In Silico Design of Antimicrobial Peptides

ARTICLE *in* METHODS IN MOLECULAR BIOLOGY (CLIFTON, N.J.) · JANUARY 2015

Impact Factor: 1.29 · DOI: 10.1007/978-1-4939-2285-7_9 · Source: PubMed

3 AUTHORS:

Giuseppe Maccari
Istituto Italiano di Tecnologia
16 PUBLICATIONS **79** CITATIONS
SEE PROFILE

Mariagrazia Di Luca
Scuola Normale Superiore di Pisa
17 PUBLICATIONS **283** CITATIONS
SEE PROFILE

Riccardo Nifosì
Italian National Research Council
50 PUBLICATIONS **963** CITATIONS
SEE PROFILE

In Silico Design of Antimicrobial Peptides

Giuseppe Maccari*[1], Mariagrazia Di Luca[2], Riccardo Nifosì[2]

[1]Center for Nanotechnology Innovation @NEST, Istituto Italiano di Tecnologia, Pisa, Italy

[2]NEST, Istituto Nanoscienze-CNR and Scuola Normale Superiore, Pisa, Italy

SUMMARY

The rapid spread of drug-resistant pathogenic microbial strains has created an urgent need for the development of new anti-infective molecules, having different mechanism of action in comparison to existing drugs. Natural antimicrobial peptides (AMPs) represent a novel class of molecules with a broad spectrum of activity and a low rate in inducing bacterial resistance. In particular, linear alpha-helical cationic antimicrobial peptides are among the most widespread membrane-disruptive AMPs in nature, representing a particularly successful structural arrangement of the innate defense against microbes. However, until now, many AMPs have failed in clinical trials because of several drawbacks that strongly limit their applicability such as degradation, cytotoxicity and high production cost. Thus, to overcome the limitations of native peptides, a rational *in-silico* approach to AMPs design becomes a promising strategy that drastically reduce production costs and the time required for evaluation of activity and toxicity.

This chapter will focus on the strategies and methods for *de-novo* design of potentially active AMPs. In particular, statistical-based design strategies and MD methods for modelling AMPs will be elucidated.

KEYWORDS

AMPs; Drug-resistance; QSAR; Molecular Dynamics; *de-novo* peptide design

1 - INTRODUCTION

The appearance and rapid spread of antibiotic-resistant bacteria represents a major global health problem. Infections caused by resistant microorganisms often fail to respond to conventional treatment, resulting in prolonged illness, greater risk of death and higher costs. The decline in effectiveness of current therapies spurs research for the identification of novel molecules endowed with antimicrobial activities and new mechanisms of action.

Antimicrobial peptides (AMPs) are small evolutionally conserved molecules, representing an exciting class of drug candidates, particularly because their mechanism of action is unlikely to induce drug resistance and some of them are also active against microbial biofilms [1]. Furthermore, AMP have been applied not only as direct antimicrobial agents, but also as potential endosomolytic moieties promoting the release of biomolecules into cells for delivery purposes [2]. Although some AMPs are already in clinical and commercial use, the future design of novel AMPs will need to minimize the toxicity against eukaryotic cells and enhance the resistance to proteolytic degradation, with a key opportunity being offered by the introduction of non-natural amino acids (AA) to contrast host resistance and increase compound's life.

AMPs belong to a vast and various class of molecules, featuring different structure, amino acid composition and chemophysical characteristics. Therefore, an understanding of AMPs physicochemical characteristics and modes of action is mandatory in order to develop proper design and optimization strategies. Despite their great variability, most AMPs act by perturbing the cytoplasmic membrane, thus determining cell death by osmotic shock. Membrane perturbation activity is usually determined by at least three mechanisms [3]. The best-characterised models, the 'barrel-stave' and the 'toroidal-pore' models, rely on the peptide ability to form transmembrane channels/pores, while in the so called 'carpet model', the peptides disrupt the bilayer in a detergent-like manner, eventually leading to the formation of micelles [4] (**Figure 1**). The mechanism of membrane disruption involves several molecular properties of the peptides, each one related to individual stages of the process:

- The process of cell attachment is facilitated by a positive net charge because of the bacteria membrane constituent.
- Aggregation facilitates the formation of a carpet on the outer side of the bacteria membrane, eventually leading to the destabilization of the lipidic bilayer. Amphipathic alpha-helical peptides better interact electrostatically with the target cell membrane.

- The overall lipophilicity rules the mechanism of permeation into the membrane, leading to a destabilization or a pore formation.

A balanced combination of these properties determines the mode of action and the overall peptide activity and cytotoxicity (**Figure 1**).

Recent research on AMPs has focused on methods to search through the constellation of known or predicted peptide sequences – either empirically or computationally – for molecules with desired properties and these approaches are continually evolving. Multi-scale approaches are increasingly applied to in silico rational design of bioactive molecules, because of their ability to study biophysical problems from multiple points of view. Multiscale approach for molecular design consists of at least two phases. The first (coarse grain) provides a fast exploration of the objective space, in order to sample its relevant regions in an approximate way. Afterwards, in a second phase, the coarse grain representation is transformed to a more detailed one, able to represent each aspect of the biological process.

Statistical-based peptide design and prediction methods are usually valid choices for unbiased screening, where speed and accuracy is a fundamental requirement. In these methods, the primary sequence information is associated with a measure of peptide activity – either quantitative or qualitative – through a series of sample sequences derived from experimentally validated peptides. A statistical model is then constructed by regression models and/or lexical methods in order to derive a rule explaining the biological activity. The derived model is then applied to stochastic or deterministic methods in order to explore the major possible number of candidates.

In contrast, computationally intensive biophysical studies are applied in order to valuate peptide folding, interaction and mode of action of a screened list of candidates. In particular, molecular dynamics (MD) has been extensively applied for the study of AMPs in order to unravel the molecular mechanisms supporting their activity. MD simulations target the motion of the molecular system by numerically solving Newton's dynamic equation. Different resolutions can be used in the simulations, varying from all–atom one, to different degrees of coarse grain, in which groups of atoms are packed into single interaction centers. From the motion of the studied systems biomolecular interactions can be inferred, and the molecular mechanisms underlying certain biological processes can be elucidated.

In this chapter, both statistical and MD design methods will be discussed. In the first part, common steps in statistical-based design strategies will be surveyed, from the dataset preparation procedure to the mathematical model training and validation. Furthermore, application of the designed model to deterministic and stochastic peptide design will be illustrated. The second part describes MD methods for modelling AMP and their interaction with the membrane. Finally, experimental procedures for in vitro validation and measure of AMP activity are listed.

2 - STATISTIC-BASED AMP DESIGN

In common statistic-based peptide design methods, a dataset of molecules is collected to extrapolate an adequate number of features in order to represent the desired activity. The dataset can contain quantitative information about the peptide activity such as MIC, or qualitative information such as active or inactive. In the latter case, the screening process will return a confidence score about peptide activity. Depending on the information available, each peptide in the dataset is encoded in some computer-friendly variables best representing the activity, and a regression or a classification algorithm is employed in order to distinguish peptide activity in a qualitative or quantitative fashion. In this paragraph, the process of dataset construction, model preparation and validation will be exhaustively outlined.

2.1 - Dataset preparation

In statistical analysis, the process of dataset preparation is one of the most delicate in model construction. During this phase, a list of peptides is collected in an ordered database and a specific activity is associated with primary and/or secondary structure information. Because of the remarkable variety of AMPs in terms of sequence and secondary structure, a rich and complete dataset of active and inactive peptides is difficult to obtain without introducing biases. For these reasons, during the years different bioinformatics methods were applied in order to collect as much information as possible, on natural and synthetic AMPs from the literature, facilitating the process of dataset preparation. Although information gathering can be automated (for example by iterative scanning of public sources [5]), because of the difficulty and sensitivity of the information crawling process, manually attended datasets are more appreciated (**Table 1**). AMPs datasets can be prepared ad-hoc by experimentally screening random peptides libraries. This method has the advantage of giving precise and uniform quantitative or qualitative information of the peptide activity [6], required by complex prediction models in order to fit the correspondingly large set of parameters. Solid-phase synthesis and high-throughput screening of large peptide arrays has become a common practice in drug discovery. However, systematic studies tent to limit the number of peptides by analyzing a fixed number of amino acids positions with a precise combination of substitution [7]. Indeed, the huge number of amino acidic combinations makes an exhaustive screening of random libraries unfeasible. For example, a full combinatorial assay of peptides with length up to 10 residues would result in 20^{10} different sequences, an unfeasible number of combinations. On the basis of the analysis of natural AMPs, the amino acidic space is limited to charged residues and moderately hydrophobic sequences; to avoid technical problems during the synthesis phase, cysteine and methionine residues are excluded, owing to potential cross linking or oxidation. In this way the number of combinations is extremely reduced, at the cost of some bias introduction, since a large number of substitutions are excluded *a priori*.

When the aim of the dataset preparation is to classify bioactive peptides, two or more different classes of sample peptides must be prepared. For the simplest case, a dataset of experimentally-validated AMPs must be compared with a

dataset of non-active peptides. Therefore, a list of inactive peptides must be compiled. Unfortunately, few peptides are annotated as non-antimicrobial in literature [8], therefore negative datasets must be inferred in different ways. One is the fuzzy and unbiased random selection of peptide fragments from datasets of known proteins. Obviously, this approach can cause the unwanted inclusion of bio-active peptides in the negative dataset. In order to reduce the possibility to introduce false negative, protein datasets can be screened with knowledge-based approaches. Gene Ontology (GO) annotations are used to mark experimentally- or computationally-known protein's functions and pathways, interactions and organelles involved in their function and activity [9]. These keywords can be combined to narrow the search process into particular districts or within specific functions. AMPs are usually released from the cell in the extracellular matrix, thus a possible strategy can be to exclude proteins and peptides marked as 'secreted' or exclusively present in specific cell compartments.

Care should be taken not to introduce bias in this process, as the bigger and wider the dataset, the more precise and complete the classification. Each particular class of protein should be represented equally, as the over-representation of a particular motif or amino-acidic combination can compromise the entire dataset. For this reason, peptide datasets are usually pruned for repetitive and over-representative sequences. CD-HIT [10] implements an algorithm that, given a threshold, clusters and trims out sequences based on their similarity. Usually, a threshold of 75 % of identity is enough to assure a proper variability in the dataset. An additional method to avoid overtraining is to split the dataset into a training set – for the model training – and a test set for the validation of the model performance.

Regression models have different requirements from classification models, as the resulting function must express a measure of AMP activity using a continuous function. A dataset containing quantitative values of experimentally tested activity is therefore mandatory. Even if precise and exhaustive datasets of AMPs with quantitative activity exists [11], their use in regression model is difficult. Data collected from different works and workgroups usually is scattered, resulting in imprecise and biased models. A solution can be to distinguish categories of AMPs, grouping together highly active peptides and low active peptides on the basis of a predetermined threshold. This choice allows for a certain tolerance, thus giving some quantitative information about peptide's activity.

2.2 – Peptide representation

In order to present the dataset to a classification or regression model, each sequence must be encoded in a computer-interpretable way, able to represent peptide's salient characteristic. Amino acids can be considered the basic unit of AMPs; therefore each peptide must be represented on the basis of its sequence composition and order. The simplest and most intuitive way to represent AMPs sequences is through a linguistic model, where sequences are considered as 'words' and amino acids are represented with one-letter code. As a consequence, text motives can be identified through

the analysis of recurrences and grammar rules, giving useful hints about the importance of specific amino acids and residue positions to peptide activity. However, such local approaches fail to account for amino acidic position-specific interactions. Furthermore, there is no understanding of the physicochemical variables influencing peptides activity. As an evolution of this grammar model, in order to introduce secondary structure information, different strategies have been adopted, like sequence alignment or position-specific scoring matrix (PSSM) [5,12]. However, these approaches are limited to natural amino acids, since there is not enough sequence information of non-natural amino acidic substitutions to build an exhaustive statistical model.

In the effort to overcome these limitations, quantitative structure-activity relationship (QSAR) models have been employed to describe the relationship between chemophysical characteristics and biological activity. These chemophysical characteristics, named *descriptors*, can be derived from experimental measures such as molecular weight, partition coefficient or HPLC retention time, but also theoretically calculated. Calculated descriptors can be related to peptide's primary structure or chemical composition, as well as secondary or tertiary structure. Moreover, single descriptors can be combined to describe different – but related – chemophysical characteristics, like polarity and hydrophobicity, in order to reduce variable hyperspace.

In AMP design and classification, the choice of representative QSAR descriptors is directly influenced by their mechanism of action. The positive net charge and hydrophobicity are important features for the attachment and the permeation of the bacteria membrane, respectively. It is likely that only those peptides which possess a balanced combination of these properties can achieve sufficient activity in each step of the concerted mechanism and attain higher levels of antimicrobial effects. Furthermore, the overall distribution of these chemicophysical properties influences the activity. As a consequence, global descriptors can be applied to account for whole-molecular properties - such as polarity, lipophilicity or molecular weight - while topological descriptors account for sequence order information and secondary structure (**Figure 2**). A measure of sequence information can be considered by analyzing the correlation between QSAR descriptors along the primary sequence. Auto and Cross-covariance (ACC) analysis is a measure originally introduced by Wold [13]. Although various methods have been employed [14–16], the concept remains that different chemophysical descriptors are correlated between each other in a given order along the primary sequence., Basically, for a given protein sequence, ACC variables describe the average interactions between residues distributed a certain *lag* apart throughout the whole sequence. Higher lag values result in describing distant interactions along the peptide sequence. Besides encoding the sequence order, ACC has the ability to transform each amino acid sequence of variable length into uniform equal-length vectors. This feature is very important in data mining methods, where a fixed length vector describing each instance is required. Even if each ACC variable is able to represent in a certain measure the amino acidic order along the primary sequence, their effectiveness should be evaluated for every

single case. A list of ACC descriptors is summarized in **Table 2**. Another method to include structure information can be integrated in the model by taking advantage of 3D structure information. Inductive QSAR descriptors are based on the intramolecular steric effects, electronegativities and intra ed inter molecular interaction energies [17]. However, it should be noted that these type of descriptors profoundly depends on AMP structure, therefore they are not suitable for the analysis of mixed datasets, where different structures are present.

2.3 - Prediction model

The development of novel AMPs and the optimization of known ones, require an understanding of how the activity is correlated to the molecular chemicophysical features. In order to develop such correlations, different statistical models and multivariate approaches can be employed. Advanced methods for data mining can be employed in connection to QSAR variables to quantitatively or qualitatively discriminate between AMP and non-AMP sequences. This paragraph is not meant to be exhaustive about this topic, however the most important and used techniques will be highlighted and discussed.

Depending on the type of dataset created and the information available, two main categories of models can be distinguished: regression models for a quantitative measure of the biological activity and classification models for the qualitative one (in this case, AMP or non-AMP). The choice of a prediction technique also involves a trade-off between model accuracy and meaningfulness. Linear methods have been widely used in AMPs design because of their simple calculation and interpretation. Principal Component Analysis (PCA) is a mathematical procedure able to transform a number of possibly correlated variables into a smaller number of uncorrelated ones called principal components. Support Vector Machine (SVM) is a linear method where two or more classes are represented in a variable hyperspace and each class is separated by critical boundary instances called *support vectors*. A linear discriminant function is then built to separate each class as widely as possible. On the other hand, nonlinear techniques, like artificial neural networks (ANN), are considered to give better results when the correlation between QSAR descriptors and biological activity is not completely clear. ANN is a mathematical model based on the simulation of some properties of biological neural networks. A network of descriptors is defined as *input nodes* or *neurons*. These nodes are connected together, forming a network that interacts in a hidden layer and sums up into an *output node*. For the purpose of classification, the nonlinear techniques are considered to give superior results, but at the cost of introducing rather opaque models that cannot easily be used to shed light on the underlying mechanisms involved.

Finally, decision trees are another method to classify an unknown instance in different classes. Each node in the tree represents a particular attribute to test. Unknown instances are routed down the tree according to the values of the attributes tested in successive nodes. The instance is then classified according to the class assigned to the leaf reached.

Random Forest (RF) is one of the most popular decision tree in biological data mining, mainly because of two important qualities: high prediction accuracy and information on variable importance for classification [18]. RF is an ensemble recursive partitioning method where many decision trees are trained using subsets of samples and descriptors with replacement. RF have been widely used in AMP prediction and optimization , with performances that compare well to other classification algorithms such as SVM and ANN [12].

RF has several properties that allow extracting relevant trends from data with complex variable relations, which are ubiquitous in data sets generated in the Life Sciences. The classification model can be analyzed *a-posteriori* to infer the similarity between samples, calculated as the number of times the two samples end up in the same terminal node of the tree [19,20]. In this way, cluster analysis can be applied to identify peptides that have similar features to other AMPs and direct the design to a particular branch of the tree.

After the choice of the statistical model, a required step can be the normalization of the descriptor set. In fact, depending on the chosen descriptors, the scale of values can be varied even of three or more logarithms. Thus, their normalization can help in improving the accuracy of the training. Some classification systems, otherwise, does not require a normalization phase. Generally speaking, decision Trees are robust enough to handle highly-varying variables, while ANN and SVM requires for descriptor hyperspace to be normalized. Another consideration is that redundant descriptors can condition the classification performance. Furthermore, in the selection of the descriptors a tradeoff should be found between the performance of the encoding and the requirement of minimizing the number of descriptors. Indeed, on equal terms of performance, a lower number of features is preferable, since the resulting model is less computationally expensive and the interpretation of resulting models is simpler. Therefore, a description selection procedure can be performed using automatic methods, such as genetic algorithms (GA) [21] or iterative methods like Incremental Feature Selection (IFS) [22].

2.4 – In silico sequence screening

Once that a sophisticate activity estimator model is constructed, an automatic method for the fast and efficient design and optimization of peptides must be adopted. AMPs design needs to explore a huge number of amino acidic combinations in order to perform an unbiased analysis of the probability space, therefore a deterministic approach would be unfeasible. Stochastic optimization methods, like Genetic Algorithms (GA) or Ant Colony Optimization have been extensively used in virtual peptide design[23,24]. In particular, GAs represents a versatile and powerful tool for AMP design. GAs are adaptive heuristic search algorithm premised on the evolutionary ideas of natural selection and genetic. The algorithm follow the principle of Nature adaptive approach to the environment, in which the evolution process is performed by successive generation or mutation and only the fittest individuals resists. Each potential AMP

candidate is treated like an entity belonging to a population, and the statistical model is used as a fitness function in order to reflect its biological activity. At the beginning of the selection process, a certain number of random sequences is generated. As the simulation goes on, the population tents to presents an increasing average fitness value, until convergence. In AMP design, sometime the simultaneous optimization of one or more conflicting objective is required, like the sequence length or a particular amino acidic composition. Multi Objective evolutional algorithms (MOEA) are a class of GA, able to optimize different objectives separately. As a result, a list of candidate solutions are screened without favouring one particular objective [25].

2.5 - Notes in statistic-based AMP design

- For the model training and validation is a good habit to have two distinct dataset, one for the training and the other one for the validation. However, when few data are available, the N-fold cross-validation is a good alternative. Basically, the dataset is divided into N parts, where N is usually set to 10. N-1are used parts for the model training, while the remaining part is used for validation. This operation is repeated N times and the average of the performance estimator (see below) is computed.
- A good choice of descriptors is imperative for a valid and non-redundant representation of the antimicrobial activity. A mix of global descriptors (describing the overall characteristics of the molecule) and topological descriptors (describing the distribution of them along the sequence) is suggested. Various methods are available for the systematic analysis of descriptor sets. However, one of the most used in literature, because of its simplicity of use is the Maximum Relevance, Minimum Redundancy (mRMR) method [22], where descriptors are sorted in descending order of importance on the basis of their relevance and redundance.
- The quality of a classification model can be measured by four parameters: true positive rate for sensitivity, false positive rate for selectivity, predictive accuracy and MCC, as defined below.

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$MCC = \frac{(TP \cdot TN) - (FN \cdot FP)}{\sqrt{(TP + FN)(TN + FP)(TP + FP)(TN + FN)}}$$

Where *TP*, *TN*, *FP* and *FN* are the number of true positive, true negative, false positive and false negative, respectively, resulting from the model. MCC is an important index used to evaluate the performance of the predictor when the dataset is not balanced. The MCC value ranges from -1 to +1, where a value above 0.5 is considered to be predictive.

For regression models, the Pearson correlation coefficient (PCC) is used as a predictive ability estimator:

$$PCC = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2} \sqrt{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}}$$

Where X_i and Y_i are the expected and predicted activity, respectively; N is the number of data points; $\overline{X}$ and $\overline{Y}$ are the average value of X and Y, respectively.

3 - MOLECULAR DYNAMICS SIMULATIONS OF AMPS

In Force-Field based Molecular Dynamics simulations, atoms in the system are propagated by numerically solving Newton's dynamic equation, with forces described by computationally amenable functions of the coordinates. The set of terms, including covalent interactions (describing bond stretching, angle bending, and dihedral torsion) and non-bonded interactions (electrostatics, hard-core repulsive and dispersive forces), is called the force field. The detail with which each molecule is described can vary from the highest resolution possible in all-atom methods, in which each atom is taken into account, to different degree of coarse grain, in which the atoms are suitably grouped into interaction centres, sometimes also grouping different small molecules together (for example 3-4 water molecules together). The result of these simulations is a trajectory (a sort of molecular "movie") recording the detailed dynamics of each molecule and how it interacts with the other components.

Current all-atom simulations of molecular systems relevant to this chapter, containing several tens of thousands of atoms, span timescale of hundreds of nanoseconds to some microseconds, the limiting factor being the small time step (1-2 fs) required to integrate Newton's equations of motion, resulting in 10^8-10^9 integration steps to reach these timescales. With coarse-grain force fields the simulation is sped up by two-three orders of magnitude thanks to i) the possibility to use longer timesteps (tens of fs) due to elimination of fast degrees of freedom, ii) fictitious speed up of the dynamics due to a smoother potential-energy surface, iii) the reduction in the number of interaction centres (though this is usually compensated by simulating systems of larger sizes).

MD simulations are playing a growing role in elucidating the mechanisms of peptide-bilayer interactions (for recent reviews see [26–30]). By computing the evolution of suitably prepared initial configurations one can in principle obtain atomic-resolution data on a vast variety of processes. However, due to the empirical nature of molecular mechanics force fields and to the necessarily limited sampling of the configuration space, MD simulations lack "absolute" prediction accuracy, and should be generally validated against experimental findings. Their role should be that of complementing experimental measurements providing the information needed to bridge the gap between the various experimental techniques.

This section provides a brief outline of issues and techniques specific to the MD simulations of AMPs. The reader is assumed to be familiar with the concepts behind MD simulations, such as the molecular mechanics force fields and the algorithms needed to solve Newton's equation of motion. For introductory material see [31].

3.1 - Force Fields

The force fields commonly employed for biomolecular simulations, and for simulation of AMPs in particular, are AMBER, CHARMM, GROMOS, and OPLS (for reviews and original references see [32][33]). Each of these is

actually a family of force fields, containing several versions of an original force field, based on a common parameterization strategy. A different version may therefore include extension to different molecules (Charmm36 contains the lipid force field, while the protein and nucleic acid part is that of Charmm27), different parameterization procedures (for example the partial charges in the ff03 Amber force field are obtained starting from a DFT quantum mechanics calculations, rather than the HF in the original version), or modification of certain torsion terms (for example with respect to Charmm22, Charmm27 contains an additional cross term for backbone torsions).

Validation studies, comparing several different force fields applied to peptide simulations [34] [35–37], have highlighted their strengths and drawbacks. Generally, the latest versions are better at reproducing a series of experimental findings such as peptide helix content, beta-hairpin formation, and NMR chemical shifts and coupling, though caution should be placed in using force fields out of the conditions in which they were parameterized (for example around standard temperature and pressure conditions, 300K and 1atm respectively).

Lipid force fields have been developed in connection to AMBER, CHARMM and GROMOS. The validation of these force fields is done by trying to reproduce physio-chemical properties of the bilayer for different lipid compositions (either homogeneous or mixtures), such as thickness, area per lipid, NMR order parameters (related to the order of the lipid alkyl chains), surface tension and isothermal area compressibility [38][39].

Force fields commonly used in peptide/lipid simulations treat electrostatic interactions using fixed partial charges sitting on the atom positions. As such, they do not account for polarization, i.e. the variation in electronic density in response to local electrostatic perturbations. The inclusion of these effects has been pursued for some time, though the use of polarizable force fields is still somewhat limited, due to higher computational costs and absence of extensive benchmarking/validation studies. Existing biomolecular force field accounting for polarization are, among others, Amoeba [40], SIBFA[41] and the polarizable versions of Amber, Amber ff02 [42]. Research is still active on these "next generation" force fields, and inclusion of polarization will be eventually needed to remedy for the deficiencies of additive (i.e. non-polarizable) force fields.

Currently available computational resources limit the size and timescales addressable with all-atoms force fields. An attractive way to speed up the calculations is to reduce the number of degrees of freedom by "coarse graining" (CG) the system, i.e. describing suitably chosen chemical group by single effective interaction centres [43][44]. Martini is a widely used coarse-grained force field for proteins and lipids [45], which has been specifically applied to peptide/bilayer simulations. The coarse graining in Martini is moderate, in that 3-4 atoms are grouped in "beads", so that single beads are assigned to the smallest amino acids such as Gly and Ala, while four beads are used to describe the biggest such as Tyr or Trp. With Martini the reachable temporal and spatial scales are expanded by 2-3 orders of magnitude, so that simulations of peptide insertion and assembly in the bilayer can be achieved. The disadvantages are

that peptide secondary structure needs be assigned *a priori*, so that no secondary structure change can be simulated. In addition the grouping of three water molecules in the same bead may conceal the observation of transient water filled pores, and implicit screening of charges may lead to overestimation of the energy required for pore formation [46]. To overcome such drawbacks multiscale approaches can be adopted, in which the resolution of the system is suitably changed from coarse grain to all atom and viceversa [47].

3.2 - Enhanced sampling schemes

Besides coarse graining, other schemes have been devised to overcome the problem of limited conformational sampling in MD simulations. These schemes may exploit collective variables tracing the relevant conformation states (umbrella sampling and metadynamics), or they may facilitate crossing of free-energy barrier through coupling with higher-temperature simulations (parallel tempering).

In Umbrella Sampling [48] a generalized coordinate $\xi(\boldsymbol{R})$ (also termed collective variable) is defined as function(s) of atom coordinates $\boldsymbol{R}$. In the context of peptide-bilayer simulations relevant coordinates may be the distance of the peptide center of mass to the bilayer center, or the peptide orientation with respect to the bilayer normal. The sought quantity is the free energy along the generalized coordinate, also called the potential of mean force $W(\xi)$, defined by

$$W(\xi) = -k_B T \ln(\rho(\xi))$$

where k_B is the Boltzmann constant and $\rho(\xi)$ is the equilibrium distribution of the coordinate. In principle a sufficiently long simulation would span the relevant configuration space, and from the distribution of ξ one could extract the potential of mean force. However the presence of free-energy barriers will generally restrain the simulation to limited free-energy basins. The umbrella sampling method forces the sampling of all relevant values of ξ by performing a sort of scan along ξ. This is accomplished by performing several simulations in which an extra term is added to the normal potential energy of the molecular system. This term may have the form

$$U = k\left(\xi(R) - \bar{\xi}_i\right)^2$$

where $\bar{\xi}_i$ are successive values of ξ, and k is a spring constant. In the case of the peptide-bilayer distance, the $\bar{\xi}_i$ may be suitably spaced value from 0 nm (peptide completely immersed in the bilayer) to 6 nm or more (peptide in the bulk solvent). For each window, $W(\xi)$ is obtained from the biased distribution of ξ during the MD simulation, $\rho_{Ui}(\xi)$, by

$$W_i(\xi) = -k_B T \ln(\rho_{Ui}(\xi)) - k(\xi - \bar{\xi}_i)^2 + \Delta_i$$

where Δ_i are unknown constants that may be found by matching together the various segment of $W(\xi)$. Clearly, for each simulation i the values of ξ will be restrained around $\bar{\xi}_i$. However, provided that there is enough overlapping between

the explored values of ξ, the continuous profile of $W(\xi)$ can be reconstructed automatically through, for example, the weighted histogram analysis method (WHAM) [49].

Though it is possible to perform multidimensional umbrella sampling, the number of needed simulation windows grows rapidly for two- and three-dimensional scans. In addition, a lot of computational time may be spent in "uninteresting" windows of ξ. The metadynamics approach[50], albeit less accurate than WHAM, at least in the original formulation, is both more amenable for treating multi collective variables and "self" regulating in the time spent exploring the various regions in the conformational space. The idea behind metadynamics is to perform an MD simulation where the system is "discouraged" to explore the same free-energy regions (described by the set of collective variables) by adding a history dependent potential that gradually fills the free-energy basins. In the original formulation, the potential energy is modified by periodically adding Gaussian functions with suitably chosen heights and widths, and centred on the current values of the ξ_i. The process is repeated until free diffusion in the collective variable space is achieved. The (one- or multi- dimensional) free-energy profile is then obtained as the negative of the sum of all added Gaussians. Several variants were based on the same idea of a history dependent potential: local elevation[51], conformational flooding[52], adaptively biased molecular dynamics[53], among others.

A common issue with both umbrella sampling and metadynamics methods is that they assume that the degrees of freedom orthogonal to the chosen collective variables be sufficiently sampled, i.e. that the relaxation times of these degrees of freedom are shorter than the time spent in each "bin" of the free energy surface. Through careful choice of the collective variables in multidimensional scans these problems can be alleviated, but still "hidden" variables coupled to the relevant reaction coordinate may play important roles. In lipid-membrane studies, a typical indicator of poor sampling in umbrella sampling simulations is the hysteresis between, for example, insertion of the peptide in the bilayer and extraction [54].

Parallel tempering, also known as replica exchange [55], enables free-energy barrier crossing by coupling the simulation at the desired temperature with higher-temperature simulations. This coupling is accomplished by exchanging the coordinates among the replica following a Metropolis scheme. More in detail, n replicas are evolved through MD, each maintained at a temperature T_i. After a number of MD steps an exchange between the coordinates of replica at T_i and T_{i+1} (the higher successive temperature in the ladder) is performed with a probability given by

$$p = \min\left(1, e^{(E_i - E_{i+1})\left(\frac{1}{k_B T_i} - \frac{1}{k_B T_{i+1}}\right)}\right)$$

i.e. the exchange is performed with probability 1 if E_{i+1} (the potential energy of replica i+1) is lower than E_i; otherwise it is performed with a probability given by the exponential term in the previous equation. This probabilistic exchange ensures the detailed balance condition, and that the MD at each temperature samples a canonical ensemble.

In a typical replica exchange molecular dynamics (REMD) simulation a set of temperatures from T_0=300K to T_n=600K-900K is chosen, and exchanges are attempted each 50-500 time steps. The spacing between successive replicas should be such as to allow for a 10%-40% successful exchanges, implying a suitable overlapping between potential energies distributions at different temperatures. Unfortunately, these distributions become narrower at increasing number of degrees of freedom, and for systems of tens or hundreds thousands atoms an unfeasible number of replica is needed (>500). A solution to this problem is provided by the so-called Hamiltonian replica exchange (HREX)[56,57] in which only a subsystem is "heated", by actually scaling its potential energy function. For example, one may choose to "heat" only the peptides, or peptides and bilayer: without the solvent degrees of freedom the number of replica is greatly reduced.

The schemes described above can be coupled together, and their simultaneous application may ensure both a sufficient sampling on the chosen coordinate, via Umbrella Sampling or Metadynamics, and rapid relaxation for the orthogonal degrees of freedom through REMD [58].

3.3 - Issues with peptide-bilayer simulations

This subsection schematically lists some of the points needing particular care in peptide-bilayer simulations. Explicit solvent simulations with periodic boundary conditions are assumed.

- Choice of the membrane model. With respect to the cellular membrane, the simulated systems contain only few lipid components and no protein or carbohydrate. They are closer to experimental studies involving artificial bilayers, with controlled lipid composition. However, for the peptide shown to destabilize the bilayer in artificial vesicles, simple bilayer models may for the most part be appropriate.
- The size of the simulated bilayer patch should be carefully chosen. Some peptide may act by selectively modifying the surface tension of the outer leaflet of the bilayer, thereby inducing curvature [33]. Such effects may be hidden by the use of periodic boundary conditions if the bilayer patch is too small.
- Different ensembles may be used in MD simulations. The microcanonical ensemble, NVE (constant energy, temperature, and particle number), is rarely used because it does not allow for temperature control and volume fluctuation. NVT, or constant temperature, is the ensemble of choice when simulating biomolecules in aqueous solvent. Several algorithms have been proposed to approximate the canonical ensemble, which may rely on stochastic terms or on the introduction of fictitious degree of freedom representing heat bath. In the isothermal-isobaric ensemble (NPT), pressure is controlled by suitably scaling the atomic coordinates, thereby changing the total volume. This can be accomplished with the Berendsen algorithm, the Nosé-Hoover Langevin piston

[59], or the Parrinello-Rahman method [60]. Bilayer simulations frequently use semi-isotropic pressure schemes, in which the control of pressure on the orientation normal to the bilayer is separated from the other two dimensions, i.e. the scaling in the lateral directions is independent from the one in the normal. This decoupling is required because of the different compressibility of water and of the bilayer. Clearly, choosing a constant volume ensemble fixes the area-per-lipid value, and this may correct for force-field artefacts. However, insertion of the peptides into the bilayer may require significant rearrangements for which flexibility in the lateral directions may be more realistic.

- Non-bonded interactions in principle require infinite summation over the pairs of particles in the periodic cells. For short-ranged potential, such as the r^{-6} attractive tail of the Lennard-Jones potential, cut-off schemes are used, i.e. only the particles within a certain distance (cut-off) are accounted for. Coulomb interactions are long ranged, so a cut-off approach (truncation schemes) may be too crude an approximation, leading to potential inconsistent behavior, such as artificial ordering [61]. The method of choice is the so-called Particle Mesh Ewald, or PME, in which the interaction is separated into a short-range and long-range part, calculated separately, the first using a cut off scheme, the second by spreading the charges on a 3D grid and accounting for all the periodic images. PME needs an overall neutral system, and failing to add neutralizing counterions may lead to serious artefact such as thinning of the bilayer [54]. The cut-off to be used for both Lennard-Jones and Coulomb interactions can vary in the range 8-12 Å, and is force-field dependent. In addition MD codes such as GROMACS or NAMD employ smoothing schemes to avoid the discontinuity at the cut-off. This schemes and the cut-off distance need be carefully tuned, and discussions on the topics can be found in the literature [38].

3.4 - Systems and processes

MD simulations can be used to predict structural properties of the peptide in different environments, such as, in order of complexity, water, organic solvents and water-organic solvent mixtures, micelles, and lipid bilayer (see Figure 3). Simulation of the monomeric AMPs, though not directly targeting their mechanism of action on the cellular membrane, are useful for extracting information such as presence and stability of secondary structure motives (alpha-helix, beta-sheets and turns), and other physicochemical characteristics such as solvent exposed surface. These quantities may be then related to antimicrobial activity and toxicity of various examined peptide sequences through multiple linear regression algorithms, such as in quantitative structure-activity relationship (QSAR) studies.

Structure predictions of peptides in solvents, water in particular, are achievable with an adequate degree of confidence, given that the force fields have been rather extensively tried and optimized for such tasks, as long as standard pressure

and temperature (around 1atm and 300K respectively) conditions are considered. In addition, the limited number of degrees of freedom of peptides allows for rather exhaustive sampling of their configuration space, at least with the enhanced sampling techniques mentioned below. Furthermore for this kind of systems the simulation protocols are rather robust and well established. Water-organic solvent mixtures can be used to asses structural properties in various environment. For example, MD simulations in pure water and water-TFE mixtures (TFE, or 2,2,2-trifluoroethanol, provides a low dielectric environment partially mimicking the conditions inside the bilayer) have been used to assist bioinformatics algorithms in designing novel AMP sequences [20]. A more realistic model of the environment inside lipid bilayer is provided by micelles, self-assembled structures of amphipatic molecules, with a highly hydrophobic interior and anionic or zwitterionic heads exposed to the solvent. Micelles mimicking the bilayer are used in NMR experiments, because of their faster relaxation times. Simulations of AMPs in micelles are at an intermediate level of complexity between those in solvent and lipid bilayer (see [62]).

The insertion of a peptide in the bilayer is a prohibitively slow process for all-atom MD simulations, and coarse-grained force fields or enhanced sampling techniques need to be used (see below). Studies of peptide structure and position inside the membrane may therefore start from an initial configuration where the peptide is already embedded in the bilayer. A different approach consists in starting from a non-formed bilayer, i.e. from a random mixture of water molecules and lipids, which are known to form a bilayer in tens to hundreds of nanoseconds[63]. In this way the self-assembly of the bilayer is simulated and the position of the peptide is not biased toward the starting configuration thanks to the high fluidity of the system during the self-assembly process.

Simulating the aggregation of several peptides in the bilayer is yet a more ambitious goal, because also the relative configurations of the various peptides need to be sampled. In MD studies of peptide aggregation in the bilayer the aggregates may be pre-assembled to study their stability and function, or the self-assembly process itself may be pursued. For example, different putative structure of a pore may be tried, and the stable ones be selected as the most probable structures [64], or peptides may be inserted at unbiased position in the bilayer and aggregation and pore formation be observed[65]. With CG force fields the whole process of peptide adsorption and pore formation [66,67], in systems of thousands of peptides and lipid patches of lateral dimension up to 0.1 μm [67].

4 - EXPERIMENTAL VALIDATION OF AMPS: MINIMAL INHIBITORY CONCENTRATION (MIC)

The *in vitro* activity of AMPs is tested using the Microtitre Broth Dilution Method in order to determine the MIC value, as recommended for the antibiotic testing by the NCLSS (National Committee of Laboratory Safety and Standards) [70].

Here, we suggest a modified version of this method as recommended by R. E. W. Hancock (University of British Columbia, Vancouver, British Columbia, Canada) for testing antimicrobial peptides (http: //www.interchg.ubc.ca/bobh/MIC.htm).

4.1 - Materials

1. Sterile tubes (15ml)
2. Mueller Hinton Broth (MHB)
3. Mueller Hinton agar plates (MHA)
4. Sterile 96-well polypropylene microtitre plates
5. Polypropylene microcentrifuge tubes
6. Sterile petri dishes
7. sterile deionized water (dH_2O)

4.2 - Methods

1. Inoculate 5 ml MHB in tubes with test strains from MHA plates and grow overnight at 37^{o}C on a shaker (160 rpm).
2. Make serial dilutions of test peptides in . sterile deionized water in polypropylene tubes:

- dissolve test peptide in dH_2O at 10 times the required maximal concentration;

- do two-fold dilutions in dH_2O to get serial dilutions of peptides at 10 times the required test concentrations, eg., 640, 320, 160, ...2.5 µg/ml.

3. Dilute overnight bacterial cultures in MHB to give 5 x 10^5 colony forming units/ml.

4. Dispense 90 µl of bacterial suspension in each well from column 1 to column 11. Do not add bacteria to column 12, and instead dispense 100 µl of MHB (sterility control and blank for the plate scanner).

5. Add 10 µl of 10x test peptide each well from column 1 to column 10 (column 11 is a control for bacteria alone, with no peptide, where 10 µl of dH_2O is added).

6. Incubate the plates at 37^{o}C for 18-24 hours.

7. MIC can be taken as the lowest concentration of drug that reduces growth by more than 50%.

8. Plate l0 µl 10^{-6} dilution of overnight cultures on MHA plates to determine a viable count. The MBC (Minimal bactericidal concentration) can be determined by plating out the contents of the first 3 wells showing no visible growth of bacteria onto MHA plates and incubating at 37^{o}C for 18 hr. MBC is defined as the lowest concentration of the peptide causing a reduction in the numbers of viable bacteria of $\geq$3 $\log_{10}$ with respect to the CFU/mL inoculated.

4.3 – Notes

- It is important that you use the material mentioned above. For example, do not substitute polystyrene for polypropylene tubes or microtitre plates. Cationic peptides bind polystyrene (especially "tissue culture treated" polystyrene).

REFERENCE

1. Di Luca M, Maccari G, Nifosì R (2014) Treatment of Microbial Biofilms in the Post Antibiotic Era: Prophylactic and Therapeutic Use of Antimicrobial Peptides and Their Design by Bioinformatics Tools. Pathog Dis. Available: http://www.ncbi.nlm.nih.gov/pubmed/24515391. Accessed 14 February 2014.

2. Salomone F, Cardarelli F, Signore G, Boccardi C, Beltram F (2013) In vitro efficient transfection by CM18-Tat11 hybrid peptide: a new tool for gene-delivery applications. PLoS One. doi:10.1371/journal.pone.0070108.

3. Bahar A, Ren D (2013) Antimicrobial Peptides. Pharmaceuticals 6: 1543–1575. Available: http://www.mdpi.com/1424-8247/6/12/1543/. Accessed 29 November 2013.

4. Shai Y, Oren Z (2001) From "carpet" mechanism to de-novo designed diastereomeric cell-selective antimicrobial peptides. Peptides 22: 1629–1641. Available: http://www.ncbi.nlm.nih.gov/pubmed/11587791. Accessed 29 December 2012.

5. Fjell CD, Hancock REW, Cherkasov A (2007) AMPer: a database and an automated discovery tool for antimicrobial peptides. Bioinformatics 23: 1148–1155. Available: http://www.ncbi.nlm.nih.gov/pubmed/17341497. Accessed 10 May 2013.

6. Rathinakumar R, Wimley WC (2008) Biomolecular engineering by combinatorial design and high-throughput screening: small, soluble peptides that permeabilize membranes. J Am Chem Soc 130: 9849–9858. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2582735&tool=pmcentrez&rendertype=abstract. Accessed 22 May 2013.

7. Marks JR, Placone J, Hristova K, Wimley WC (2011) Spontaneous membrane-translocating peptides by orthogonal high-throughput screening. J Am Chem Soc 133: 8995–9004. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3118567&tool=pmcentrez&rendertype=abstract. Accessed 3 January 2013.

8. Wang P, Hu L, Liu G, Jiang N, Chen X, et al. (2011) Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. PLoS One 6: e18476. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3076375&tool=pmcentrez&rendertype=abstract. Accessed 15 March 2012.

9. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25: 25–29. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3037419&tool=pmcentrez&rendertype=abstract. Accessed 21 January 2014.

10. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22: 1658–1659. Available: http://www.ncbi.nlm.nih.gov/pubmed/16731699. Accessed 30 July 2012.

11. Piotto SP, Sessa L, Concilio S, Iannelli P (2012) YADAMP: yet another database of antimicrobial peptides. Int J Antimicrob Agents 39: 346–351. Available: http://www.ncbi.nlm.nih.gov/pubmed/22325123. Accessed 23 August 2012.

12. Thomas S, Karnik S, Barai RS, Jayaraman VK, Idicula-Thomas S (2010) CAMP: a useful resource for research on antimicrobial peptides. Nucleic Acids Res 38: D774–80. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2808926&tool=pmcentrez&rendertype=abstract. Accessed 23 August 2012.

13. Wold S, Jonsson J, Sjörström M, Sandberg M, Rännar S (1993) DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 277: 239–253. Available: http://linkinghub.elsevier.com/retrieve/pii/000326709380437P. Accessed 23 July 2012.

14. Sokal RR, Thomson BA (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. Am J Phys Anthropol 129: 121–131. Available: http://www.ncbi.nlm.nih.gov/pubmed/16261547. Accessed 13 February 2014.

15. Horne DS (1988) Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities. Biopolymers 27: 451–477. Available: http://www.ncbi.nlm.nih.gov/pubmed/3359010. Accessed 13 February 2014.

16. Feng ZP, Zhang CT (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. J Protein Chem 19: 269–275. Available: http://www.ncbi.nlm.nih.gov/pubmed/11043931. Accessed 13 February 2014.

17. Jaiswal K, Naik PK (2008) Distinguishing compounds with anticancer activity by ANN using inductive QSAR descriptors. Bioinformation 2: 441–451. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2561164&tool=pmcentrez&rendertype=abstract. Accessed 13 February 2014.

18. Michaelson JJ, Sebat J (2012) forestSV: structural variant discovery through statistical learning. Nat Methods 9: 819–821. Available: http://www.ncbi.nlm.nih.gov/pubmed/22751202. Accessed 24 August 2012.

19. Touw WG, Bayjanov JR, Overmars L, Backus L, Boekhorst J, et al. (2012) Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform. Available: http://www.ncbi.nlm.nih.gov/pubmed/22786785. Accessed 17 July 2012.

20. Maccari G, Di Luca M, Nifosí R, Cardarelli F, Signore G, et al. (2013) Antimicrobial peptides design by evolutionary multiobjective optimization. PLoS Comput Biol 9: e1003212. Available: http://www.ploscompbiol.org/article/metrics/info:doi/10.1371/journal.pcbi.1003212. Accessed 23 September 2013.

21. Hansen L, Lee EA, Hestir K, Williams LT, Farrelly D (2009) Controlling feature selection in random forests of decision trees using a genetic algorithm: classification of class I MHC peptides. Comb Chem High Throughput Screen 12: 514–519. Available: http://www.ncbi.nlm.nih.gov/pubmed/19519331. Accessed 14 February 2014.

22. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27: 1226–1238. Available: http://www.ncbi.nlm.nih.gov/pubmed/16119262. Accessed 23 July 2012.

23. Hiss JA, Bredenbeck A, Losch FO, Wrede P, Walden P, et al. (2007) Design of MHC I stabilizing peptides by agent-based exploration of sequence space. Protein Eng Des Sel 20: 99–108. Available: http://www.ncbi.nlm.nih.gov/pubmed/17314106. Accessed 14 February 2014.

24. Fjell CD, Jenssen H, Cheung WA, Hancock REW, Cherkasov A (2011) Optimization of antibacterial peptides by genetic algorithms and cheminformatics. Chem Biol Drug Des 77: 48–56. Available: http://www.ncbi.nlm.nih.gov/pubmed/20942839. Accessed 25 May 2012.

25. Deb K, Pratap A, Agarwal S, Meyarivan T (2002) A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans Evol Comput 6: 182–197. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=996017. Accessed 14 July 2012.

26. Bocchinfuso G, Bobone S, Mazzuca C, Palleschi A, Stella L (2011) Fluorescence spectroscopy and molecular dynamics simulations in studies on the mechanism of membrane destabilization by antimicrobial peptides. Cell Mol Life Sci 68: 2281–2301. Available: http://www.ncbi.nlm.nih.gov/pubmed/21584808. Accessed 6 August 2013.

27. Gurtovenko AA, Anwar J, Vattulainen I (2010) Defect-mediated trafficking across cell membranes: insights from in silico modeling. Chem Rev 110: 6077–6103. Available: http://www.ncbi.nlm.nih.gov/pubmed/20690701. Accessed 7 August 2013.

28. Marrink SJ, de Vries AH, Tieleman DP (2009) Lipids on the move: simulations of membrane pores, domains, stalks and curves. Biochim Biophys Acta 1788: 149–168. Available: http://www.ncbi.nlm.nih.gov/pubmed/19013128. Accessed 7 August 2013.

29. Bolintineanu DS, Kaznessis YN (2011) Computational studies of protegrin antimicrobial peptides: a review. Peptides 32: 188–201. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3013618&tool=pmcentrez&rendertype=abstract. Accessed 7 August 2013.

30. Chen L, Gao L (2012) How the Antimicrobial Peptides Kill Bacteria: Computational Physics Insights. Commun Comput Phys. Available: http://www.global-sci.com/issue/abstract/readabs.php?vol=11&page=709&issue=3&ppage=725&year=2012. Accessed 7 August 2013.

31. Leach A (2001) Molecular Modelling: Principles and Applications (2nd Edition). 2nd ed. Prentice Hall.

32. Ponder JW, Case DA (2003) Force Fields for Protein Simulations. In: Daggett V, editor. Protein Simulations. Academic Press, Vol. 66. pp. 27–85. doi:10.1016/S0065-3233(03)66002-X.

33. Mackerell AD (2004) Empirical force fields for biological macromolecules: Overview and issues. J Comput Chem 25: 1584–1604. doi:10.1002/jcc.20082.

34. Lange OF, van der Spoel D, de Groot BL (2010) Scrutinizing molecular mechanics force fields on the submicrosecond timescale with NMR data. Biophys J 99: 647–655. doi:10.1016/j.bpj.2010.04.062.

35. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, et al. (2012) Systematic validation of protein force fields against experimental data. PLoS One 7: e32131. Available: http://dx.plos.org/10.1371/journal.pone.0032131. Accessed 21 May 2013.

36. Beauchamp KA, Lin Y-S, Das R, Pande VS (2012) Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. J Chem Theory Comput 8: 1409–1414. doi:10.1021/ct2007814.

37. Cino E a, Choy W-Y, Karttunen M (2012) Comparison of Secondary Structure Formation Using 10 Different Force Fields in Microsecond Molecular Dynamics Simulations. J Chem Theory Comput 8: 2725–2740. doi:10.1021/ct300323g.

38. Piggot TJ, Piñeiro Á, Khalid S (2012) Molecular Dynamics Simulations of Phosphatidylcholine Membranes: A Comparative Force Field Study. J Chem Theory Comput 8: 4593–4609. doi:10.1021/ct3003157.

39. Jämbeck JPM, Lyubartsev AP (2012) An Extension and Further Validation of an All-Atomistic Force Field for Biological Membranes. J Chem Theory Comput 8: 2938–2948. doi:10.1021/ct300342n.

40. Shi Y, Xia Z, Zhang J, Best R, Wu C, et al. (2013) The Polarizable Atomic Multipole-based AMOEBA Force Field for Proteins. J Chem Theory Comput 9: 4046–4063. doi:10.1021/ct4003702.

41. Guo, H., Gresh, N., Roques, B. P., and Salahub DR (2000) No Title. J Phys Chem B 104: 9746–9754.

42. Cieplak P, Caldwell J, Kollman P (2001) Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/. J Comput Chem 22: 1048–1057. doi:10.1002/jcc.1065.

43. Tozzini V (2005) Coarse-grained models for proteins. Curr Opin Struct Biol 15: 144–150. Available: http://www.ncbi.nlm.nih.gov/pubmed/15837171. Accessed 3 June 2013.

44. Baaden M, Marrink SJ (2013) Coarse-grain modelling of protein-protein interactions. Curr Opin Struct Biol 23: 878–886. doi:10.1016/j.sbi.2013.09.004.

45. Monticelli L, Kandasamy SK, Periole X, Larson RG, Tieleman DP, et al. (2008) The MARTINI Coarse-Grained Force Field: Extension to Proteins. J Chem Theory Comput 4: 819–834. doi:10.1021/ct700324x.

46. Bennett WFD, Tieleman DP (2011) Water Defect and Pore Formation in Atomistic and Coarse-Grained Lipid Membranes : Pushing the Limits of Coarse Graining. 12: 2981–2988.

47. Ayton GS, Noid WG, Voth G a (2007) Multiscale modeling of biomolecular systems: in serial and in parallel. Curr Opin Struct Biol 17: 192–198. doi:10.1016/j.sbi.2007.03.004.

48. Torrie GM, Valleau JP (1977) Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. J Comput Phys 23: 187–199. doi:10.1016/0021-9991(77)90121-8.

49. Roux B (1995) The calculation of the potential of mean force using computer simulations. Comput Phys Commun 91: 275–282. doi:DOI: 10.1016/0010-4655(95)00053-I.

50. Laio A, Parrinello M (2002) Escaping free-energy minima. Proc Natl Acad Sci U S A 99: 12562–12566. doi:10.1073/pnas.202427399.

51. Huber T, Torda AE, Gunsteren WF (1994) Local elevation: A method for improving the searching properties of molecular dynamics simulation. J Comput Aided Mol Des 8: 695–708. doi:10.1007/BF00124016.

52. Grubmüller H (1995) Predicting slow structural transitions in macromolecular systems: Conformational flooding. Phys Rev E 52: 2893–2906. doi:10.1103/PhysRevE.52.2893.

53. Adamson S, Kharlampidi D, Dementiev A (2008) Stabilization of resonance states by an asymptotic Coulomb potential. J Chem Phys 128: 024101. doi:10.1063/1.2821102.

54. Yesylevskyy S, Marrink S-J, Mark AE (2009) Alternative mechanisms for the interaction of the cell-penetrating peptides penetratin and the TAT peptide with lipid bilayers. Biophys J 97: 40–49. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2711361&tool=pmcentrez&rendertype=abstract. Accessed 6 August 2013.

55. Sugita Y, Yuko Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314: 141–151.

56. Sugita Y, Okamoto Y (2000) Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. 329.

57. Wang L, Friesner RA, Berne BJ (2011) Replica exchange with solute scaling: a more efficient version of replica exchange with solute tempering (REST2). J Phys Chem B 115: 9431–9438. doi:10.1021/jp204407d.

58. Bussi G, Gervasio FL, Laio A, Parrinello M (2006) Free-energy landscape for beta hairpin folding from combined parallel tempering and metadynamics. J Am Chem Soc 128: 13435–13441. doi:10.1021/ja062463w.

59. Feller SE, Zhang Y, Pastor RW, Brooks BR (1995) Constant pressure molecular dynamics simulation: The Langevin piston method. J Chem Phys 103: 4613. doi:10.1063/1.470648.

60. Parrinello M (1981) Polymorphic transitions in single crystals: A new molecular dynamics method. J Appl Phys 52: 7182. doi:10.1063/1.328693.

61. Patra M, Karttunen M, Hyvönen MT, Falck E, Vattulainen I (2004) Lipid Bilayers Driven to a Wrong Lane in Molecular Dynamics Simulations by Subtle Changes in Long-Range Electrostatic Interactions. J Phys Chem B 108: 4485–4494. doi:10.1021/jp031281a.

62. Langham A, Kaznessis YN (2010) Molecular simulations of antimicrobial peptides. Methods Mol Biol 618: 267–285. doi:10.1007/978-1-60761-594-1_17.

63. Venturoli M, Smit B (1999) Simulating the self-assembly of model membranes. PhysChemComm 2: 45. doi:10.1039/a906472i.

64. Peter Tieleman D, Hess B, Sansom MSP (2002) Analysis and Evaluation of Channel Models: Simulations of Alamethicin. Biophys J 83: 2393–2407. Available: http://linkinghub.elsevier.com/retrieve/pii/S0006349502752533. Accessed 7 August 2013.

65. Thøgersen L, Schiøtt B, Vosegaard T, Nielsen NC, Tajkhorshid E (2008) Peptide aggregation and pore formation in a lipid bilayer: a combined coarse-grained and all atom molecular dynamics study. Biophys J 95: 4337–4347. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2567951&tool=pmcentrez&rendertype=abstract. Accessed 7 August 2013.

66. Gkeka P, Sarkisov L (2009) Spontaneous formation of a barrel-stave pore in a coarse-grained model of the synthetic LS3 peptide and a DPPC lipid bilayer. J Phys Chem B 113: 6–8. doi:10.1021/jp808417a.

67. Woo H-J, Wallqvist A (2011) Spontaneous buckling of lipid bilayer and vesicle budding induced by antimicrobial peptide magainin 2: a coarse-grained simulation study. J Phys Chem B 115: 8122–8129. Available: http://www.ncbi.nlm.nih.gov/pubmed/21651300. Accessed 7 August 2013.

68. Perrin BS, Tian Y, Fu R, Grant C V, Chekmenev EY, et al. (2014) High-Resolution Structures and Orientations of Antimicrobial Peptides Piscidin 1 and Piscidin 3 in Fluid Bilayers Reveal Tilting, Kinking, and Bilayer Immersion. J Am Chem Soc. Available: http://www.ncbi.nlm.nih.gov/pubmed/24410116. Accessed 14 February 2014.

69. Parton DL, Akhmatskaya E V, Sansom MSP (2012) Multiscale simulations of the antimicrobial peptide maculatin 1.1: water permeation through disordered aggregates. J Phys Chem B 116: 8485–8493. doi:10.1021/jp212358y.

70 National Committee for Clinical Laboratory Standards. 2000. Methods for dilution antimicrobial susceptibility tests for bacteria that grow aerobically; M7-A5, 5th ed. National Committee for Clinical Laboratory Standards, Wayne, Pa.

FIGURES

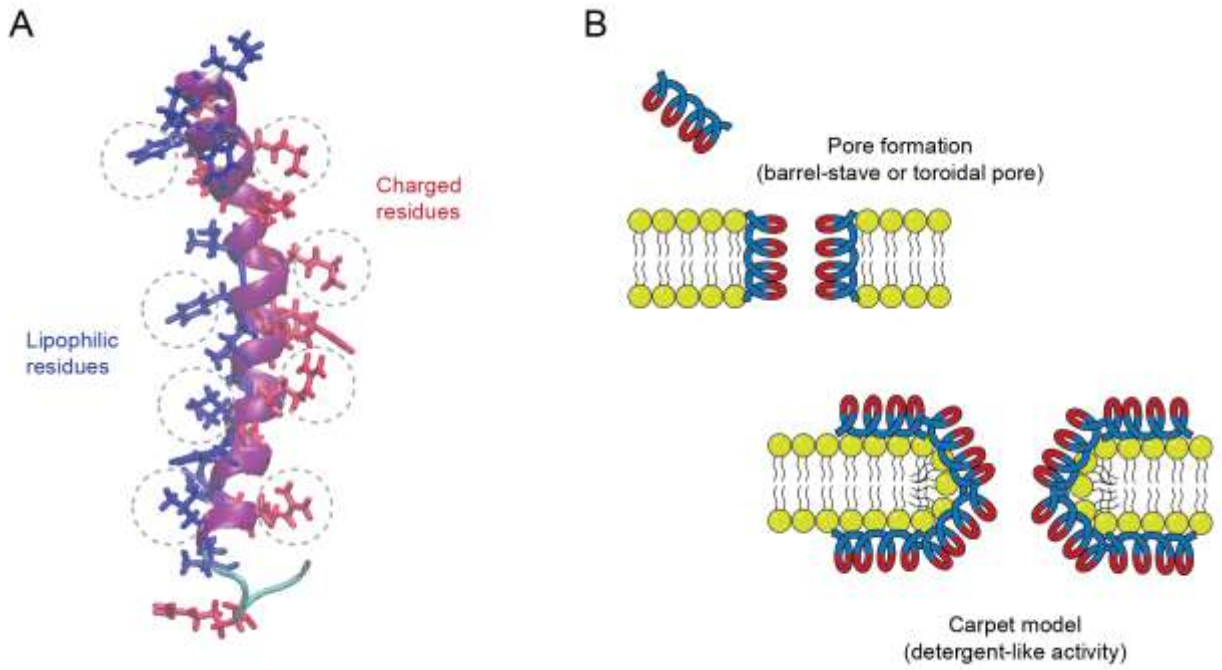

Figure 1 – AMPs chemophysical features and mode of action. A) NMR structure of LL-37 (PMID: 18818205). In red are highlighted charged residues, while in blue lipophilic ones. B) AMPs membrane perturbation activity. Top left, the 'barrel-stave' and 'toroidal-pore' models; bottom-right, the carpet model.

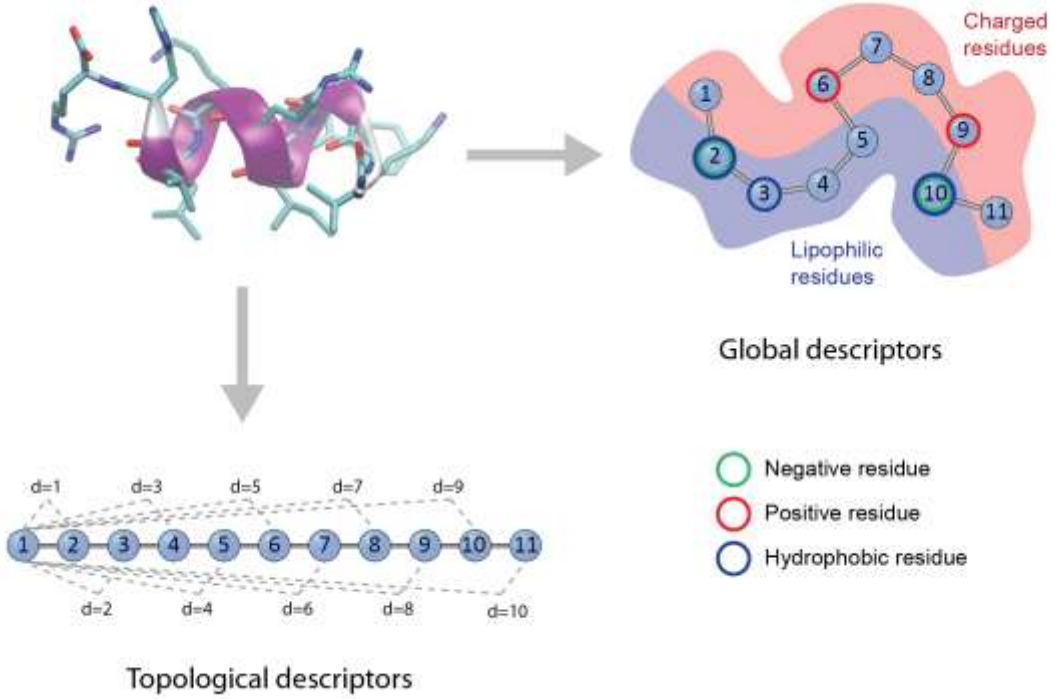

Figure 2 – Schematic representation of the feature selection process. Global and topological features are selected in order to represent the overall chemophysical characteristics and their distribution, respectively.

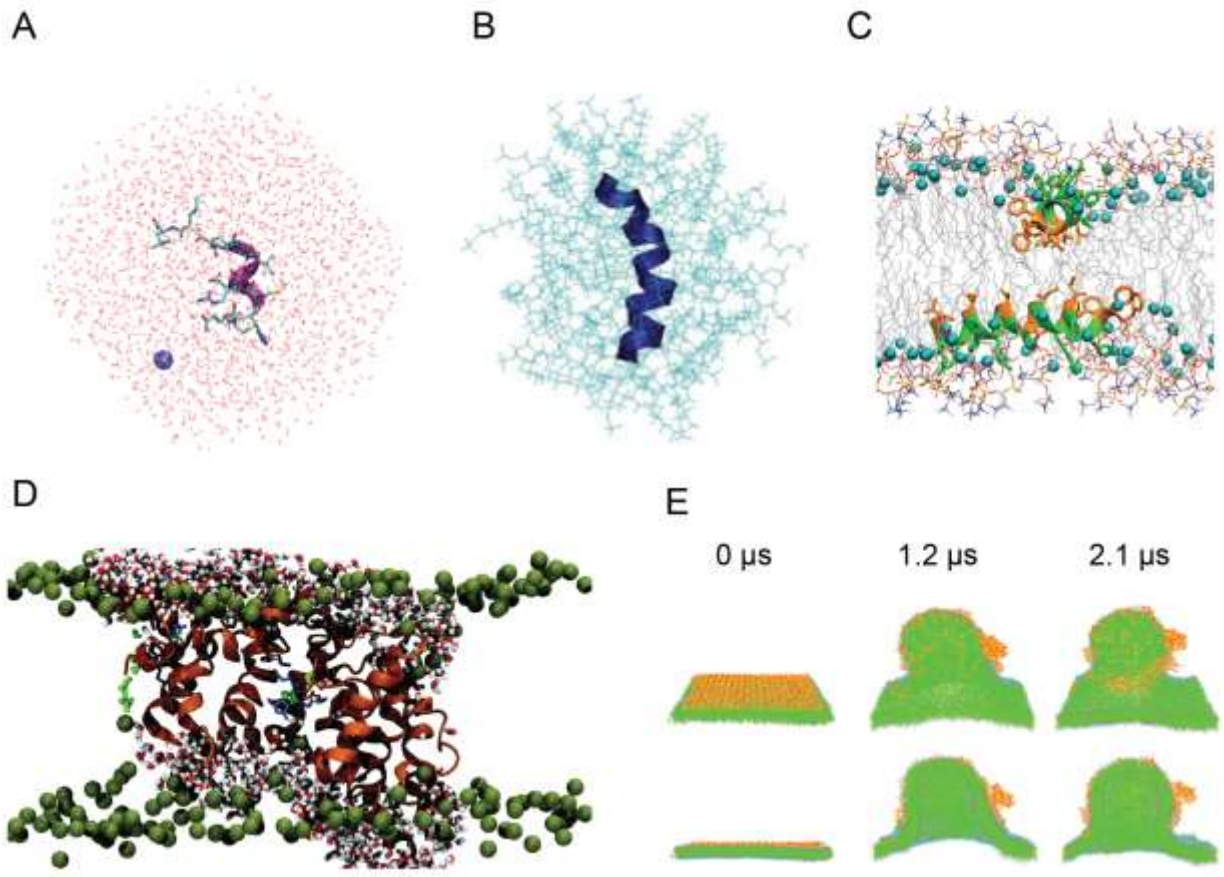

Figure 3 – MD simulation of AMPs. A) Snapshots from all-atom simulation of a peptide in water and B) in a micelle C) Snapshot from all-atom simulation of AMP Piscidin 1 and Piscidin 3 in lipid bilayers. Reprinted (adapted) with permission from [68]. Copyright 2011 American Chemical Society. D) All atom MD simulation of pore formation by a cluster of 16 Maculatin 1.1 peptides (orange) in a lipid bilayer. Reprinted (adapted) with permission from [69], copyright (2012) American Chemical Society E) Snapshots from coarse grained simulations of several Magainine 2 peptides (orange) placed on one side of a pure DPPC bilayer (in green). Reprinted with permission from [67]. Copyright 2011 American Chemical Society.

TABLES

Year	Database	Web site	Content
2002	AMSDb	http://www.bbcm.univ.trieste.it/~tossi/pag1.htm	Plant and Animal AMPs
2007	AMPer	http://marray.cmdr.ubc.ca/cgi-bin/amp.pl	Plant and Animal AMPs
2007	BACTIBASE	http://bactibase.pfba-lab-tun.org/main.php	Bacteriocins
2008	RAPD	http://faculty.ist.unomaha.edu/chen/rapd/	Recombinant AMPs
2009	PhytAMP	http://phytamp.pfba-lab-tun.org/main.php	Plant AMPs
2009	APD2	http://aps.unmc.edu/AP/main.php	Natural AMPs
2010	CAMP	http://www.bicnirrh.res.in/antimicrobial/	All AMPs
2012	DAMPD	http://apps.sanbi.ac.za/dampd/	All AMPs
2012	YADAMP	http://yadamp.unisa.it/	All AMPs
2014	BAAMPS	http://www.baamps.it/	Biofilm-active AMPs

Table 1 – A chronological list of AMPs databases.

Name	**Formula**	**Description**
Normalize Moreau-Broto autocorrelation [16]	$F(d)=\sum_{i=1}^{N-d} P_i \cdot P_{i+d}$	Properties values are used as a measure of spatial autocorrelation.
Moran autocorrelation [15]	$F(d)=\frac{\frac{1}{N-d}\sum_{i=1}^{N-d}\left(P_i-\overline{P}\right)\left(P_{i+d}-\overline{P}\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(P_i-\overline{P}\right)^2};\ \overline{P}=\frac{\sum_{i=1}^{N} P_i}{N}$	Property deviations from the average values as a measure of spatial autocorrelation
Geary autocorrelation [14]	$F(d)=\frac{\frac{1}{2(N-d)}\sum_{i=1}^{N-d}\left(P_i-P_{i+d}\right)^2}{\frac{1}{N-1}\sum_{i=1}^{N}\left(P_i-\overline{P}\right)^2}$	Square difference of property values as a measure of spatial autocorrelation

Table 2. Auto- and cross- correlation descriptors. d is defined as the lag of the autocorrelation; Pi and Pi+d are the normalized properties of the amino acid at position i and i+d respectively;