaxation [133].

Chapter 2

Methods

2.1 Motif Discovery

2.1.1 Protein Database

For the purpose of this master thesis project, the UniProt database [69] was used. UniProt has an intuitive and easy to use interface. Also, it is thoroughly annotated and reviewed. One feature that came to be very handy for the current project is the keywords annotation. As stated in UniProt help manual: "UniProtKB Keywords constitute a controlled vocabulary with a hierarchical structure. Keywords summarise the content of a UniProtKB entry and facilitate the search for proteins of interest." [69].

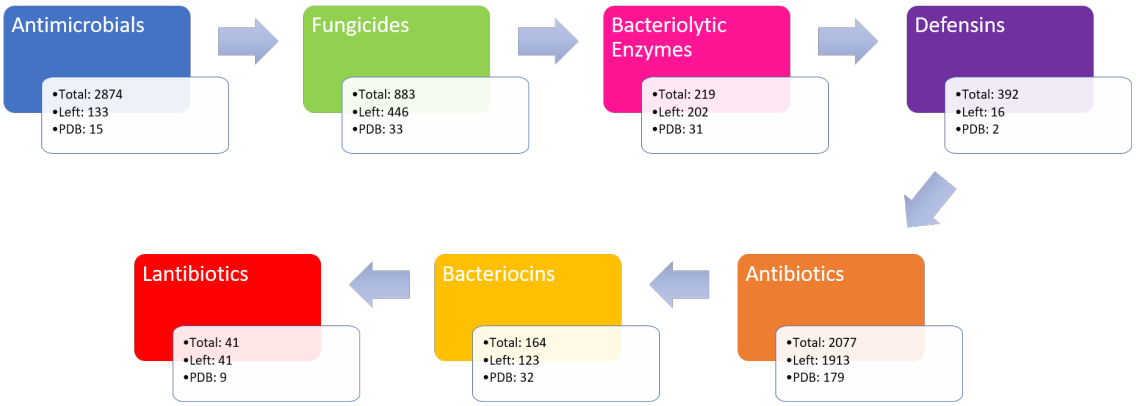


Figure 2.1: Priority order. The numbers represent: Total - the total number of records under the keyword, Left - number of records left after removing the records belonging to keyword with higher priority, PDB - number of records that are left after removing records belonging to keyword with higher priority and also have a PDB cross-reference.

2.1.2 Workflow

Figure 2.2 is a flowchart of the process used for finding the motifs. The first step is to download all of the records under each of the keywords. Downloading is done using the Biopython ExPASy interface which is part of the SeqIO submodule of Biopython [134] version 1.68. After that, the sequence in every record is trimmed, to the part

that is within the CHAIN annotation. If no CHAIN annotation is present, the PEPTIDE annotation is used for the same purpose. Next, all records are separated according to the existence of PDB (3D structure) cross-reference. If the record has a PDB cross-reference, it goes in the first pool of records.
In the first pool, for every record, a vector of structure features is created. This vector includes the length of the sequence, the percentage of each of the three secondary structures: alpha helix, beta sheet and turn and the number of disulphide bonds. So, a vector of five elements is created for every record. An Unweighted Pair Group Method with Arithmetic Mean (UPGMA) tree is constructed based on this features vector. In this tree, every leaf is a UniProt record. In the next step, a recursive traversal of the tree is done, using a depth-first search. At the current node, all the sequences, of the records that are under this node, are aligned.
The alignment is done using T-COFFEE software suite [135] version 11.00.8cbe486. Two substitution matrices and four methods are used. The substitution matrices are Blosum62 and Blosum40, and the methods are PSI-COFFEE, Local Alignment (as implemented in T-COFFEE), M-COFFEE and standard T-COFFEE alignment. PSI-COFFEE associates each sequence with a profile of homologous sequences and then aligns the profiles. M-COFFEE is a meta-aligner that uses eight different aligners: ClustalW, POA, MUSCLE, ProbCons, MAFFT, Dialing-T, PCMA and standard T-COFFEE. The final alignment is a combination of all methods. For further information on each method refer to the T-COFFEE User Manual [136].
When all alignments are ready, the best one is determined using the trimAl tool [137]. Then, every single pair of sequences in this alignment is evaluated for similarity. If all of the pairwise similarities are above 20%, the records corresponding to the sequences form a new cluster. This cluster is then used to form a group. Else the traversal continues with the children nodes of the current node. If the traversal continues, without finding a cluster, until it reaches a leaf (UniProt record), then this single record forms a group of its own.
The result of the traversal is a list of groups. Some of them have a single record for a seed, others a cluster. By a seed, we mean the starting entity from which the group was formed. The next step is to extend each group. For the purpose of this project, an extension of a group will mean to add additional records to the group based on BLAST search [138]. For the groups, that stem from a single record a protein-protein BLAST (BLASTP) is used [139]. If the group stems from a cluster of records, then PSI-BLAST is used [140]. In both cases, the search is against a BLAST database containing all records that don't have PDB (3D) reference. After all of the groups are extended, the motif search begins. For that purpose, the MEME SUITE version 4.11.2 for motif discovery and searching is used [141]. In particular, the program Multiple Em for Motif Elicitation (MEME) [142]. The settings for MEME are to find the best 8 motifs in each group, given that their E-value is above 0.005. If there are fewer than 8 motifs with this E-value, then MEME will return fewer motif hits.

This process is repeated for all of the keywords. Because most of the records are under more than one keyword, every keyword has a priority. The order is shown schematically in fig. 2.1. It begins with the keyword furthest away from the root keyword - Lantibiotic. Lantibiotic is also the only level 3 keyword. Then continues to the parent keywords - Bacteriocin (level 2) then Antibiotic (level 1). After that, it goes through the rest level 1 keywords: Defensin, Bacteriolytic enzyme, Fungicide. In the end, it finishes at the parent node - Antimicrobial.

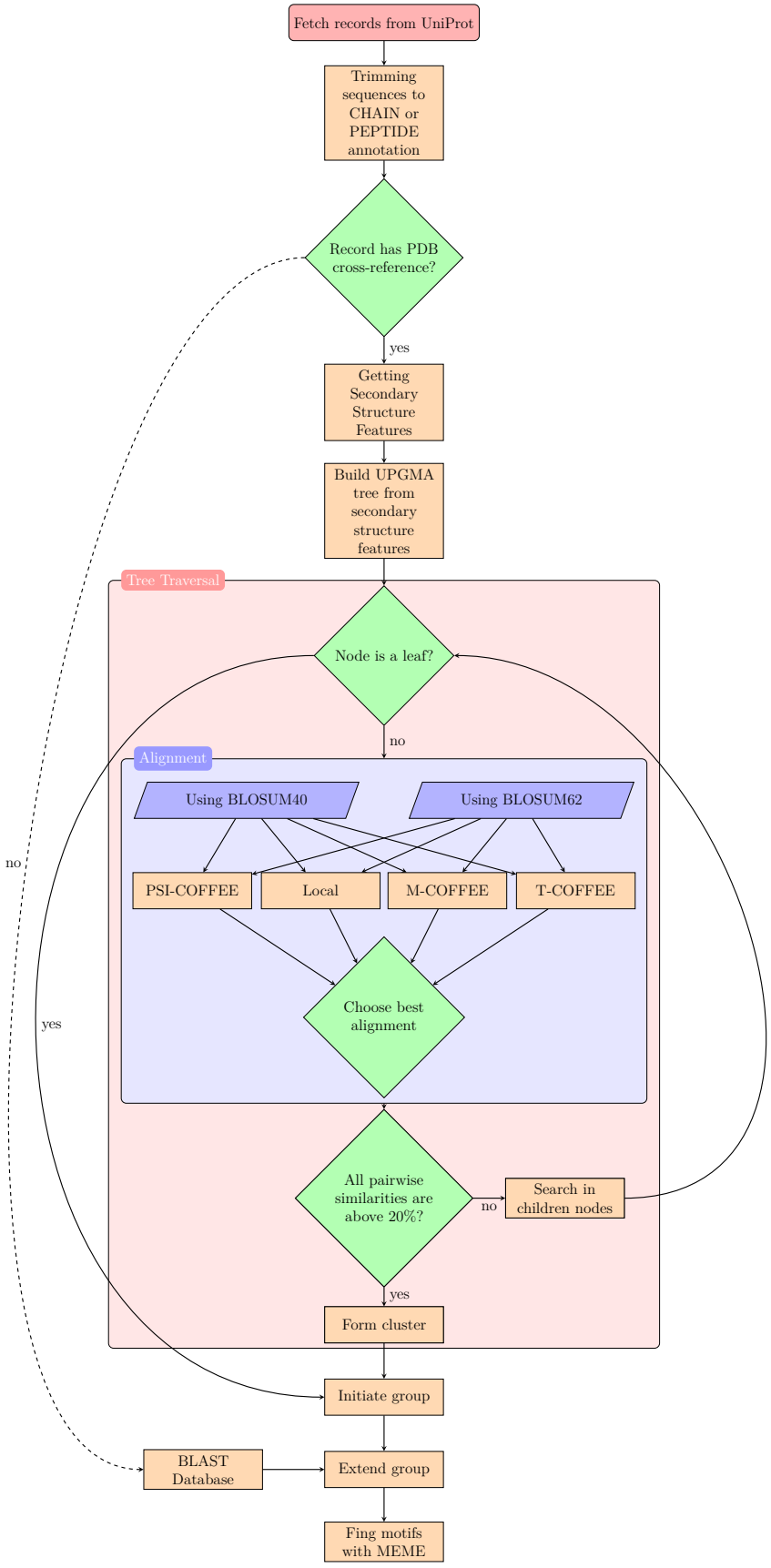


Figure 2.2: Flowchart of the process used for finding the motifs.

2.2 Searching in UniRef

Next part of the thesis project was to find occurrences of the motifs, found in the previous step, in sequence clusters from the UniRef database with 90% identity. For this search first, the motif groups are filtered for Keyword by excluding all Bacteriolytic enzymes. This is done because we are primarily interested in peptides acting on the bacterial membrane by a non-catalytic mode of action. The motifs are then filtered to a maximum length of 40 amino acids and disulphide bonds for the group not more than two. The motifs, that are left, are sorted by the number of occurrences in the set. The four top motif groups are then used to find the occurrences of the motifs in the UniRef90 database. The Motif Alignment & Search Tool (MAST) program from the MEME Suite was used. The UniRef90 database was downloaded from UniProtKB ftp server in fasta format. All records of clusters containing sequences with amino acids ambiguity codes (like B and Z) or unknown (X) were excluded from the fasta file, using an in-house Python script. In MAST an E-value of maximum 1 was used for showing an UniRef90 cluster in the output. Furthermore, the sequence p-values were used (option -seqp) instead of the default position p-value and the p-values and E-values were adjusted for sequence composition of every sequence (option -comp). For further elaboration on the options please refer to the reference manual at `http://meme-suite.org`. The results in XML format were imported in Microsoft Excel. Every table was sorted by the number of motif hits and the top sequence clusters will be used to find sequences containing the antimicrobial motifs more than ones.

2.3 Molecular Dynamics

Four representative structure, one for each motif group, were assayed for their molecular action against a membrane model of E. coli inner membrane. The simulations are done in 2 steps. First, a CG with a subsequent all-atom simulation.

2.3.1 Coarse grained - MARTINI

Initially a CG MARTINI system [118] was build using the Martini Maker module [143] of the CHARMM-GUI web application [144]. The system contains six hundred lipid molecules, with composition and ratio between the lipids in the system taken from the works of Dowhan [145] and Picas et al. [146]. There are three classes of lipids in the system Phosphoethanolamines (PEs), Phosphoglycerols (PGs) and Cardiolipins (CDLs) in ratio 74:19:3. For 37 degrees Celsius the ratio between saturated and unsaturated fatty acids chains is reported to be 1:1. In the end, the lipid composition of the membrane is calculated for six hundred lipids (three hundred lipids in every leaflet) - see Table 2.1 The system has a water on top and on the bottom of the system with height of 2 nm[1]. From every group a representative peptide structure is downloaded in pdb format from the Protein Data Bank `http://rcsb.org`. If for some of the groups the structure, for the entire peptide, is not known, the SWISS-MODEL protein structure homology-modelling server [147] is used to construct 3D structure.

[1] With one exception. There are two runs with the cathelicidin LL-37, one of which is with 4 nm

Lipid species	Number of lipid molecules
1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine (POPE)	81
1,2-dioleoyl-sn-glycero-3-phosphoethanolamine (DOPE)	75
1,2-dipalmitoyl-sn-glycero-3-phosphoethanolamine (DPPE)	75
1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol (POPG)	22
1,2-dioleoyl-sn-glycero-3-phosphoglycerol (DOPG)	19
1,2-dipalmitoyl-sn-glycero-3-phosphoglycerol (DPPG)	19
1', 3'-Bis-[1-palmitoyl-2-vaccenoyl-sn-glycero-3-phospho]-sn-glycerol (PVCL)	9

Table 2.1: Lipid composition of each of the leaflets of the E. coli inner membrane mimicking system

The pdb file is converted to MARTINI CG structure and topology using the CHARMM-GUI Martini Maker. The protein is then inserted, at a random position in a 4 by 4 grid, 1.5 nm above the membrane upper leaflet. The boundary of the membrane is defined as the average Z position of the phosphate groups of the lipids. The insertion is done using the GROMACS version 2016.3 [148] insert-molecules module. The lipid to protein ration is 50:1, i.e. every simulation system has twelve protein molecules. Electrostatic forces are simulated with Fast smooth PME. The distance cut-off is 1.1 nm and Potential-shift-Verlet Coulomb modifier is applied. The relative dielectric constant is set to 2.5 because we use polarisable water model with MARTINI. Van der Waals (VdW) forces are simulated with Twin range cut-offs with neighbour list cut-off and VdW cut-off. As for the electrostatics a Potential-shift-Verlet modifier is applied and VdW cut-off 1.1 nm. These settings are in accordance with the one recommended in MARTINI website for newer versions of GROMACS [132].

The system is energy minimised in two steps. The first is soft-core steepest descent minimisation and then normal steepest descent minimisation. The energy minimised system is run through a single NPT equilibration step. The equilibration step is followed by a production run with a length of 2 µs [2] (see table 2.2 for more information on the settings).

Step	Ensemble	Time step [fs]	Duration [ns]	T-coup	tau-T	ref T [K]	P-coup	tau-P	Compressibility	ref P [bar]
Soft-Core Minimization	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Minimization	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Equilibration	NPT	20	10	v-rescale	1.0	310.15	berendsen	5.0	3e-4	1.0
Production run	NPT	20	2000	v-rescale	1.0	310.15	Parrinello-Rahman	12.0	3e-4	1.0

Table 2.2: Molecular dynamics parameters for course-grained simulations

backwards.py - Reverse Transformation Script

The research group behind MARTINI force field has designed a script using the second approach, for reverse transformation, and made it publicly available on the force field website. The script uses geometrical rules to place the atoms in certain positions in the vicinity of the CG beads. The position of the atoms can be tuned my changing the corresponding beads in a special map file the script uses as an input. The format of the map files is documented in the website. Every molecule needs one mapping file to fully describe the transformation. The script also has an option for a target force field topology file. With this topology provided the

[2] Note all systems were running for this amount of time. Check section 3.3

script is capable of including atoms that are not listed in the mapping file. This is accomplished by placing the missing atom within a small random spatial displacement in comparison to the previous atom. To avoid overlapping atoms there is a small random displacement, the magnitude of which can be adjusted by the user, for every atom in the structure. A graphical representation of the backmapping is provided in fig. 2.3.

The final frame of every CG structure is then converted to an all-atom representation of the system.

2.3.2 All atom simulations

CHARMM36m [113] force field is used for the all-atom simulations. The protein topology files are build using GROMACS pdb2gmx utility from the original all-atom representation of the peptide. The topology files for the lipids are taken from CHARMM-GUI All-Atom Converter from the Martini Maker Module. Unfortunately, there were a few problems with the Automatic All-atom converter at CHARMM-GUI website and it was not used for the backmapping. Firstly, it is not converting back the CDL molecules at their original position. Maybe, there is a problem with the CDL mapping files at the site. That is why a new mapping file, for backward.py, was written. The file backmaps generic CDL molecule from MARTINI to PVCL molecule, for which we have CHARMM36 topology file from CHARMM-GUI web portal. The all-atom simulations are then passing energy minimisation, 2 NVT and 4 NPT equilibration runs. Table 2.3 shows the parameters for each equilibration run. The equilibration is followed by 20 ns production run [3]. Positional and dihedral restraints were applied for the energy minimisation and equilibration runs. For the lipids the position restrains are applied on the phosphate group and dihedral restraints are applied on the glicerol head groups and on the double bonds of the fatty acids, if present in the lipid. The values can be seen in table 2.4 and are implemented as in CHARMM-GUI output files. For the proteins only position restrains are applied, as provided by pdb2gmx, with a value of 1000 kJ mol^{-1} nm^{-2} and are not changed between the equilibration steps.

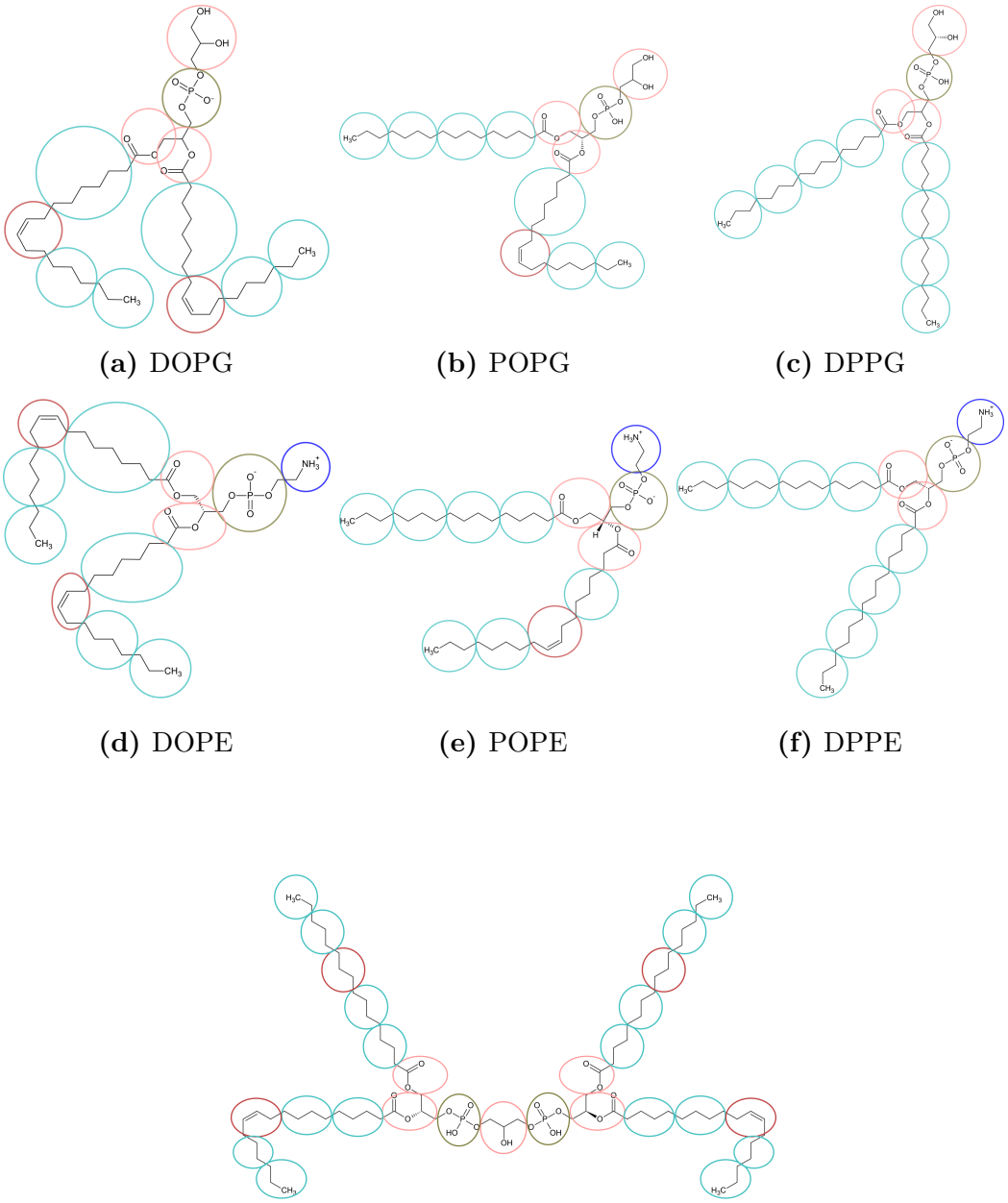
Step		Ensemble	Time step [fs]	Duration [ps]	T-coup	tau-T	ref T [K]	P-coup	tau-P	Compressibility	ref P [bar]
0	Energy Minimization	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
0		NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1	Equilibration	NVT	1	25	Berendsen	1.0	310.15	None	NA	NA	NA
2		NVT	1	25	Berendsen	1.0	310.15	None	NA	NA	NA
3		NPT	1	25	Berendsen	1.0	310.15	Berendsen	5.0	4.5e-5	1.0
4		NPT	2	100	Berendsen	1.0	310.15	Berendsen	5.0	4.5e-5	1.0
5		NPT	2	100	Berendsen	1.0	310.15	Berendsen	5.0	4.5e-5	1.0
6		NPT	2	100	Berendsen	1.0	310.15	Berendsen	5.0	4.5e-5	1.0
7	Production run	NPT	2	20000	Nose-Hoover	1.0	310.15	Parrinello-Rahman	5.0	4.5e-5	1.0

Table 2.3: Molecular dynamics parameters for all-atom simulations

2.3.3 Analysis

Several properties are calculated for both types of simulations. For the CG simulations these are area A and area per lipid A_l. The area per lipid is calculated by dividing the total area A by the number of lipid molecules in leaflet $A_l = \frac{A}{300}$. Partial densities are calculated with *gmx density* for 4 groups:

[3] The reference membrane system is run only for 10 ns



(a) DOPG **(b)** POPG **(c)** DPPG

(d) DOPE **(e)** POPE **(f)** DPPE

(g) PVCL

Figure 2.3: Backmapping of the lipid molecules. CG beads are colour coded: NH3, PO, GL, C, D
. The radius of the circles does not represent the bead radius during a simulation

	Step						
	0	1	2	3	4	5	6
Position	1000	1000	1000	400	200	40	0
Dihedral	1000	1000	400	200	200	100	0

Table 2.4: Restraints values for lipids during all atom minimisation and equilibration. For positional restraints the values are in $\mathrm{kJ\,mol^{-1}\,nm^{-2}}$ and for dihedral restraints in $\mathrm{kJ\,mol^{-1}}$

- Solvent - includes water and ions
- Phosphate
- Membrane - includes all lipids
- Peptides

The box is divided into 100 slices perpendicular to the Z axis. Finally, the minimal distance and number of contacts between the peptides and the 3 types of lipids are calculated using *gmx mindist*. In order to count as a contact, the atoms of the two molecules must be less than 0.20 nm apart for all-atom and 0.5 nm for CG. The analysis for both CG and all-atom simulation is done in the same manner. For the reference system, without peptides, the analysis of the peptides is omitted. All molecular visualisation is done using VMD 1.9.3 [149]. All statistical analysis is done using R [150] and the graphics using ggplot2 [151].

Chapter 3

Results and Discussion

3.1 Motif Discovery

In this work, the Antimicrobial keyword (KW-0929) is a starting point. As already mentioned UniProtKB Keywords have a hierarchical organisation. Child nodes of the Antimicrobial keyword are Antibiotic, Bacteriolytic enzyme, Defensin and Fungicide. There are also the Bacteriocin and Lantibiotic keywords which are child nodes of Antibiotic. Under each of the keywords, there is a certain number of UniProt records. Figure 3.2(a) shows the exact number of records under each keyword and the overlap between them.

The number of motifs found with the workflow elaborated in section 2.1.2 is two hundred. These two hundred motifs are separated among all keywords. Only the Defensin keyword has no motifs associated only with it. Graphical representation of the motifs is shown in fig. 3.3. In this chord diagram, we can see several properties of every motif. If we start from the outermost layer, we can see the keyword that all motifs belong to. Inner to that is the index number of every motif around a ribbon showing the primary association of every motif. Every ribbon is coloured in colour associated with the keyword: red for Lantibiotic, gold for Bacteriocin, dark orange for Antibiotic, blue for Antimicrobial, pink for Bacteriolytic enzyme, lime for Fungicide and purple for Defensin. The light grey overlay on top of each of the ribbons is indicative of a group of motifs. All motifs, which belong to the same overlay, came from one group of records. Every group is one of the groups that are formed by the workflow. The inner layer of the ribbons is a tiles layer. In this layer, there are additional tiles for each keyword. They represent an association of this motif with another keyword, except the primary. The primary keyword is indicated in the ribbon. For example, motif 69 which belongs to the antibiotics is also present in the defensins and the fungicides. In other words, at least one of the sequences in which this motif is present is tagged with at least one of the three keywords: Antibiotic, Defensin, Fungicide. To correctly distinguish the groups remember the colour code.
The next inner layer represents the percentage of records in a group that have this motif. This layer is a heat map with four levels. Dark green means that more that 75% of the records in the group have this motif, while red means that less than 25% have it. This heat map gives a good indication of whether the motif is widespread, and therefore essential for the antimicrobial activity, or it is more probably a sequence motif restricted to a smaller number of sequences within the

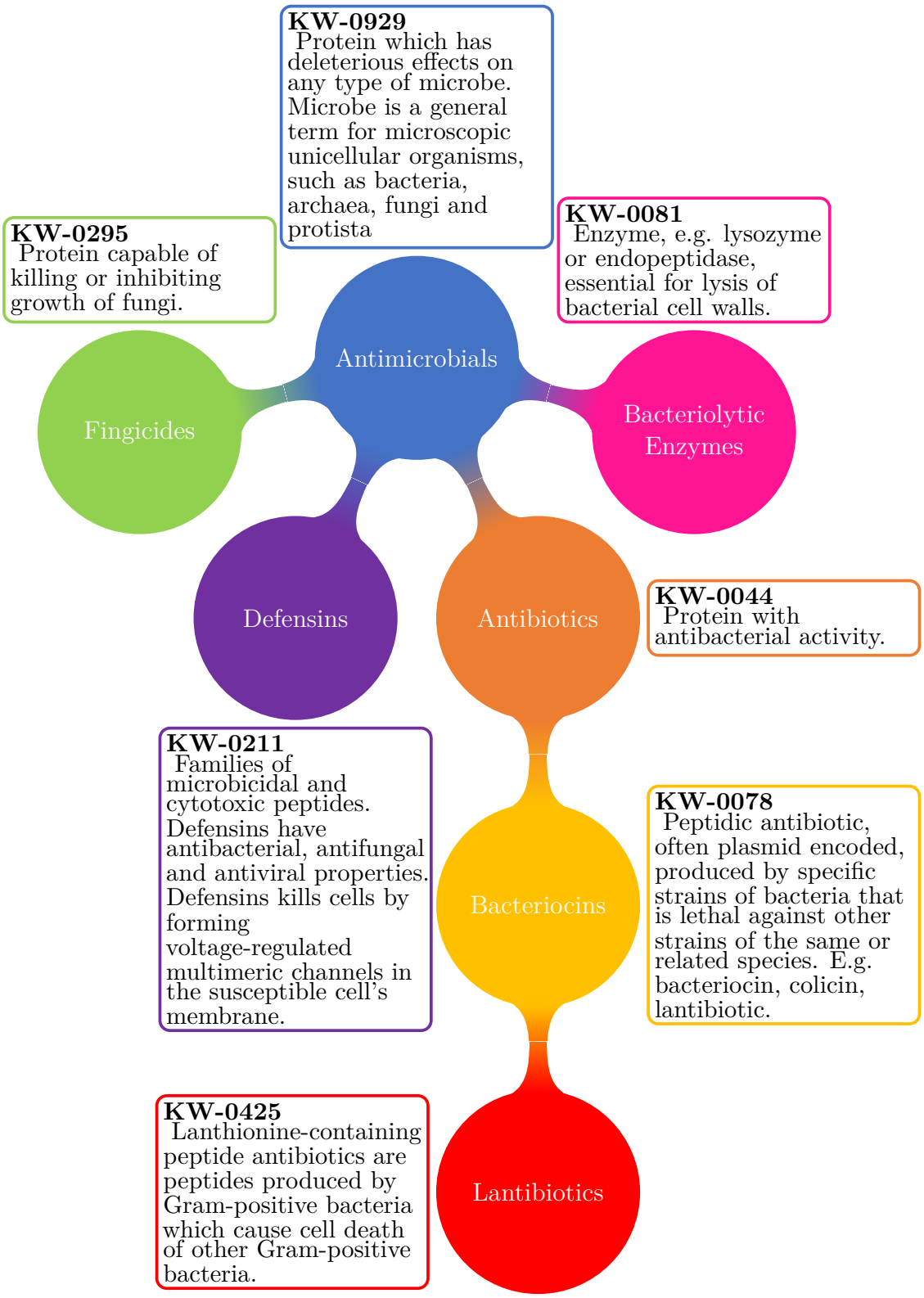
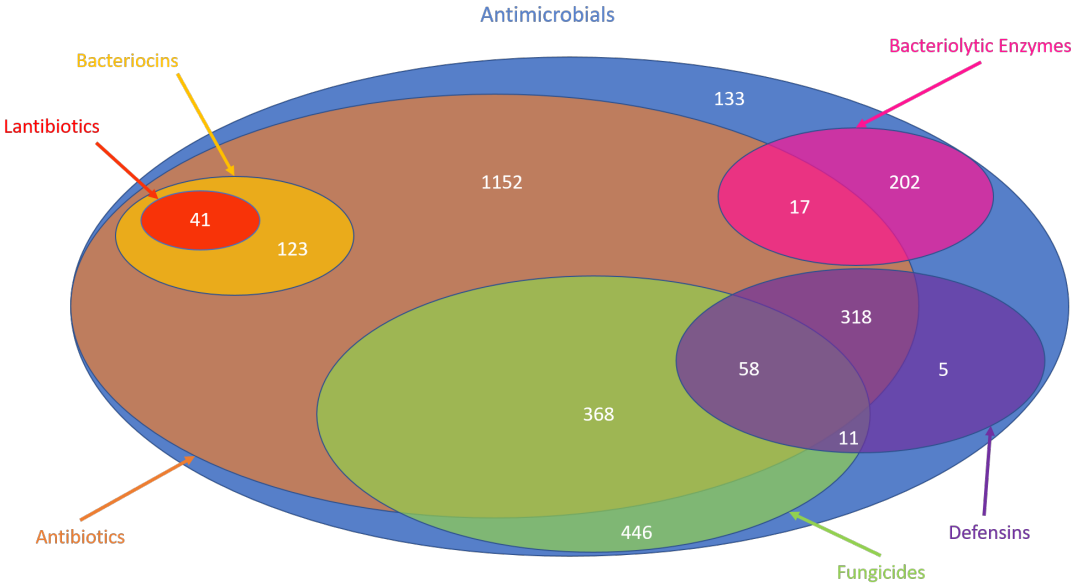
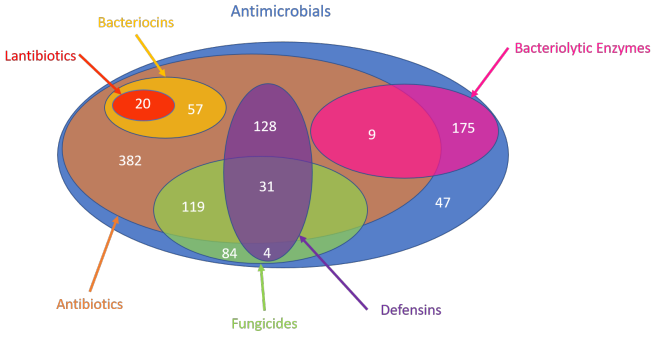


Figure 3.1: Keywords hierarchy in UniProt starting with Antimicrobial.



(a) All Keywords used in the project with number of records.



(b) All Keywords used in the project with number of records that have motifs in them.

Figure 3.2: Eulers diagrams

group. We believe that these motifs restricted to fewer sequences in the group are motifs belonging to very closely related sequences. These sequences have some other function, not connected with the antimicrobial activity, e.g., signal peptide or taxon-specific sequence conveying another information.

The next three layers are bar plots of a particular property of the motif. They are labelled in the diagram. The first one is Length, followed by the absolute number of sequences that have this motif or Hits, and finally the innermost is the Significance of every motif (the absolute logarithmic scale of the E-value). Absolute logarithmic scale means that every bar represents the E-value after the absolute value of the logarithm with base 10 is taken. The transformation above allows us to plot more significant motifs with a higher bar. Furthermore, before plotting any of the bar plots, the data points were searched for outliers using the median absolute deviation and modified Z-score of four. That way outliers will not produce bars with highs overshadowing the rest of the bars.
The inner circle shows the links between the different motifs. By links, I mean the distance, in the sequence space, between them, defined as 1−Pearson Correlation Coefficient (PCC). The less the distance is, the thicker and redder the link between the two motifs. This is valid for distances up to 0.36 where the colour is green, and the transparency is above 88%. The scale used is from red to green. From the diagram, we can see that most of the bacteriocins' motifs are connected to some degree to each other. Also, some of the groups under Antibiotic are also connected. Most prominently, groups of records 37 to 44 and 85 to 92. Furthermore, within the antibiotics, the motifs that also belong to the defensins show some connection too. For example, motifs 81 to 83 and 94 to 96.

Lantibiotics show just one group with a single motif. Bacteriocins are a source of eight groups and a total of thirty-six motifs. The groups under the Bacteriocin keyword are the ones with most relations among them. This could occur because when we are adding records, without PDB cross-reference, to the groups in the extension part of the workflow, it is possible for one or more records to be added to different groups at the same time. Then, in the motif finding step, if two or more records are present in more than one group, MEME will find the same motifs, but in different groups. An indication of that could be seen in the heat map, that shows the percentage of records in which the motif is present. Under Bacteriocin, most of the motifs have a red or orange indication for ratio percentage. Respectively less than 25% or less than 50% of the records in the group have these motifs.
Antibiotic is the keyword that gives the most of the groups and the motifs. Under Antibiotic, there are 35 groups with 95 motifs in total. We can see a strong association between groups of records 37 to 44 and 85 to 92. As with the bacteriocins, we see that under keyword Antibiotic the ratio percentage of the motifs is lower. Furthermore, both groups have the same number of motifs and the same number of motifs that belong to records that also belong to the fungicides. A clear sign that most probably these two groups are erroneously segregated by the workflow and have in them records that have similar features. Two other groups that show strong association are those of records 81 to 83 and 94 to 96. In this groups, all records have high values of ratio percentage. Furthermore, they all belong to three of the keywords: Antibiotic, Fungicide and Defensin. There are some other examples of associated groups, like the single motif groups of records 67 and 120.
The antimicrobials, which are comprised of two groups with a total of six records

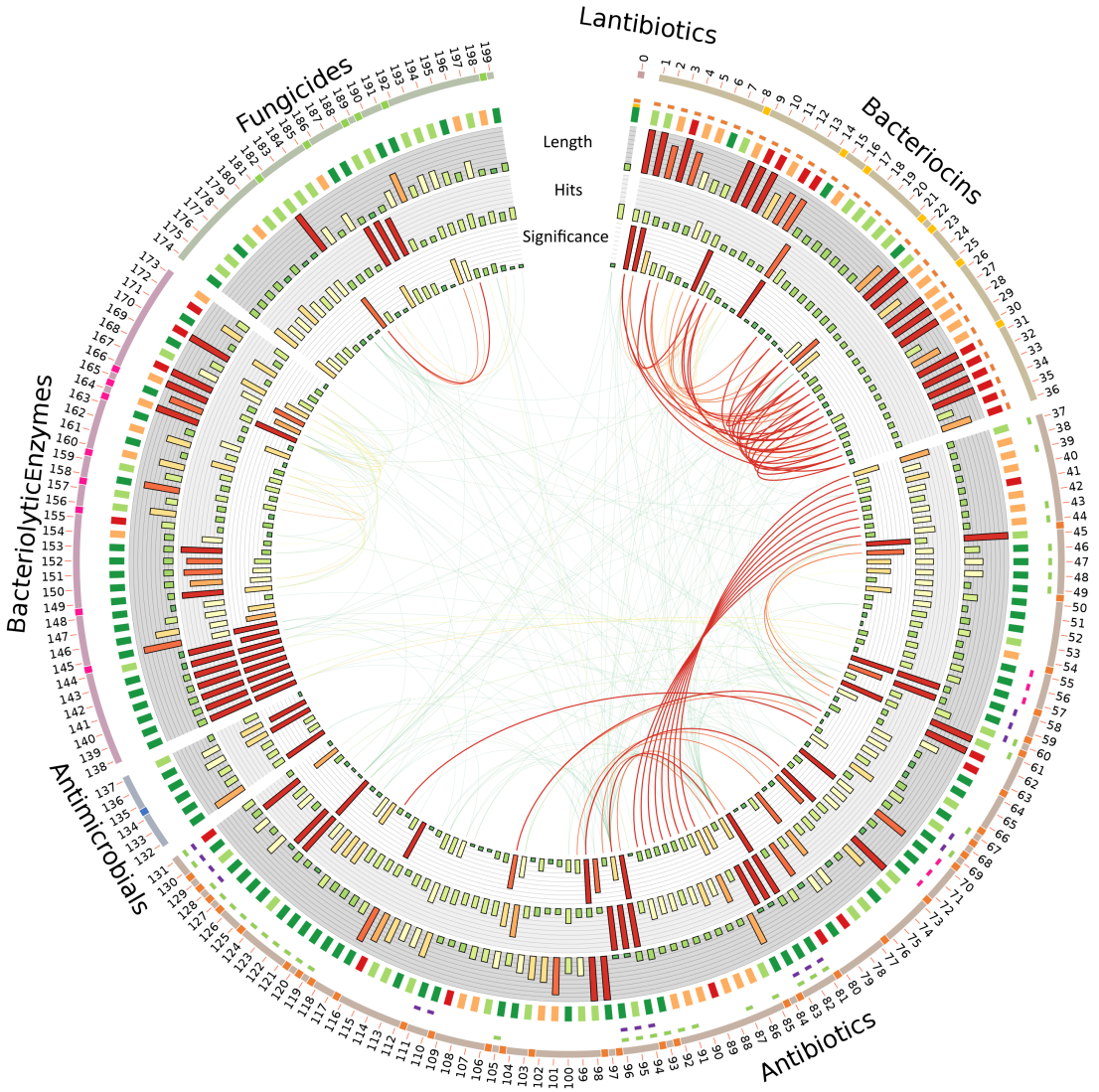


Figure 3.3: Chord diagram of All discovered motifs. For each of the motifs there is (from inside to outside): Significance - given by MEME, Hits - number if sequences where this motif is found, Length - of the motif, colormap of the percentage of sequences in the group that have this motif (dark green - 75% – 100%, light green - 50% – 74%, orange - 25% – 49%, red - 0% – 24%). The outer two layers show to which keyword the records, which have the motifs belong to. All the colours are consistent with the colour code from fig. 3.1

show no strong association with anything. The next keyword, the Bacteriolytic enzyme has nine groups with thirty-six records in total. Again, there are no strong associations. The last keyword is Fungicide. It has 26 motifs distributed among 7 groups. There are no associated groups, just one strong association between records 182 and 194.
In general, we can see that there are two hundred motifs, distributed in 45 groups in total. There are no strong inter-keyword associations between the motifs and the groups. Nevertheless, there are some very prominent intra-keyword associations, but these may well be artefacts from the extension procedure, that is employed to add records with an unknown 3D structure to the groups already formed. However, because both blast searches find the same sequence, there is a possibility that records in these groups are distantly related.
Moreover, the lack of further association and links between the motifs and the groups shows us that there are many different independent antimicrobial peptide sequences and structures. This diversity is present even in this subset of the data present in UniProt. We cannot know how complete the data set is, or how many more antimicrobial sequences are yet to be added to the databases. Because every group has a seed formed by a record or records, with a known 3D structure, we have some idea of the structure of the motifs. This will allow us to make a better prediction for their mechanism.

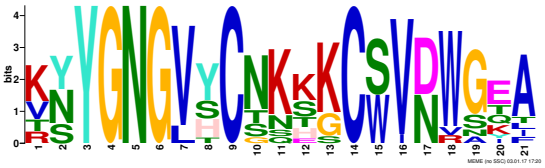
3.2 UniRef Clusters

(a) Motif 125(1)

(b) Motif 126(2)

(c) Sequences

Figure 3.4: Antibiotic-84 Motifs



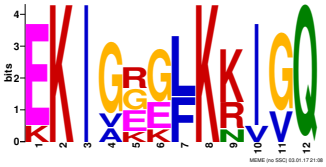
(a) Motif 15(1)

(b) Motif 16(2)

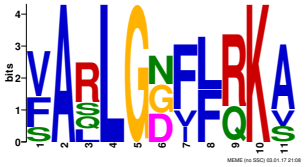
(c) Sequences

Figure 3.5: Bacteriocin-5 Motifs

(a) Motif 64(1)



(b) Motif 65(2)



(c) Motif 66(3)

(d) Sequences

Figure 3.6: Antibiotic-39 Motifs

Antibiotic-84 - Cecropin Family

. This motif group has two motifs: motif 125 (1) and motif 126 (2) (fig. 3.4). The seed for this motif group is the **Cecropin-A** molecule isolated from *Hyalophora cecropia* with UniProt AC P01507. The sequence is 37 amino acids long. The records in this group belong to the Cecropin family of AMPs. The sequences in this group have an average positive charge of 5 and are on average 43 amino acid residues long (table C.1)

Motif 125 is 21 amino acids long and is present in all 18 members of the group. Motif 126 is shorter - just 8 amino acids and is absent in one of the members, with this statistical thresholds. Motif 126 in all sequences is closer to the C-terminus than Motif 125, and in most of the sequences immediately after Motif 125. Exceptions are sequences Q0Q027 - Motif 2 is not present and P83403, P67792, Q06590 - there is a gap of 2 amino acids between the motifs.

Bacteriocin-5 - Bacteriocin class IIA/YGNGV Family

The second motif group is the Bacteriocin-5 group. The records in this group belong to the Bacteriocin class IIA/YGNGV family. Representative protein of the group is **Bacteriocin curvacin-A** from *Lactobacillus curvatus*, UniProt AC P0A311. It has a positive charge of 3 and length of 41 residues. On average the group members have a charge of +3 and are 42 amino acids long (table C.2).

There are again 2 motifs in this group Motif 15 (1) and Motif 16 (2) fig. 3.5. Motif 1 is present in all 13 sequences of the group. Motif 2 is present, after motif one, in just 5 of the sequences P80925, P81053, P29430, A9Q0M7 and B3A0N4. In 4 of them, there is a gap of 4 amino acids and in a single one of 5 amino acids.

Antibiotic-65 and Antibiotic-39 - Cathelicidin Family

These two groups belong to the same family and as we will see group Antibiotic-39 is a subgroup of Antibiotic-65. Group Antibiotic-65 includes motifs 106 to 109 and group Antibiotic-39 includes motifs 64 to 66. Both motif groups represent the Cathelicidin family of AMPs.

Antibiotic-65 has 4 motifs Motif 106 (1), Motif 107 (2), Motif 108 (3), Motif 109 (4) fig. B.1. The representative sequence for the group is **Antimicrobial protein CAP18** from *Oryctolagus cuniculus* (Rabbit) with UniProt AC P25230. Motif 106 is the most represented. It is present in the 8 of the 12 sequences in the group. Furthermore, in 7 of them, Motif 106 is the only motif and it comprises of the entire sequences. This motif is also known as Antimicrobial protein CAP7 or LL-37. The rest of the motifs are present in 5 of the sequence. These sequences are part of the cathelin sequence. Only the representative sequence member of the group P25230 include all 4 motifs in it.

Antibiotic-39 is a group with 3 motifs: Motif 64 (1), Motif 65 (2), Motif 66 (3) fig. 3.6. The members of the groups are actually the members of group Antibiotic-65 which had only Motif 106. Motifs 64 to 66 are sub-motifs of Motif 106. This connection can be seen also in fig. 3.3 and confirm the idea that the workflow is working correctly in finding similar motifs. A representative member of the group is the **Cathelicidin antimicrobial peptide** from *Homo sapiens* with UniProt AC P49913. This peptide is 37 amino acids long, a charge of +6 and is known as LL-37. All members of the group have all 3 motifs with the exception of P51437. The average length of the peptides in this group is 37 amino acids and bear an average charge of +7.

The idea behind searching the motifs in the UniRef90 databases is to find proteins in which our antimicrobial motifs are present, as a group or individually, more than ones. For example, group Antibiotic-84 has 2 motifs, then we are searching for UniRef90 clusters in which there are more than 2 motif hits. In the case, we find such proteins, we can speculate that this protein is some kind of AMP precursor. This precursor might have different purpose and activity. Following proteolytic cleavage, parts of this protein would be released in the environment as AMPs. In the ideal case, this protein would be secreted in high quantities and from an organism or organisms that are easily cultivated. If both the precursor protein and peptidase are commercially available and cheap this could help us mass produce the AMP in hand.

The Bacteriocin-5 search returned 87 hits and the Antibiotic-39 returned 23. None of both groups has a motif count more than the number of motifs in the respective

groups. By motif count I mean the number of motif hits in the target sequence cluster of UniRef90. The search with Antibiotic-84 motifs returned 139 hits. The top hits, that have more than 2 hits are three and can be seen in table 3.1. The Antibiotic-65 search returned 706 hits from the UniRef90 database. From all of this hits 6 have a higher motif count than 4 and can be seen in table 3.2. UniRef90_M7BBJ0 is the cluster with highest motif count of all hits. An overview of this hit can be seen in fig. B.2. Unfortunately, as a result of the scans against the UniRef90 database, none of the top motif groups gave meaningful hits. Meaningful hit here means hit against a protein that is easily produced. The top hit from the Antibiotic-65 group, indeed, has the cathelin motifs 4 times in itself but is not considered a viable target for further research. Furthermore, exactly this motifs are not the ones that bring the antimicrobial activity but are common proregion.

3.3 MD Simulations

For the first group (Antibiotic-84) this is the Cecropin A molecule from *Hyalophora cecropia* fig. 3.8(a). The 3D model is build using Swiss-Model and the template is Papiliocin from *Papilio xuthus* [41]. There are two α-helices between residues Lys 3 - Lys 21 and Ala 25 - Val 36. Between them is a hinge with sequence Ala-Gly-Pro. The overall charge of the molecule at neutral pH is +6 (8 positively and 2 negatively charged amino acids). The N-terminal helix bears more polar residues and is more hydrophobic than the C-terminal helix which is amphipathic.

The representative structure for the second group (Bacteriocin-5) is the Bacteriocin Curvacin A from *Lactobacillus curvatus*. The structure of the molecule was determined by Haugen et al. [47]. It the presence of membrane or membrane-mimicking environment this peptide forms an S-shaped β-sheet-like structure that is supported by a disulphide bridge between residues Cys 10 - Cys 15 fig. 3.8(b). This sheet-like structure is followed by two α-helices. A short one, 6 residues long Arg 19 - Gln 24 and a longer one between Gly 29 - Ala 39. The longer C-terminal helix is more amphipathic and is believed to be the one that anchors the peptide to the lipid membrane of the target cell. These three structures are separated by two hinges: W-V-N and I-I-G that are flexible and provide for movement between the structural elements. The molecule has a charge of +3 (4 positively and 1 negatively charged residues)

The interesting structure of the cathelicidin family is the human LL-37 peptide. The structure of this molecule is taken from the work of Wang [152]. He reports an amphipathic α-helical region covering from residues Leu 2 - Leu 30. This helix is followed by a short unstructured hydrophilic region fig. 3.8(d). The molecule has a charge of +6 (11 positively charged and 5 negatively charged residues)

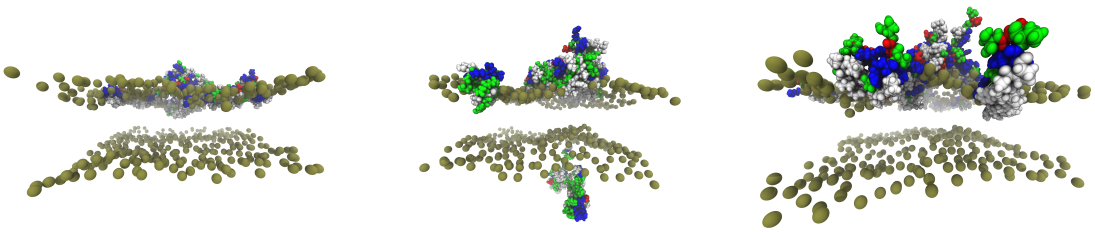
All three of the molecules show their amphipathic properties by quickly interacting with the membrane or with the other peptides in the simulation box. As can be seen from the simulation snapshots fig. A.1 and **????**, the molecules are aggregating in the first 10 ns of the CG simulation. On visual inspection, we can see that the peptides have different propensity to stay together on the surface of the membrane. After the initial aggregation and contact with the membrane, the Cecropin A molecules continue to float on top of the surface, forming aggregates of several molecules fig. A.1. These aggregates appear stable and neither fuse together nor break apart. There are also some single molecules floating on the surface of the membrane.

ID	Tax	Tax Common Name	TaxID	Representative ID	Sequences in cluster	Comment	Motif Count	Length
UniRef90_UPI00067C1674	Amyelois transitella	Navel orangeworm moth	680683	UPI00067C1674	1	PREDICTED: uncharacterized protein LOC106130664	4	166
UniRef90_A0A0L0BZD8	Lucilia cuprina	Green bottle fly	7375	A0A0L0BZD8_LUCCU	1	Uncharacterized protein	3	130
UniRef90_A0A0L0C212	Lucilia cuprina	Green bottle fly	7375	A0A0L0C212_LUCCU	1	Uncharacterized protein	3	180

Table 3.1: Antibiotic-84 Cluster hits with more than 2 motifs

ID	Tax	Tax Common Name	TaxID	Representative ID	Sequences in cluster	Comment	Motif Count	Length
UniRef90_M7BBJ0	Chelonia mydas	Green sea-turtle	8469	M7BBJ0_CHEMY	1	Uncharacterized protein	13	538
UniRef90_UPI000440726D	Balaenoptera acutorostrata scammoni	North Pacific minke whale	310752	UPI000440726D	1	PREDICTED: uncharacterized protein LOC103016278	5	329
UniRef90_L9KXF7	Tupaia chinensis	Chinese tree shrew	246437	L9KXF7_TUPCH	1	Brain-specific homeobox protein like protein	5	1699
UniRef90_S9W698	Camelus ferus	Wild bactrian camel	419612	S9W698_CAMFR	1	Cathelicidin antimicrobial peptide-like protein	5	247
UniRef90_A0A151P502	Alligator mississippiensis	American alligator	8496	A0A151P502_ALLMI	1	Cathelicidin-related peptide Oh-Cath-like	5	240
UniRef90_A0A1I8G1K2	Macrostomum lignano	None (flatworm)	282301	A0A1I8G1K2_9PLAT	1	Uncharacterized protein	5	2981

Table 3.2: Antibiotic-65 Cluster hits with more than 4 motifs



(a) Cecropin A **(b)** Curvacin A **(c)** LL-37

Figure 3.7: Snapshots of the last frame of all-atom simulations

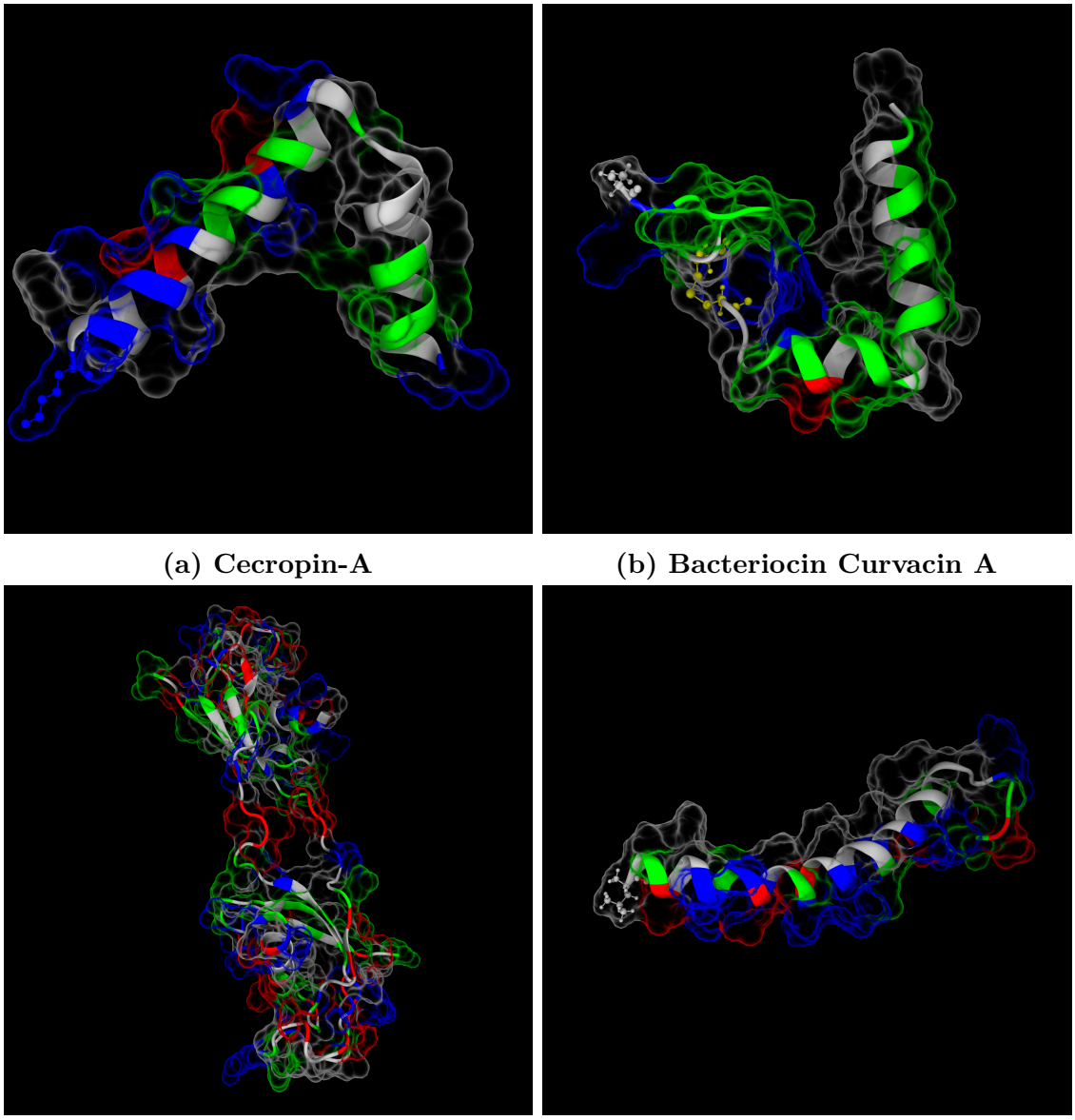
The Curvacin A molecules show different behaviour fig. A.2. They have the tendency to continue to aggregate during the simulation, forming clusters that merge together. In the first 1 µs of the simulation all molecules, except one are bound in a single globular (amorphic) cluster that continues to float on top of the membrane.

The LL-37 molecules show similar behaviour to that of their Curvacin counterparts fig. A.3. However, they don't form a globular cluster but most of them tend to form an elongated fiber-like structure on top of the membrane. This structure becomes seemingly stable holding its shape until the end of the simulation. The few molecules that are not included in the fiber-like structure still appear to be associated with it, orientating themselves perpendicular to it.

After the backmapping and during the all-atom simulations there are not significant changes in the peptide structure and appearance. Thus the simulation window seems to be short. It remains interesting to find out whether during a longer all-atom simulation the peptides will develop a different kind of interaction with themselves and the lipids. Furthermore, the all-atom simulation appears to give much greater detail in the dynamics of the membrane. The peptides at the top leaflet appear to promote the formation of a dent on the opposite side and we can speculate that during longer simulation this dent can be transformed in a pore of some kind - fig. 3.7.

3.3.1 Thickness and area of the membrane

The thickness of the membrane is taken every single ns for the duration of both CG and all-atom simulations. Graphical representation of the thickness with the evolution of the systems is shown in fig. 3.9. For the CG simulations we see that all systems start with a thickness of around 3.9 nm. The reference system that has no peptides in it has a thickness of 3.94 nm. In all other systems, the membrane is thinner. The membrane with Curvacin molecules has a thickness of 3.88 nm, followed by the LL-37 system with shorter Z axis - 3.84 nm. The Cecropin A-treated membrane has a thickness of 3.82 nm. The thinnest membrane is the one exposed to LL-37 in a cubic box with the long Z axis. Its thickness is 3.76 nm. The fact that the LL-37 molecules are making the membrane thinner than the rest is also visible in the box plot of the thickness fig. 3.9(d). We observe that 'LL-37 short' has a long tail of lower laying outliers and despite the fact that the average thickness of this sample is bigger than the one for Cecropin A it could be an artefact from the box size. An Analysis of Variance (ANOVA) gives a p-value of less than $2e^{-16}$ showing a statistical significance of the difference in the thickness. Furthermore, all pairwise between a AMP-treated membrane system and the reference system give again p-value of less than $2e^{-16}$. In fig. 3.9(b) we can get an idea of the timescale of



(a) Cecropin-A **(b) Bacteriocin Curvacin A**

(c) Cathelin **(d) LL-37**

Figure 3.8: Cartoon and surface representations of the reference peptides from each of the working groups. (b) **Bacteriocin Curvacin A** RCSB PDB record 2A2B. (a) **Cecropin-A** SwissModel Homology Model with RCSB PDB 2LA2 as a template. (c) **Cathelin** SwissModel Homology Model with RCSB PDB 1LXE as a template. (d) **LL-37** RCSB PDB record 2K6O. The colour code is: blue - positively charged side chain, red - negatively charged side chain, green - polar uncharged side chain, white - hydrophobic side chain. In (b), (a) and (d) the N-terminal amino acid and the cystein residues forming double bonds in (b) are also represented with a CPK model.

the membrane thinning. As expected there are fluctuations in the thickness but it is visible that the thinning is happening in the first 250 ns of the simulation. After that point, the thickness for every system is fluctuating around the average value.

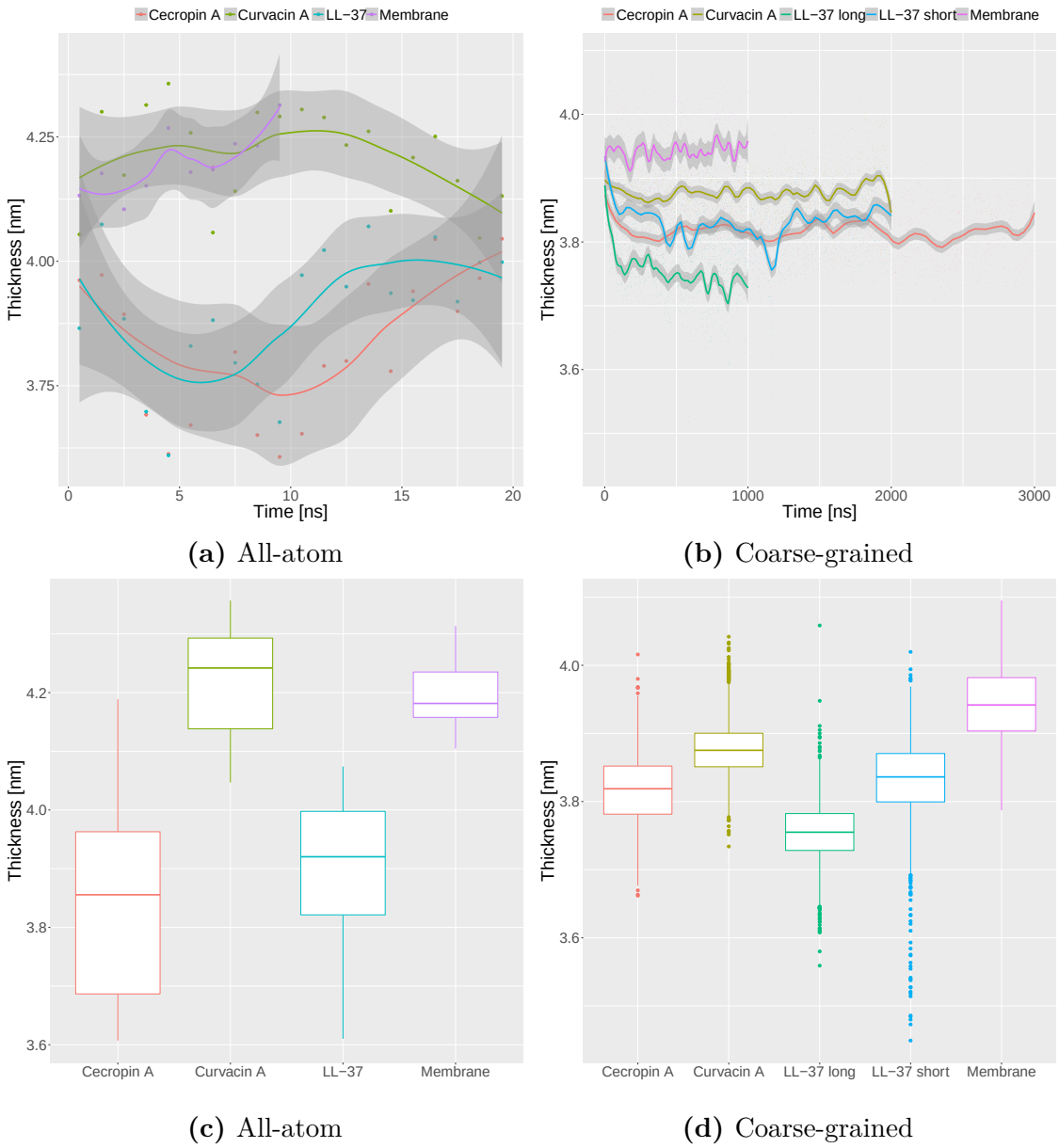
After backmapping of the CG systems to the all-atom representation, the difference between the thickness of the membranes is largely preserved. An unexpected result is the Curvacin-treated membrane which shows greater thickness (4.24 nm) compared to the reference membrane (4.18 nm). This could be attributed to the different length of the simulations. In fig. 3.9(a), the thickness of the reference membrane is increasing and has not reached an equilibrium. The same can also be said for the rest of the systems. This means that longer time scales should be reached in order to assure correct properties for the system. The other two systems, Cecropin and LL-37, are again with thinner membranes 3.86 nm and 3.92 nm, respectively.

The value for the thickness of the all-atom representation of the reference membrane is in good agreement with the result of Venable et al. [153]. They report the P-P thickness of bilayer build only from POPE to be 4.16 nm. The two systems are not equivalent because the one presented here is a mixture of seven lipid species. Nevertheless, the model membrane presented here has the same fatty acid composition, with the exception of the vaccenoyl fatty acid in PVCL and 77% of the lipid head groups are phosphoethanolamines. Thus the systems are close enough for us to be able to compare them.

The area per lipid is calculated by simply dividing the total area of the simulation box by the number of lipids in a leaflet. This simple approximation works well for membrane composed of one lipid. In our case, there are three factors that distort the measurement. First, in the reported simulations there is a mixture of lipids. Second, this calculation does not take into account the peptides that are inserted in the upper leaflet. In theory, these peptides should increase the area of the upper membrane. And lastly, it is not taking into account the bending or the curvature of the membrane. If the lipids are not parallel to the Z axis we can expect the measurement to be incorrect. Nevertheless, this method of calculation of the membrane is very easy and straightforward, allowing us to compare between the different systems. Furthermore, because it is derived directly from the total area of the membrane we can think of it as a scaled version of this total area.

In the graphs of the area fig. 3.10 we see that the presence of AMPs brings with it increase in the area and area per lipid. In the CG simulations the Cecropin and LL-37 have the same average area of 0.73 nm^2.They are followed by the Curvacin with an average area of 0.71 nm^2. The reference membrane has an area of 0.69 nm^2. The LL-37 simulation conducted with shorter Z axis was deliberately omitted from the area graph as it showed great disturbance of the membrane during all-atom simulation and thus was considered inadequate. The all-atom simulation brings decrease in the area but the order of the systems remains almost the same. The Cecropin and the LL-37 have larger areas 0.64 nm^2 and 0.65 nm^2, respectively. The Curvacin and the reference membrane have the same area of 0.59 nm^2. Venable et al. [153] reports all-atom simulations with the value of 0.59 nm^2 for POPE only membrane and 0.68 nm^2 for POPG only membrane. The membrane system in the current project are in good agreement with them and the experimental values of 0.60 to 0.61 nm^2 for POPE [154], and 0.64 nm^2 [155] - 0.66 nm^2 [156] for POPG

The distribution of the Curvacin, however, seems to extend to lower values than that of the reference membrane. This effect can again be attributed to the fact that



(a) All-atom **(b)** Coarse-grained

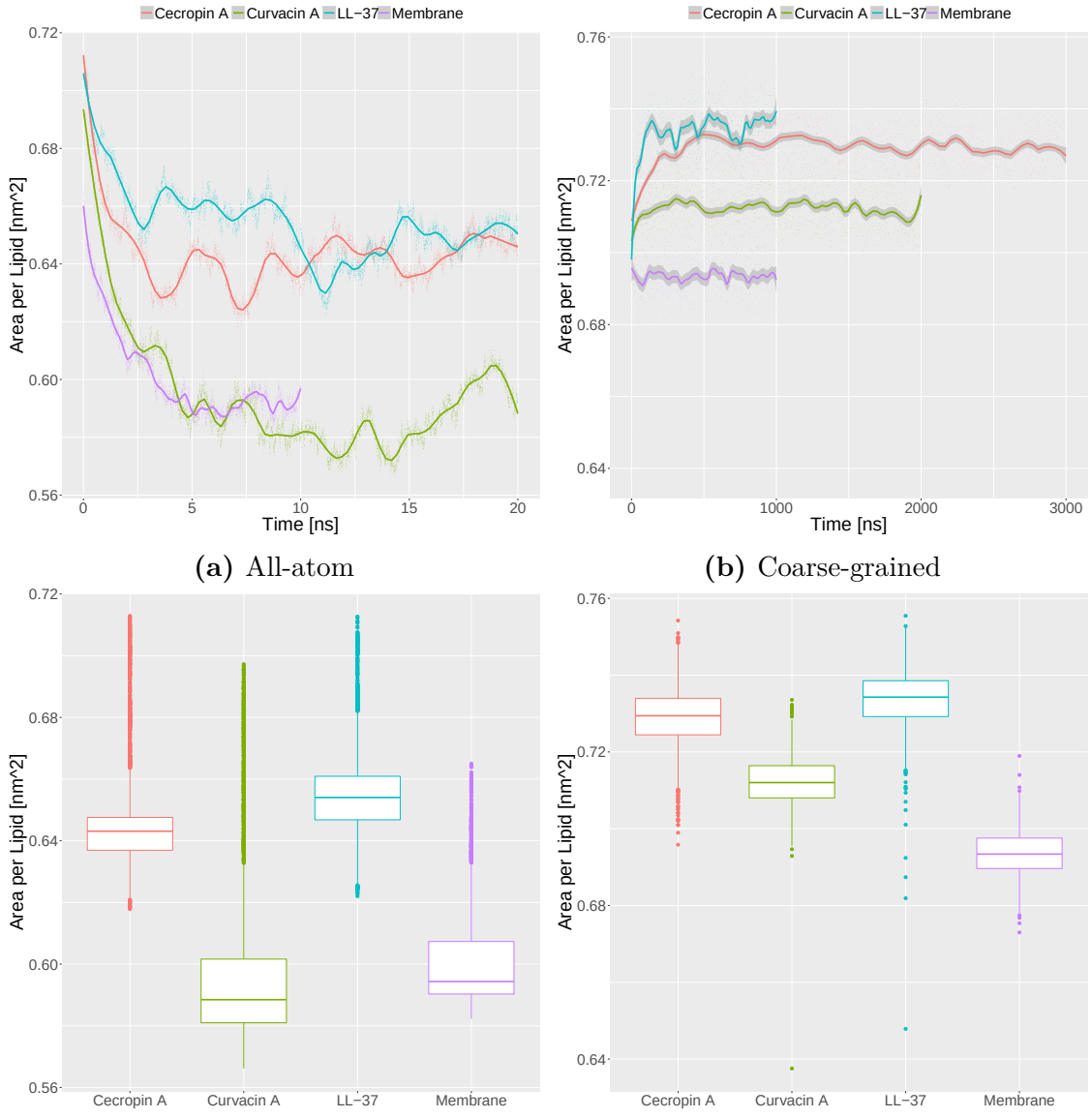
(c) All-atom **(d)** Coarse-grained

Figure 3.9: Thickness (P-P distance) of the membrane during the simulations

the reference membrane simulation is half as long as the rest all-atom simulation. In fig. 3.10(a) we can also see the relaxation time of the area after the backmapping. It appears that in the first 5 ns of the simulations the membrane is reaching an equilibrium. The effect of this is visible on the box plots fig. 3.10(c) as a big number of outliers with higher than average value. As with the thickness analysis, there is high statistical significance for the difference between the simulation systems (p-value $< 2e^{-16}$ for all cases).

An observation we can make is that all thinner membranes have a larger area and vice versa. This gives confidence that these relatively simple calculations of the thickness and of the area of the membrane are reliable and are not distorted by bending or twisting of the membrane. In conclusion, it appears that the more positively charged peptides, Cecropin and LL-37, have greater membrane-thinning properties. In the CG simulations, the Curvacin-treated membrane shows lesser

propensity to get thinner. Interestingly, after the backmapping, the Curvacin system appears to be with a thicker membrane than the reference system. This, however, might be an artefact from the difference in simulation time and to be proven requires analysis of longer trajectories.



(a) All-atom **(b)** Coarse-grained

(c) All-atom **(d)** Coarse-grained

Figure 3.10: Area per Lipid during the simulations

3.3.2 Contacts counting

If there are no preferable interactions between the peptides and the three different lipid classes we can assume that on average the lipids in contact with the peptides will follow the same ratio as the overall ratio of lipid classes in the membrane. The lipid classes ratio, as already mentioned in section 2.3.1 is 74 PEs : 19 PGs : 3 CDLs. From the assumption we made follows that if there is any preference for some of the lipid classes to be in contact with the protein we will see a difference this ratio.

Figure 3.11 shows a stacked bar plot of the lipid classes in contact with the proteins. The Ref group represents the distribution of lipid classes in the membrane. We can clearly see that in all simulated systems the average ratio of CDLs in contact with the proteins is at least 5 times higher than the reference ratio. For the LL-37, there is almost 5 fold increase in the number of CDL molecules than there would be if no interaction is happening. The Cecropin and Curvacin systems express an even higher increase in the association of CDLs with the peptides.

This effect is expected as the peptides in this study are cationic. The Cecropin and the LL-37 molecules both have a positive charge of 6 and the Curvacin positive of 3. In the same time, the CDL molecule has a negative charge of -2. Electrostatic interaction favour the contact between the CDLs and the peptides. The PGs also bear a negative charge of -1. Yet they are not as associated as the CDLs with the peptides. Furthermore, if the association between a peptide and a lipid was purely based on electrostatic attraction then we would expect for the Cecropin and LL-37, with their charge of +6, to recruit more CDLs than the Curvacin, which has half that charge. This leads to the idea that the interplay between the peptides and the lipids is more complicated.

The recruitment of one class of lipids, especially the CDLs may have consequences for the bacterial cell. Their recruitment and binding to the AMPs means that they will be depleted in the rest of the membrane, especially as we know that they comprise just 3% of the lipids in the inner membrane. The interplay between this and other suggested antibacterial mechanisms might enhance the activity of the AMPs.

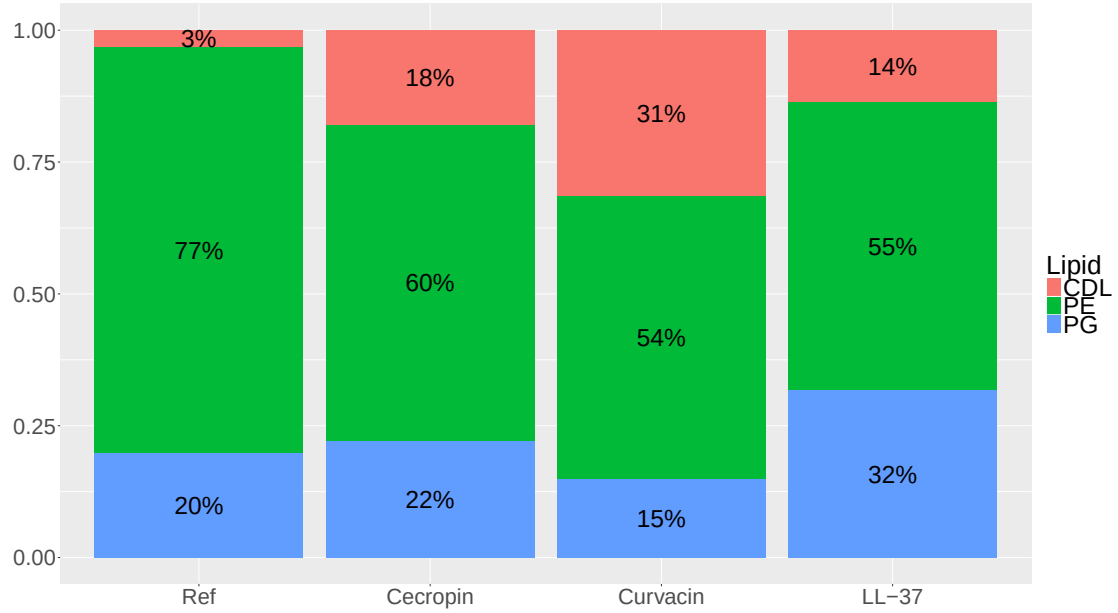


Figure 3.11: Average percentage of contacts between protein molecules and lipids

3.4 Conclusions

In overall we can draw the following conclusions for the project:

- Based on the motif search the peptides listed as Antimicrobial in the UniProt

database show great diversity in structure and sequence with minimal overlap between the different groups found in the project.

- Based on the search of the motifs against the UniRef90 database no peptide clusters having repeated occurrences of the motifs, defined by the three groups of interest, were found.
- In MD simulations of the peptides interacting with a model of *E. coli* inner membrane a statistically significant reduction in the thickness of the membrane and increase in the area is observed. The exception to that is the bacteriocin Curvacin A which shows inconclusive results and will need further research.
- The cationic peptides are discriminating between the different lipid types in the model membrane of *E. coli* predominantly binding themselves to cardiolipins.

Acronyms

AMP Antimicrobial Peptide.

ANOVA Analysis of Variance.

BMW Big Multipole Water.

CDL Cardiolipin.

CG Coarse-Grained.

CMAP Correction Map.

CMC Critical Micelle Concentration.

DNA Deoxyribonucleic acid.

DOPE 1,2-dioleoyl-sn-glycero-3-phosphoethanolamine.

DOPG 1,2-dioleoyl-sn-glycero-3-phosphoglycerol.

DPPE 1,2-dipalmitoyl-sn-glycero-3-phosphoethanolamine.

DPPG 1,2-dipalmitoyl-sn-glycero-3-phosphoglycerol.

GPU Graphics Processing Unit.

HMM Hidden Markov Models.

LJ Lennard-Jones.

LPS Lipopolysaccharide.

MAST Motif Alignment & Search Tool.

MD Molecular Dynamics.

MEME Multiple Em for Motif Elicitation.

MIC Minimum Inhibitory Concentration.

PCC Pearson Correlation Coefficient.

PE Phosphoethanolamine.

PG Phosphoglycerol.

PME Particle-Mesh Ewald.

POPE 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine.

POPG 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoglycerol.

PVCL 1', 3'-Bis-[1-palmitoyl-2-vaccenoyl-sn-glycero-3-phospho]-sn-glycerol.

TNF Tumor Necrosis Factor.

UPGMA Unweighted Pair Group Method with Arithmetic Mean.

VdW Van der Waals.

Bibliography

[1] Liu, S., L. Fan, J. Sun, X. Lao, and H. Zheng
2017. Computational resources and tools for antimicrobial peptides. *Journal of Peptide Science*, 23(1):4–12.

[2] Haney, E., S. Mansour, and R. Hancock
2017. Antimicrobial peptides: An introduction. *Methods in Molecular Biology*, 1548:3–22.

[3] Fjell, C., J. Hiss, R. Hancock, and G. Schneider
2012. Designing antimicrobial peptides: Form follows function. *Nature Reviews Drug Discovery*, 11(1):37–51.

[4] Jenssen, H., P. Hamill, and R. Hancock
2006. Peptide antimicrobial agents. *Clinical Microbiology Reviews*, 19(3):491–511.

[5] Baumann, G. and P. Mueller
1974. A molecular model of membrane excitability. *Journal of Supramolecular and Cellular Biochemistry*, 2(5-6):538–557.

[6] Boheim, G.
1974. Statistical analysis of alamethicin channels in black lipid membranes. *The Journal of Membrane Biology*, 19(1):277–303.

[7] Ehrenstein, G. and H. Lecar
1977. Electrically gated ionic channels in lipid bilayers. *Quarterly Reviews of Biophysics*, 10(1):1–34.

[8] Laver, D.
1994. The barrel-stave model as applied to alamethicin and its analogs reevaluated. *Biophysical Journal*, 66(2 I):355–359.

[9] He, K., S. Ludtke, H. Huang, and D. Worcester
1995. Antimicrobial peptide pores in membranes detected by neutron in-plane scattering. *Biochemistry*, 34(48):15614–15618.

[10] Spaar, A., C. Münster, and T. Salditt
2004. Conformation of peptides in lipid membranes, studied by x-ray grazing incidence scattering. *Biophysical Journal*, 87(1):396–407.

[11] Brogden, K.
2005. Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria? *Nature Reviews Microbiology*, 3(3):238–250.

[12] Ahmad, A., E. Ahmad, G. Rabbani, S. Haque, M. Arshad, and R. Khan
2012. Identification and design of antimicrobial peptides for therapeutic applications. *Current Protein and Peptide Science*, 13(3):211–223.

[13] Matsuzaki, K., K.-I. Sugishita, M. Harada, N. Fujii, and K. Miyajima
1997. Interactions of an antimicrobial peptide, magainin 2, with outer and inner membranes of gram-negative bacteria. *Biochimica et Biophysica Acta - Biomembranes*, 1327(1):119–130.

[14] Matsuzaki, K., K.-I. Sugishita, N. Ishibe, M. Ueha, S. Nakata, K. Miyajima, and R. Epand
1998. Relationship of membrane curvature to the formation of pores by magainin 2. *Biochemistry*, 37(34):11856–11863.

[15] Yang, L., T. Harroun, T. Weiss, L. Ding, and H. Huang
2001. Barrel-stave model or toroidal model? a case study on melittin pores. *Biophysical Journal*, 81(3):1475–1485.

[16] Henzler Wildman, K., D.-K. Lee, and A. Ramamoorthy
2003. Mechanism of lipid bilayer disruption by the human antimicrobial peptide, LL-37. *Biochemistry*, 42(21):6545–6558.

[17] Leontiadou, H., A. Mark, and S. Marrink
2006. Antimicrobial peptides in action. *Journal of the American Chemical Society*, 128(37):12156–12161.

[18] Yoneyama, F., Y. Imura, K. Ohno, T. Zendo, J. Nakayama, K. Matsuzaki, and K. Sonomoto
2009. Peptide-lipid huge toroidal pore, a new antimicrobial mechanism mediated by a lactococcal bacteriocin, lacticin Q. *Antimicrobial Agents and Chemotherapy*, 53(8):3211–3217.

[19] Shenkarev, Z., S. Balandin, K. Trunov, A. Paramonov, S. Sukhanov, L. Barsukov, A. Arseniev, and T. Ovchinnikova
2011. Molecular mechanism of action of β-hairpin antimicrobial peptide arenicin: Oligomeric structure in dodecylphosphocholine micelles and pore formation in planar lipid bilayers. *Biochemistry*, 50(28):6255–6265.

[20] Cruz, J., C. Ortiz, F. Guzmán, R. Fernández-Lafuente, and R. Torres
2014. Antimicrobial peptides: Promising compounds against pathogenic microorganisms. *Current Medicinal Chemistry*, 21(20):2299–2321.

[21] Axelsen, P.
2008. A chaotic pore model of polypeptide antibiotic action. *Biophysical Journal*, 94(5):1549–1550.

[22] Gregory, S., A. Cavenaugh, V. Journigan, A. Pokorny, and P. Almeida
2008. A quantitative model for the all-or-none permeabilization of phospholipid vesicles by the antimicrobial peptide cecropin A. *Biophysical Journal*, 94(5):1667–1680.

[23] Wu, M., E. Maier, R. Benz, and R. Hancock
1999. Mechanism of interaction of different classes of cationic antimicrobial peptides with planar bilayers and with the cytoplasmic membrane of escherichia coli. *Biochemistry*, 38(22):7235–7242.

[24] Shai, Y.
1999. Mechanism of the binding, insertion and destabilization of phospholipid bilayer membranes by α-helical antimicrobial and cell non-selective membrane-lytic peptides. *Biochimica et Biophysica Acta - Biomembranes*, 1462(1-2):55–70.

[25] Ladokhin, A. and S. White
2001. 'Detergent-like' permeabilization of anionic lipid vesicles by melittin. *Biochimica et Biophysica Acta - Biomembranes*, 1514(2):253–260.

[26] Shai, Y.
1995. Molecular recognition between membrane-spanning polypeptides. *Trends in Biochemical Sciences*, 20(11):460–464.

[27] Gazit, E., A. Boman, H. Boman, and Y. Shai
1995. Interaction of the mammalian antibacterial peptide cecropin P1 with phospholipid vesicles. *Biochemistry*, 34(36):11479–11488.

[28] Silvestro, L. and P. Axelsen
2000. Membrane-induced folding of cecropin A. *Biophysical Journal*, 79(3):1465–1477.

[29] Pokorny, A., T. Birkbeck, and P. Almeida
2002. Mechanism and kinetics of δ-lysin interaction with phospholipid vesicles. *Biochemistry*, 41(36):11044–11056.

[30] Pokorny, A. and P. Almeida
2004. Kinetics of dye efflux and lipid flip-flop induced by δ-lysin in phosphatidylcholine vesicles and the mechanism of graded release by amphipathic, α-helical peptides. *Biochemistry*, 43(27):8846–8857.

[31] Powers, J.-P. and R. Hancock
2003. The relationship between peptide structure and antibacterial activity. *Peptides*, 24(11):1681–1691.

[32] Mahlapuu, M., J. Håkansson, L. Ringstad, and C. Björn
2016. Antimicrobial peptides: An emerging category of therapeutic agents. *Frontiers in Cellular and Infection Microbiology*, 6(DEC).

[33] Bierbaum, G. and H.-G. Sahl
1987. Autolytic system of Staphylococcus simulans 22: Influence of cationic peptides on activity of N-acetylmuramoyl-L-alanine amidase. *Journal of Bacteriology*, 169(12):5452–5458.

[34] Zhao, H. and P. Kinnunen
2003. Modulation of the activity of secretory phospholipase A2 by antimicrobial peptides. *Antimicrobial Agents and Chemotherapy*, 47(3):965–971.

[35] Nicolas, P.
2009. Multifunctional host defense peptides: Intracellular-targeting antimicrobial peptides. *FEBS Journal*, 276(22):6483–6496.

[36] Hultmark, D., H. Steiner, T. Rasmuson, and H. Boman
1980. Insect immunity. purification and properties of three inducible bactericidal proteins from hemolymph of immunized pupae of Hyalophora cecropia. *European Journal of Biochemistry*, 106(1):7–16.

[37] Steiner, H., D. Hultmark, A. Engström, H. Bennich, and H. Boman
1981. Sequence and specificity of two antibacterial proteins involved in insect immunity. *Nature*, 292(5820):246–248.

[38] Lee, J.-Y., A. Boman, S. Chuanxin, M. Andersson, H. Jornvall, V. Mutt, and H. Boman
1989. Antibacterial peptides from pig intestine: Isolation of a mammalian cecropin. *Proceedings of the National Academy of Sciences of the United States of America*, 86(23):9159–9162.

[39] Bechinger, B. and K. Lohner
2006. Detergent-like actions of linear amphipathic cationic antimicrobial peptides. *Biochimica et Biophysica Acta - Biomembranes*, 1758(9):1529–1539.

[40] Hong, R., M. Shchepetov, J. Weiser, and P. Axelsen
2003. Transcriptional profile of the escherichia coli response to the antimicrobial insect peptide cecropin A. *Antimicrobial Agents and Chemotherapy*, 47(1):1–6.

[41] Kim, J.-K., E. Lee, S. Shin, K.-W. Jeong, J.-Y. Lee, S.-Y. Bae, S.-H. Kim, J. Lee, S. Kim, D. Lee, J.-S. Hwang, and Y. Kim
2011. Structure and function of papiliocin with antimicrobial and anti-inflammatory activities isolated from the swallowtail butterfly, Papilio xuthus. *Journal of Biological Chemistry*, 286(48):41296–41311.

[42] Wade, D., A. Boman, B. Wåhlin, C. Drain, D. Andreu, H. Boman, and R. Merrifield
1990. All-D amino acid-containing channel-forming antibiotic peptides. *Proceedings of the National Academy of Sciences of the United States of America*, 87(12):4761–4765.

[43] Merrifield, R., E. Merrifield, P. Juvvadi, D. Andreu, and H. Boman
1994. Design and synthesis of antimicrobial peptides. *Ciba Foundation symposium*, 186:5–20; discussion 20.

[44] Vunnam, S., P. Juvvadi, K. Rotondi, and R. Merrifield
1998. Synthesis and study of normal, enantio, retro, and retroenantio isomers of cecropin A-melittin hybrids, their end group effects and selective enzyme inactivation. *Journal of Peptide Research*, 51(1):38–44.

[45] Quesada, H., S. Ramos-Onsins, and M. Aguadé
2005. Birth-and-death evolution of the cecropin multigene family in drosophila. *Journal of Molecular Evolution*, 60(1):1–11.

[46] Cleveland, J., T. Montville, I. Nes, and M. Chikindas
2001. Bacteriocins: Safe, natural antimicrobials for food preservation. *International Journal of Food Microbiology*, 71(1):1–20.

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

[48] Papagianni, M. and S. Anastasiadou
2009. Pediocins: The bacteriocins of Pediococci. sources, production, properties and applications. *Microbial Cell Factories*, 8(1).

[49] Moll, G., W. Konings, and A. Driessen
1999. Bacteriocins: Mechanism of membrane insertion and pore formation. *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*, 76(1-4):185–198.

[50] Chikindas, M., M. Garcia-Garcera, A. Driessen, A. Ledeboer, J. Nissen-Meyer, I. Nes, T. Abee, W. Konings, and G. Venema
1993. Pediocin PA-1, a bacteriocin from Pediococcus acidilactici PAC1.0, forms hydrophilic pores in the cytoplasmic membrane of target cells. *Applied and Environmental Microbiology*, 59(11):3577–3584.

[51] Zanetti, M., R. Gennaro, and D. Romeo
1995. Cathelicidins: a novel protein family with a common proregion and a variable c-terminal antimicrobial domain. *FEBS Letters*, 374(1):1–5.

[52] Lehrer, R. and T. Ganz
2002. Cathelicidins: A family of endogenous antimicrobial peptides. *Current Opinion in Hematology*, 9(1):18–22.

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

[54] Gordon, Y., L. Huang, E. Romanowski, K. Yates, R. Proske, and A. McDermott
2005. Human cathelicidin (LL-37), a multifunctional peptide, is expressed by ocular surface epithelia and has potent antibacterial and antiviral activity. *Current Eye Research*, 30(5):385–394.

[55] Pütsep, K., G. Carlsson, H. Boman, and M. Andersson
2002. Deficiency of antibacterial peptides in patients with morbus Kostmann: An observation study. *Lancet*, 360(9340):1144–1149.

[56] Boman, H., B. Agerberth, and A. Boman
1993. Mechanisms of action on Escherichia coli of cecropin P1 and PR-39, two antibacterial peptides from pig intestine. *Infection and Immunity*, 61(7):2978–2984.

[57] Skerlavaj, B., M. Scocchi, R. Gennaro, A. Risso, and M. Zanetti
2001. Structural and functional analysis of horse cathelicidin peptides. *Antimicrobial Agents and Chemotherapy*, 45(3):715–722.

[58] Dorschner, R., V. Pestonjamasp, S. Tamakuwala, T. Ohtake, J. Rudisill, V. Nizet, B. Agerberth, G. Gudmundsson, and R. Gallo
2001. Cutaneous injury induces the release of cathelicidin anti-microbial peptides active against group a streptococcus. *Journal of Investigative Dermatology*, 117(1):91–97.

[59] Turner, J., Y. Cho, N.-N. Dinh, A. Waring, and R. Lehrer
1998. Activities of LL-37, a cathelin-associated antimicrobial peptide of human neutrophils. *Antimicrobial Agents and Chemotherapy*, 42(9):2206–2214.

[60] Wong, J., T. Ng, A. Legowska, K. Rolka, M. Hui, and C. Cho
2011. Antifungal action of human cathelicidin fragment (LL13-37) on Candida albicans. *Peptides*, 32(10):1996–2002.

[61] Gallo, R., M. Ono, T. Povsic, C. Page, E. Eriksson, M. Klagsbrun, and M. Bernfield
1994. Syndecans, cell surface heparan sulfate proteoglycans, are induced by a proline-rich antimicrobial peptide from wounds. *Proceedings of the National Academy of Sciences of the United States of America*, 91(23):11035–11039.

[62] Li, J., M. Post, R. Volk, Y. Gao, M. Li, C. Metais, K. Sato, J. Tsai, W. Aird, R. Rosenberg, T. Hampton, J. Li, F. Sellke, P. Carmeliet, and M. Simons
2000. Pr39, a peptide regulator of angiogenesis. *Nature Medicine*, 6(1):49–55.

[63] Niyonsaba, F., K. Iwabuchi, A. Someya, M. Hirata, H. Matsuda, H. Ogawa, and I. Nagaoka
2002. A cathelicidin family of human antibacterial peptide LL-37 induces mast cell chemotaxis. *Immunology*, 106(1):20–26.

[64] De Yang, B., Q. Chen, A. Schmidt, G. Anderson, J. Wang, J. Wooters, J. Oppenheim, and O. Chertov
2000. LL-37, the neutrophil granule- and epithelial cell-derived cathelicidin, utilizes formyl peptide receptor-like 1 (FPRL1) as a receptor to chemoattract human peripheral blood neutrophils, monocytes, and T cells. *Journal of Experimental Medicine*, 192(7):1069–1074.

[65] Scott, M., D. Davidson, M. Gold, D. Bowdish, and R. Hancock
2002. The human antimicrobial peptide LL-37 is a multifunctional modulator of innate immune responses. *Journal of Immunology*, 169(7):3883–3891.

[66] Bowdish, D., D. Davidson, Y. Lau, K. Lee, M. Scott, and R. Hancock
2005. Impact of LL-37 on anti-infective immunity. *Journal of Leukocyte Biology*, 77(4):451–459.

[67] Niyonsaba, F., A. Someya, M. Hirata, H. Ogawa, and I. Nagaoka
2001. Evaluation of the effects of peptide antibiotics human β-defensins-1/-2 and LL-37 on histamine release and prostaglandin D2 production from mast cells. *European Journal of Immunology*, 31(4):1066–1075.

[68] Guaní-Guerra, E., T. Santos-Mendoza, S. Lugo-Reyes, and L. Terán
2010. Antimicrobial peptides: General overview and clinical implications in human health and disease. *Clinical Immunology*, 135(1):1–11.

[69] Bateman, A., M. Martin, C. O'Donovan, M. Magrane, R. Apweiler, E. Alpi, R. Antunes, J. Arganiska, B. Bely, M. Bingley, C. Bonilla, R. Britto, B. Bursteinas, G. Chavali, E. Cibrian-Uhalte, A. Da Silva, M. De Giorgi, T. Dogan, F. Fazzini, P. Gane, L. Castro, P. Garmiri, E. Hatton-Ellis, R. Hieta, R. Huntley, D. Legge, W. Liu, J. Luo, A. Macdougall, P. Mutowo, A. Nightingale, S. Orchard, K. Pichler, D. Poggioli, S. Pundir, L. Pureza, G. Qi, S. Rosanoff, R. Saidi, T. Sawford, A. Shypitsyna, E. Turner, V. Volynkin, T. Wardell, X. Watkins, H. Zellner, A. Cowley, L. Figueira, W. Li, H. McWilliam, R. Lopez, I. Xenarios, L. Bougueleret, A. Bridge, S. Poux, N. Redaschi, L. Aimo, G. Argoud-Puy, A. Auchincloss, K. Axelsen, P. Bansal, D. Baratin, M.-C. Blatter, B. Boeckmann, J. Bolleman, E. Boutet, L. Breuza, C. Casal-Casas, E. De Castro, E. Coudert, B. Cuche, M. Doche, D. Dornevil, S. Duvaud, A. Estreicher, L. Famiglietti, M. Feuermann, E. Gasteiger, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz-Gumowski, U. Hinz, C. Hulo, F. Jungo, G. Keller, V. Lara, P. Lemercier, D. Lieberherr, T. Lombardot, X. Martin, P. Masson, A. Morgat, T. Neto, N. Nouspikel, S. Paesano, I. Pedruzzi, S. Pilbout, M. Pozzato, M. Pruess, C. Rivoire, B. Roechert, M. Schneider, C. Sigrist, K. Sonesson, S. Staehli, A. Stutz, S. Sundaram, M. Tognolli, L. Verbregue, A. Veuthey, C. Wu, C. Arighi, L. Arminski, C. Chen, Y. Chen, J. Garavelli, H. Huang, K. Laiho, P. McGarvey, D. Natale, B. Suzek, C. Vinayaka, Q. Wang, Y. Wang, L.-S. Yeh, M. Yerramalla, and J. Zhang
2015. UniProt: A hub for protein information. *Nucleic Acids Research*, 43(D1):D204–D212.

[70] Rose, P., A. Prlić, A. Altunkaya, C. Bi, A. Bradley, C. Christie, L. Di Costanzo, J. Duarte, S. Dutta, Z. Feng, R. Green, D. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. Rose, C. Shao, Y.-P. Tao, Y. Valasatava, M. Voigt, J. Westbrook, J. Woo, H. Yang, J. Young, C. Zardecki, H. Berman, and S. Burley
2017. The RCSB protein data bank: Integrative view of protein, gene and 3D structural information. *Nucleic Acids Research*, 45(D1):D271–D281.

[71] Torrent, M., M. Nogués, and E. Boix
2012. Discovering new in silico tools for antimicrobial peptide prediction. *Current Drug Targets*, 13(9):1148–1157.

[72] Aguilera-Mendoza, L., Y. Marrero-Ponce, R. Tellez-Ibarra, M. Llorente-Quesada, J. Salgado, S. Barigye, and J. Liu
2015. Overlap and diversity in antimicrobial peptide databases: Compiling a non-redundant set of sequences. *Bioinformatics*, 31(15):2553–2559.

[73] Wang, G., X. Li, and Z. Wang
2016. Apd3: The antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research*, 44(D1):D1087–D1093.

[74] Qureshi, A., N. Thakur, H. Tandon, and M. Kumar
2014. AVPdb: A database of experimentally validated antiviral peptides targeting medically important viruses. *Nucleic Acids Research*, 42(D1):D1147–D1153.

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

[76] Hammami, R., A. Zouhir, C. Le Lay, J. Ben Hamida, and I. Fliss
2010. BACTIBASE second release: A database and tool platform for bacteriocin characterization. *BMC Microbiology*, 10.

[77] de Jong, A., A. van Heel, J. Kok, and O. Kuipers
2010. BAGEL2: Mining for bacteriocins in genomic data. *Nucleic Acids Research*, 38(SUPPL. 2):W647–W651.

[78] Waghu, F., R. Barai, P. Gurung, and S. Idicula-Thomas
2016. $CAMP_{R3}$: A database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Research*, 44(D1):D1094–D1097.

[79] Wang, C., Q. Kaas, L. Chiche, and D. Craik
2008. CyBase: A database of cyclic protein sequences and structures, with applications in protein discovery and engineering. *Nucleic Acids Research*, 36(SUPPL. 1):D206–D210.

[80] Novković, M., J. Simunić, V. Bojović, A. Tossi, and D. Juretić
2012. DADP: The database of anuran defense peptides. *Bioinformatics*, 28(10):1406–1407.

[81] Brahmachary, M., S. Krishnan, J. Koh, A. Khan, S. Seah, T. Tan, V. Brusic, and V. Bajic
2004. ANTIMIC: A database of antimicrobial sequences. *Nucleic Acids Research*, 32(DATABASE ISS.):D586–D589.

[82] Sundararajan, V., M. Gabere, A. Pretorius, S. Adam, A. Christoffels, M. Lehvašlaiho, J. Archer, and V. Bajic
2012. DAMPD: A manually curated antimicrobial peptide database. *Nucleic Acids Research*, 40(D1):D1108–D1112.

[83] Pirtskhalava, M., A. Gabrielian, P. Cruz, H. Griggs, R. Squires, D. Hurt, M. Grigolava, M. Chubinidze, G. Gogoladze, B. Vishnepolsky, V. Alekseev, A. Rosenthal, and M. Tartakovsky
2016. DBAASP v.2: An enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Research*, 44(D1):D1104–D1112.

[84] Seebah, S., A. Suresh, S. Zhuo, Y. Choong, H. Chua, D. Chuon, R. Beuerman, and C. Verma
2007. Defensins knowledgebase: A manually curated database and information source focused on the defensins family of antimicrobial peptides. *Nucleic Acids Research*, 35(SUPPL. 1):D265–D268.

[85] Fan, L., J. Sun, M. Zhou, J. Zhou, X. Lao, H. Zheng, and H. Xu
2016. DRAMP: A comprehensive data repository of antimicrobial peptides. *Scientific Reports*, 6.

[86] Gautam, A., K. Chaudhary, S. Singh, A. Joshi, P. Anand, A. Tuknait, D. Mathur, G. Varshney, and G. Raghava
2014. Hemolytik: A database of experimentally determined hemolytic and non-hemolytic peptides. *Nucleic Acids Research*, 42(D1):D444–D449.

[87] Qureshi, A., N. Thakur, and M. Kumar
2013. HIPdb: A database of experimentally validated HIV inhibiting peptides. *PLoS ONE*, 8(1).

[88] Zhao, X., H. Wu, H. Lu, G. Li, and Q. Huang
2013. LAMP: A database linking antimicrobial peptides. *PLoS ONE*, 8(6).

[89] Théolier, J., I. Fliss, J. Jean, and R. Hammami
2014. MilkAMP: A comprehensive database of antimicrobial peptides of dairy origin. *Dairy Science and Technology*, 94(2):181–193.

[90] Whitmore, L. and B. Wallace
2004. The peptaibol database: A database for sequences and structures of naturally occurring peptaibols. *Nucleic Acids Research*, 32(DATABASE ISS.):D593–D594.

[91] Hammami, R., J. Ben Hamida, G. Vergoten, and I. Fliss
2009. PhytAMP: A database dedicated to antimicrobial plant peptides. *Nucleic Acids Research*, 37(SUPPL. 1):D963–D968.

[92] Singh, S., K. Chaudhary, S. Dhanda, S. Bhalla, S. Usmani, A. Gautam, A. Tuknait, P. Agrawal, D. Mathur, and G. Raghava
2015. SATPdb: A database of structurally annotated therapeutic peptides. *Nucleic Acids Research*, 44(D1):D1119–D1126.

[93] Li, J., X. Qu, X. He, L. Duan, G. Wu, D. Bi, Z. Deng, W. Liu, and H.-Y. Ou
2012. ThioFinder: A web-based tool for the identification of thiopeptide gene clusters in DNA sequences. *PLoS ONE*, 7(9).

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

[95] Wong-ekkabut, J. and M. Karttunen
2016. The good, the bad and the user in soft matter simulations. *Biochimica et Biophysica Acta - Biomembranes*, 1858(10):2529–2538.

[96] MacKerell Jr., A., D. Bashford, M. Bellott, R. Dunbrack Jr., J. Evanseck, M. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. Lau, C. Mattos, S. Michnick, T. Ngo, D. Nguyen, B. Prodhom, W. Reiher III, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus
1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *Journal of Physical Chemistry B*, 102(18):3586–3616.

[97] Case, D., T. Cheatham III, T. Darden, H. Gohlke, R. Luo, K. Merz Jr., A. Onufriev, C. Simmerling, B. Wang, and R. Woods
2005. The amber biomolecular simulation programs. *Journal of Computational Chemistry*, 26(16):1668–1688.

[98] Robertson, M., J. Tirado-Rives, and W. Jorgensen
2015. Improved peptide and protein torsional energetics with the OPLS-AA force field. *Journal of Chemical Theory and Computation*, 11(7):3499–3509.

[99] Best, R., X. Zhu, J. Shim, P. Lopes, J. Mittal, M. Feig, and A. MacKerell Jr.
2012. Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. *Journal of Chemical Theory and Computation*, 8(9):3257–3273.

[100] MacKerell Jr., A., J. Wiórkiewicz-Kuczera, and M. Karplus
1995. An all-atom empirical energy function for the simulation of nucleic acids. *Journal of the American Chemical Society*, 117(48):11946–11975.

[101] Foloppe, N. and A. MacKerell Jr.
2000. All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *Journal of Computational Chemistry*, 21(2):86–104.

[102] MacKerell Jr., A. and N. Banavali
2000. All-atom empirical force field for nucleic acids: II. application to molecular dynamics simulations of dna and rna in solution. *Journal of Computational Chemistry*, 21(2):105–120.

[103] Feller, S. and A. MacKerell Jr.
2000. An improved empirical potential energy function for molecular simulations of phospholipids. *Journal of Physical Chemistry B*, 104(31):7510–7515.

[104] Feller, S., K. Gawrisch, and A. MacKerell Jr.
2002. Polyunsaturated fatty acids in lipid bilayers: Intrinsic and environmental contributions to their unique physical properties. *Journal of the American Chemical Society*, 124(2):318–326.

[105] Klauda, J., R. Venable, J. Freites, J. O'Connor, D. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A. MacKerell Jr., and R. Pastor
2010. Update of the charmm all-atom additive force field for lipids: Validation on six lipid types. *Journal of Physical Chemistry B*, 114(23):7830–7843.

[106] Kuttel, M., J. Brady, and K. Naidoo
2002. Carbohydrate solution simulations: Producing a force field with experimentally consistent primary alcohol rotational frequencies and populations. *Journal of Computational Chemistry*, 23(13):1236–1243.

[107] Guvench, O., S. Greenr, G. Kamath, J. Brady, R. Venable, R. Pastor, and A. Mackerell Jr.
2008. Additive empirical force field for hexopyranose monosaccharides. *Journal of Computational Chemistry*, 29(15):2543–2564.

[108] Hatcher, E., O. Guvench, and A. MacKerell Jr.
2009. CHARMM additive all-atom force field for acyclic polyalcohols, acyclic carbohydrates, and inositol. *Journal of Chemical Theory and Computation*, 5(5):1315–1327.

[109] Guvench, O., E. Hatcher, R. Venable, R. Pastor, and A. MacKerell Jr.
2009. CHARMM additive all-atom force field for glycosidic linkages between hexopyranoses. *Journal of Chemical Theory and Computation*, 5(9):2353–2370.

[110] Vanommeslaeghe, K., E. Hatcher, C. Acharya, S. Kundu, S. Zhong, J. Shim, E. Darian, O. Guvench, P. Lopes, I. Vorobyov, and A. Mackerell Jr.
2010. CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, 31(4):671–690.

[111] Freddolino, P. and K. Schulten
2009. Common structural transitions in explicit-solvent simulations of villin headpiece folding. *Biophysical Journal*, 97(8):2338–2347.

[112] Brooks, B., C. Brooks III, A. Mackerell Jr., L. Nilsson, R. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. Pastor, C. Post, J. Pu, M. Schaefer, B. Tidor, R. Venable, H. Woodcock, X. Wu, W. Yang, D. York, and M. Karplus
2009. CHARMM: The biomolecular simulation program. *Journal of Computational Chemistry*, 30(10):1545–1614.

[113] Huang, J., S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. De Groot, H. Grubmüller, and A. MacKerell
2016. CHARMM36m: An improved force field for folded and intrinsically disordered proteins. *Nature Methods*, 14(1):71–73.

[114] Liwo, A., S. Ołdziej, M. Pincus, R. Wawak, S. Rackovsky, and H. Scheraga
1997. A united-residue force field for off-lattice protein-structure simulations. I. functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *Journal of Computational Chemistry*, 18(7):849–873.

[115] Maupetit, J., P. Tuffery, and P. Derreumaux
2007. A coarse-grained protein force field for folding and structure prediction. *Proteins: Structure, Function and Genetics*, 69(2):394–408.

[116] Bereau, T. and M. Deserno
2009. Generic coarse-grained model for protein folding and aggregation. *Journal of Chemical Physics*, 130(23).

[117] Wan, C.-K., W. Han, and Y.-D. Wu
2012. Parameterization of PACE force field for membrane environment and simulation of helical peptides and helix-helix association. *Journal of Chemical Theory and Computation*, 8(1):300–313.

[118] Marrink, S., H. Risselada, S. Yefimov, D. Tieleman, and A. De Vries
2007. The MARTINI force field: Coarse grained model for biomolecular simulations. *Journal of Physical Chemistry B*, 111(27):7812–7824.

[119] Monticelli, L., S. Kandasamy, X. Periole, R. Larson, D. Tieleman, and S.-J. Marrink
2008. The MARTINI coarse-grained force field: Extension to proteins. *Journal of Chemical Theory and Computation*, 4(5):819–834.

[120] De Jong, D., G. Singh, W. Bennett, C. Arnarez, T. Wassenaar, L. Schäfer, X. Periole, D. Tieleman, and S. Marrink
2013. Improved parameters for the martini coarse-grained protein force field. *Journal of Chemical Theory and Computation*, 9(1):687–697.

[121] Barnoud, J. and L. Monticelli
2014. Coarse-grained force fields for molecular simulations. *Methods in Molecular Biology*, 1215:125–149.

[122] López, C., A. Rzepiela, A. de Vries, L. Dijkhuizen, P. Hünenberger, and S. Marrink
2009. Martini coarse-grained force field: Extension to carbohydrates. *Journal of Chemical Theory and Computation*, 5(12):3195–3210.

[123] Lee, H., A. De Vries, S.-J. Marrink, and R. Pastor
2009. A coarse-grained model for polyethylene oxide and polyethylene glycol: Conformation and hydrodynamics. *Journal of Physical Chemistry B*, 113(40):13186–13194.

[124] Rossi, G., L. Monticelli, S. Puisto, I. Vattulainen, and T. Ala-Nissila
2011. Coarse-graining polymers with the MARTINI force-field: Polystyrene as a benchmark case. *Soft Matter*, 7(2):698–708.

[125] Milani, A., M. Casalegno, C. Castiglioni, and G. Raos
2011. Coarse-grained simulations of model polymer nanofibres. *Macromolecular Theory and Simulations*, 20(5):305–319.

[126] Uusitalo, J., H. Ingólfsson, P. Akhshi, D. Tieleman, and S. Marrink
2015. Martini coarse-grained force field: Extension to DNA. *Journal of Chemical Theory and Computation*, 11(8):3932–3945.

[127] Vögele, M., C. Holm, and J. Smiatek
2015. Coarse-grained simulations of polyelectrolyte complexes: MARTINI models for poly(styrene sulfonate) and poly(diallyldimethylammonium). *Journal of Chemical Physics*, 143(24).

[128] Marrink, S. and D. Tieleman
2013. Perspective on the martini model. *Chemical Society Reviews*, 42(16):6801–6822.

[129] Yesylevskyy, S., L. Schäfer, D. Sengupta, and S. Marrink
2010. Polarizable water model for the coarse-grained MARTINI force field. *PLoS computational biology*, 6(6).

[130] Wu, Z., Q. Cui, and A. Yethiraj
2010. A new coarse-grained model for water: The importance of electrostatic interactions. *Journal of Physical Chemistry B*, 114(32):10524–10529.

[131] Michalowsky, J., L. Schäfer, C. Holm, and J. Smiatek
2017. A refined polarizable water model for the coarse-grained MARTINI force field with long-range electrostatic interactions. *Journal of Chemical Physics*, 146(5).

[132] De Jong, D., S. Baoukina, H. Ingólfsson, and S. Marrink
2016. Martini straight: Boosting performance using a shorter cutoff and GPUs. *Computer Physics Communications*, 199:1–7.

[133] Wassenaar, T., K. Pluhackova, R. Böckmann, S. Marrink, and D. Tieleman
2014. Going backward: A flexible geometric approach to reverse transformation from coarse grained to atomistic models. *Journal of Chemical Theory and Computation*, 10(2):676–690.

[134] Cock, P., T. Antao, J. Chang, B. Chapman, C. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. De Hoon
2009. Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

[135] Notredame, C., D. Higgins, and J. Heringa
2000. T-coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology*, 302(1):205–217.

[136]
2016. *T-COFFEE User Manual*.

[137] Capella-Gutiérrez, S., J. Silla-Martínez, and T. Gabaldón
2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973.

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

[139] Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. Madden
2009. BLAST+: Architecture and applications. *BMC Bioinformatics*, 10.

[140] Altschul, S., T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman
1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402.

[141] Bailey, T., M. Boden, F. Buske, M. Frith, C. Grant, L. Clementi, J. Ren, W. Li, and W. Noble
2009. MEME suite: Tools for motif discovery and searching. *Nucleic Acids Research*, 37(SUPPL. 2):W202–W208.

[142] Bailey, T. and C. Elkan
1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / . International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36.

[143] Qi, Y., H. Ingólfsson, X. Cheng, J. Lee, S. Marrink, and W. Im
2015. CHARMM-GUI martini maker for coarse-grained simulations with the martini force field. *Journal of Chemical Theory and Computation*, 11(9):4486–4494.

[144] Jo, S., T. Kim, V. Iyer, and W. Im
2008. CHARMM-GUI: A web-based graphical user interface for charmm. *Journal of Computational Chemistry*, 29(11):1859–1865.

[145] Dowhan, W.
1997. Molecular basis for membrane phospholipid diversity: Why are there so many lipids? *Annual Review of Biochemistry*, 66:199–232.

[146] Picas, L., C. Suárez-Germà, M. Montero, O. Domènech, and J. Hernández-Borrell
2012. Miscibility behavior and nanostructure of monolayers of the main phospholipids of Escherichia coli inner membrane. *Langmuir*, 28(1):701–706.

[147] Biasini, M., S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. Cassarino, M. Bertoni, L. Bordoli, and T. Schwede
2014. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(W1):W252–W258.

[148] Abraham, M., T. Murtola, R. Schulz, S. Páll, J. Smith, B. Hess, and E. Lindah
2015. Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, 1-2:19–25.

[149] Humphrey, W., A. Dalke, and K. Schulten
1996. VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, 14(1):33–38.

[150] R Core Team
2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[151] Wickham, H.
2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[152] Wang, G.
2008. Structures of human host defense cathelicidin LL-37 and its smallest antimicrobial peptide KR-12 in lipid micelles. *Journal of Biological Chemistry*, 283(47):32637–32643.

[153] Venable, R., F. Brown, and R. Pastor
2015. Mechanical properties of lipid bilayers from molecular dynamics simulation. *Chemistry and Physics of Lipids*, 192:60–74.

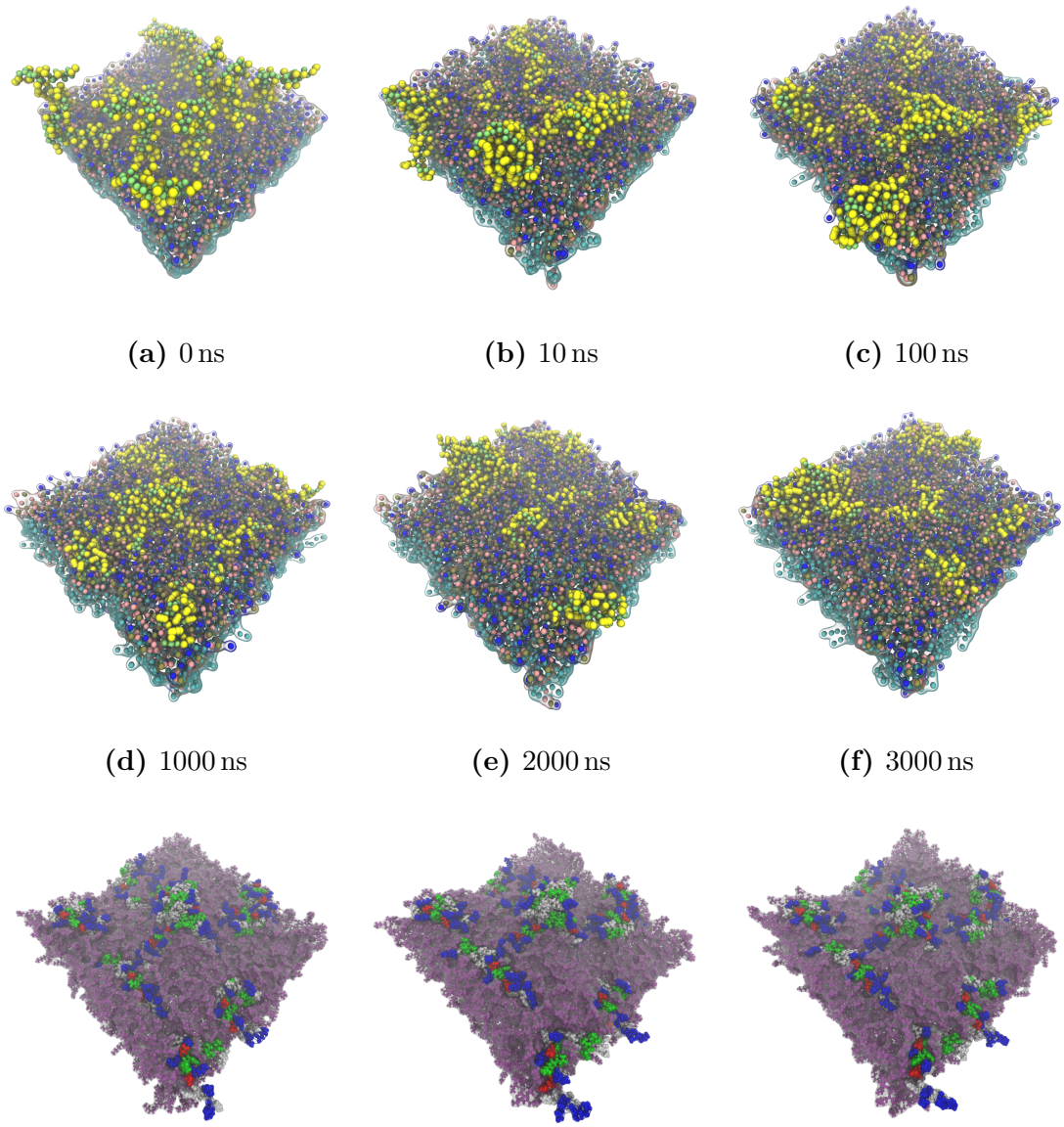
[154] Rappolt, M., A. Hickel, F. Bringezu, and K. Lohner
2003. Mechanism of the lamellar/inverse hexagonal phase transition examined by high resolution x-ray diffraction. *Biophysical Journal*, 84(5):3111–3122.

[155] Pan, J., D. Marquardt, F. Heberle, N. Kučerka, and J. Katsaras
2014. Revisiting the bilayer structures of fluid phase phosphatidylglycerol lipids: Accounting for exchangeable hydrogens. *Biochimica et Biophysica Acta - Biomembranes*, 1838(11):2966–2969.

[156] Kučerka, N., B. Holland, C. Gray, B. Tomberli, and J. Katsaras
2012. Scattering density profile model of popg bilayers as determined by molecular dynamics simulations and small-angle neutron and x-ray scattering experiments. *Journal of Physical Chemistry B*, 116(1):232–239.

Appendix A

Simulation Snapshots

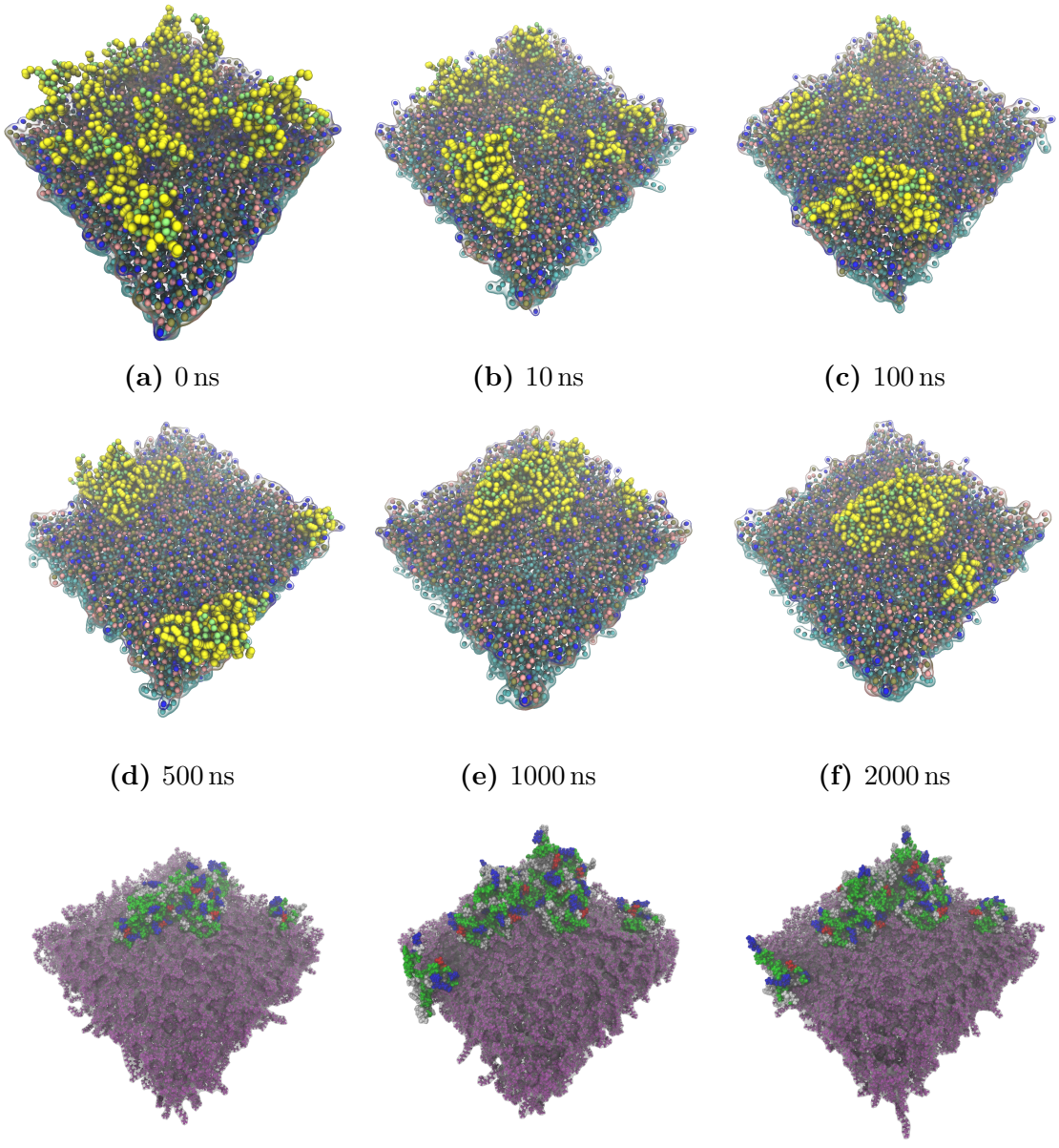


(a) 0 ns **(b)** 10 ns **(c)** 100 ns

(d) 1000 ns **(e)** 2000 ns **(f)** 3000 ns

(g) 1 ns **(h)** 10 ns **(i)** 20 ns

Figure A.1: Snapshots of the Cecropin A simulation

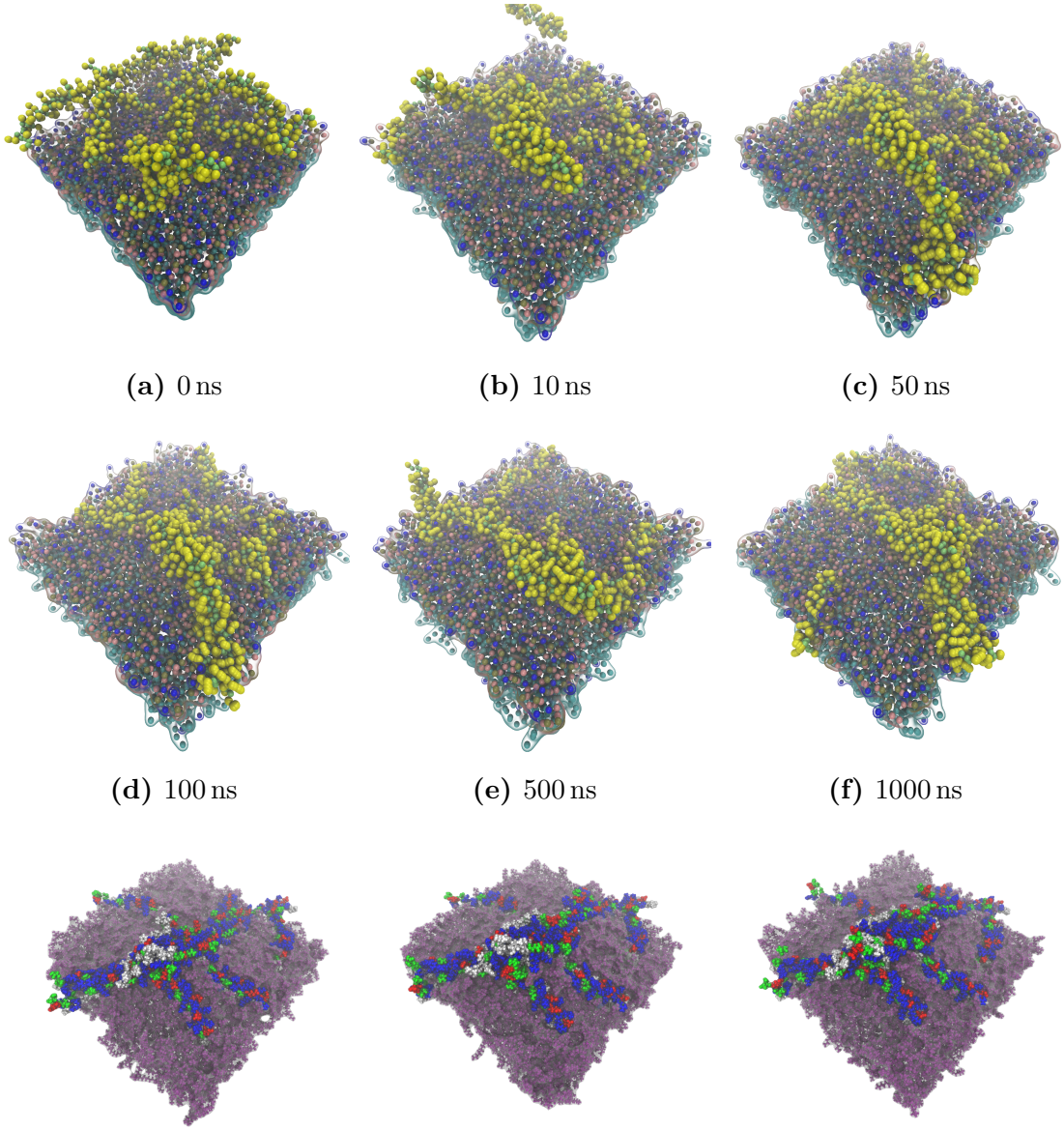


(a) 0 ns **(b)** 10 ns **(c)** 100 ns

(d) 500 ns **(e)** 1000 ns **(f)** 2000 ns

(g) 1 ns **(h)** 10 ns **(i)** 20 ns

Figure A.2: Snapshots of the bacteriocin Curvacin A simulation



(a) 0 ns **(b)** 10 ns **(c)** 50 ns

(d) 100 ns **(e)** 500 ns **(f)** 1000 ns

(g) 1 ns **(h)** 10 ns **(i)** 20 ns

Figure A.3: Snapshots of the LL-37 simulation

Appendix B

Additional Figures

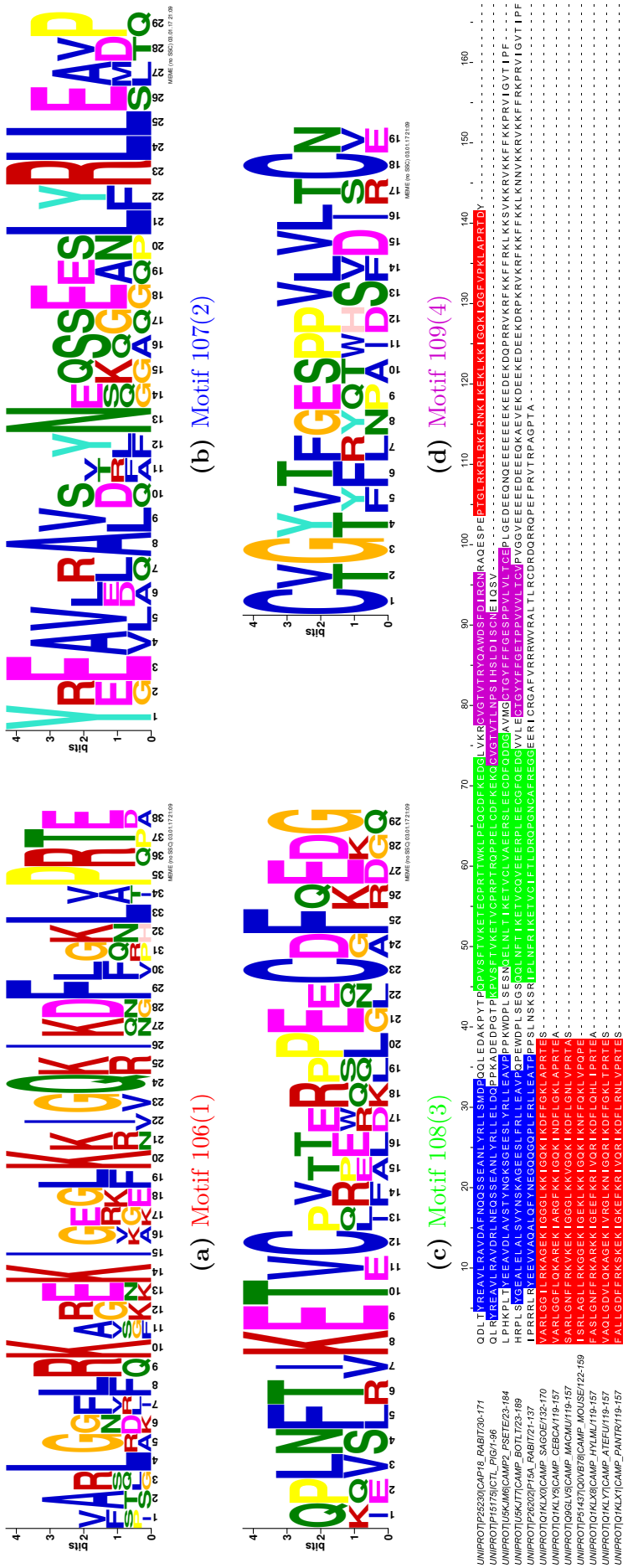


Figure B.1: Antibiotic-65 Motifs

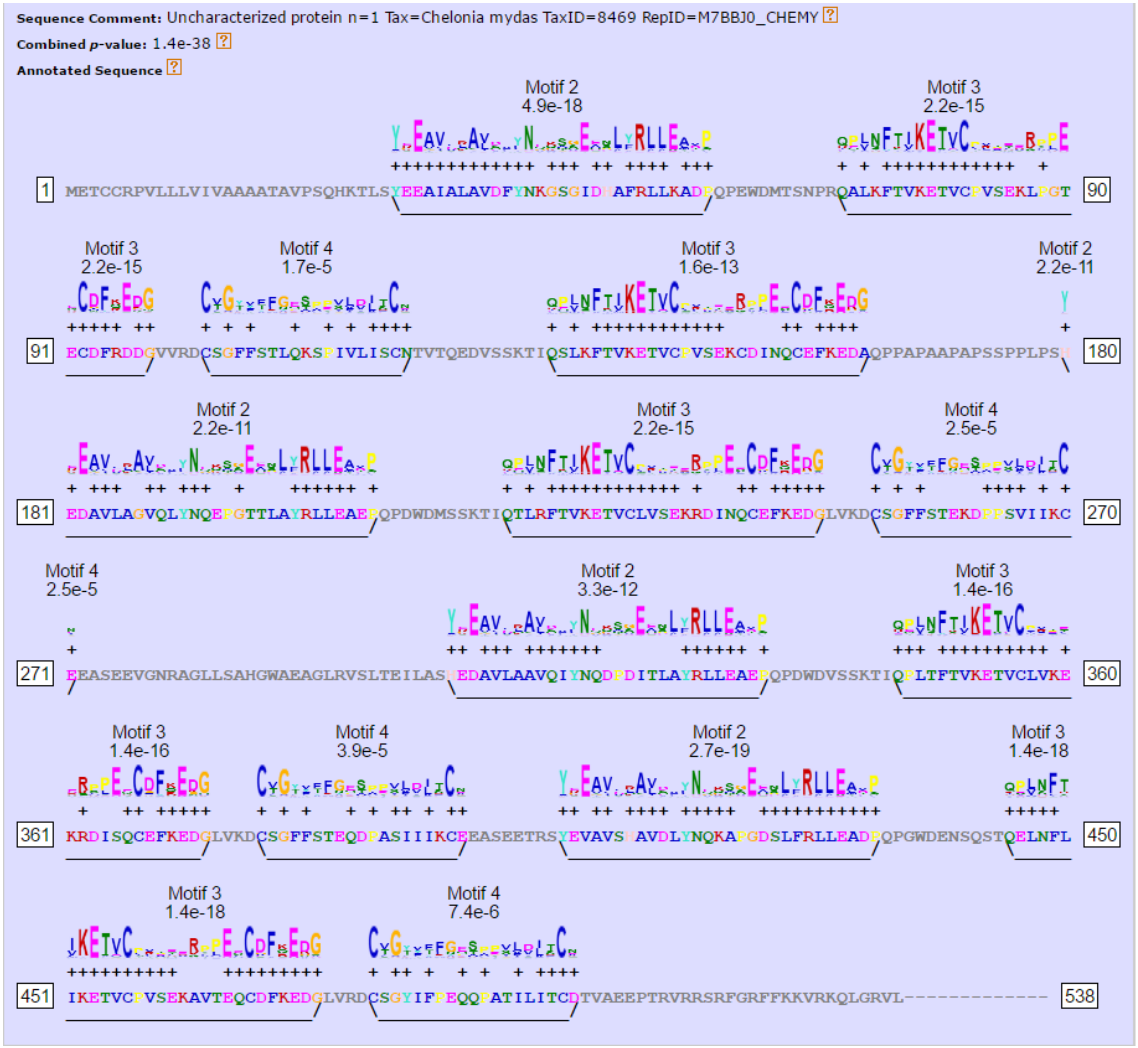


Figure B.2: Antibiotic-65 motifs in UniRef90_M7BBJ0

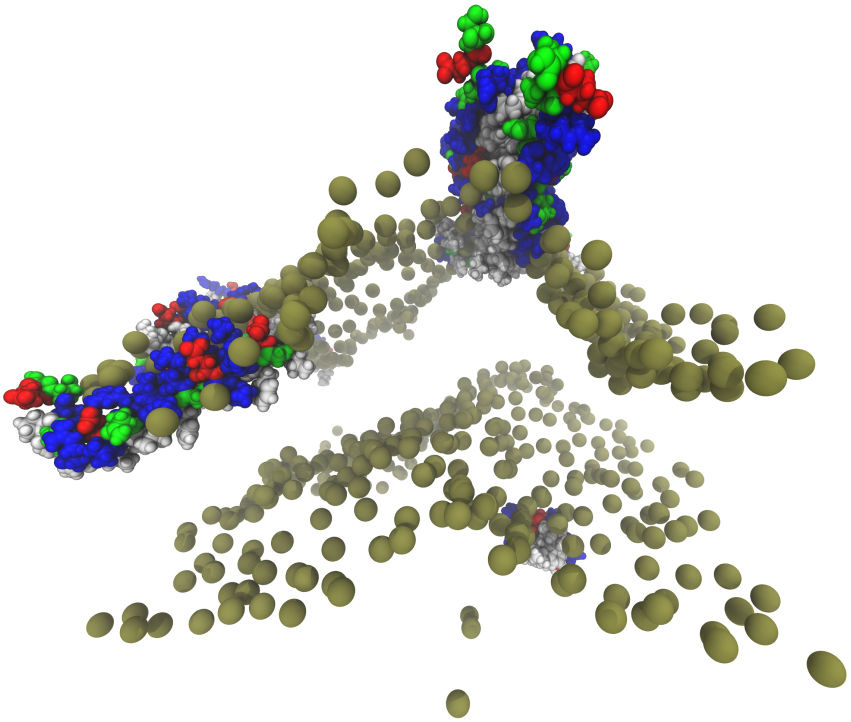


Figure B.3: LL-37 Short Z axis

Appendix C

Additional Tables

Table C.1: Antibiotic-84 Basic Data

UniProt AC	(-) Charges	(+) Charges	Overall Charge	Residues	Molecular Weight	Average residue weight	Isoelectric point
P01507	2	8	6	37	4004.82	108.238	11.1847
Q27239	2	9	7	35	3861.68	110.334	11.8225
P50723	2	9	7	35	3843.62	109.818	11.7770
P01509	2	9	7	35	3817.67	109.076	11.4938
P50720	2	8	6	35	3875.67	110.733	11.7770
P04142	3	9	6	35	3894.73	111.278	11.3945
P14663	3	5	2	37	3848.38	104.010	11.3051
P83420	2	8	6	32	3434.09	107.315	11.7770
P48821	3	5	2	37	3882.44	104.931	11.3051
P85210	6	6	0	39	4255.84	109.124	6.7214
P01510	3	5	2	36	3793.38	105.372	10.4944
Q5MGD8	5	8	3	40	4549.22	113.730	11.0056
O76146	4	6	2	36	3817.42	106.039	10.3258
Q0Q027	8	18	10	144	16382.25	113.766	9.7616
P83403	3	8	5	39	4205.88	107.843	11.3101
Q8MUF4	1	10	9	34	3854.70	113.373	12.1620
P67792	2	8	6	40	4297.98	107.449	12.1021
Q06590	2	8	6	39	4200.90	107.715	11.3363
Average	3	8	5	43	4656.70	108.897	11.0587

Table C.2: Bacteriocin-5 Basic Data

UniProt AC	(-) Charges	(+) Charges	Overall Charge	Residues	Molecular Weight	Average residue weight	Isoelectric point
P0A311	1	4	3	41	4308.89	105.095	9.3672
P34034	1	4	3	37	3932.32	106.279	8.7725
P38580	1	5	4	48	4969.5	103.531	9.9639
Q0Z8B6	3	7	4	44	5093.78	115.768	9.2523
P80925	1	5	4	43	4289.81	99.763	9.8139
P81053	1	4	3	43	4598.04	106.931	8.7677
P86394	0	4	4	35	3629.08	103.688	9.4978
P29430	1	6	5	44	4628.19	105.186	8.6675
P86291	3	4	1	29	3190.55	110.019	7.2502
P84962	2	5	3	43	4525.21	105.237	8.6471
O30434	2	4	2	44	4630.17	105.231	8.223
A9Q0M7	1	5	4	49	5274.85	107.65	9.3466
B3A0N4	1	5	4	43	4448.93	103.463	9.3268
Average	1	5	3	42	4424.56	105.988	8.9920

Table C.3: Antibiotic-39 Basic Data

UniProt AC	(-) Charges	(+) Charges	Overall Charge	Residues	Molecular Weight	Average residue weight	Isoelectric point
P49913	5	11	6	37	4493.32	121.441	11.3469
Q1KLX8	4	11	7	37	4471.28	120.845	11.6503
Q9GLV5	2	10	8	37	4100.91	110.835	11.8609
Q1KLX0	3	10	7	37	3968.75	107.263	11.4624
Q1KLY5	3	10	7	37	4156.94	112.350	11.7508
Q1KLY7	4	8	4	37	4141.83	111.941	10.9482
P51437	3	10	7	38	4291.20	112.926	11.2476
Average	3	10	7	37	4232.03	113.943	11.4667

	Estimate	Std. Error	t value	Pr(>|t|)
Membrane	3.941768	0.001645	2396.88	<2e-16 ***
Cecropin A	-0.125288	0.001899	-65.97	<2e-16 ***
Curvacin A	-0.064191	0.002014	-31.87	<2e-16 ***
LL-37 long	-0.188817	0.002326	-81.19	<2e-16 ***
LL-37 short	-0.110827	0.002014	-55.02	<2e-16 ***

Table C.4: Summary table of linear model Thickness vs. System for CG systems

	Estimate	Std. Error	t value	Pr(>|t|)
Membrane	4.19766	0.04028	104.203	< 2e-16 ***
Cecropin A	-0.35073	0.04934	-7.109	1.04e-09 ***
Curvacin A	0.01389	0.04934	0.282	0.779
LL-37	-0.30254	0.04934	-6.132	5.46e-08 ***

Table C.5: Summary table of linear model Thickness vs. System for all-atom systems

	Estimate	Std. Error	t value	Pr(>|t|)
Membrane	0.6934589	0.0002207	3142.35	<2e-16 ***
Cecropin A	0.0356591	0.0002548	139.93	<2e-16 ***
Curvacin A	0.0187521	0.0002703	69.38	<2e-16 ***
LL-37	0.0401760	0.0003121	128.73	<2e-16 ***

Table C.6: Summary table of linear model Area per Lipid vs. System for CG systems

	Estimate	Std. Error	t value	Pr(>|t|)
Membrane	0.6010036	0.0002300	2612.76	<2e-16 ***
Cecropin A	0.0425905	0.0002817	151.18	<2e-16 ***
Curvacin A	-0.0061627	0.0002817	-21.88	<2e-16 ***
LL-37	0.0535994	0.0002817	190.25	<2e-16 ***

Table C.7: Summary table of linear model Area per Lipid vs. System for all-atom systems