Prediction of Antimicrobial Activity of Peptides using Relational Machine Learning

Andrea Szabóová, Ondřej Kuželka, Filip Železný
Department of Cybernetics
Czech Technical University in Prague
Prague, Czech Republic
Email: {szaboand, kuzelon2, zelezny}@fel.cvut.cz

***Abstract*—We apply relational machine learning techniques to predict antimicrobial activity of peptides. We follow our successful strategy (Szabóová et al., MLSB 2010) to prediction of DNA-binding propensity of proteins from structural features. We exploit structure prediction methods to obtain peptides' spatial structures, then we construct the structural relational features. We use these relational features as attributes in a regression model. We apply this methodology to antimicrobial activity prediction of peptides achieving better predictive accuracies than a state-of-the-art approach.**

***Keywords*-Antimicrobial activity prediction, peptides, relational machine learning, data mining.**

I. Introduction

Antimicrobial peptides are molecules responsible for defence against microbial infections in the first stages of the immunological response. Recently antimicrobial peptides have been recognized as a potential replacement of conventional antibiotics for which some microorganisms had already acquired resistance. Although, there are theories about the mechanisms by which antimicrobial peptides kill pathogenic microorganisms, the process has not been fully uncovered yet. Several computational approaches have been developed in past years to predict antimicrobial activity of peptides [1], [2].

In this paper we are concerned with prediction of antimicrobial activity from modelled spatial structure information. We utilize our relational learning method [3] which has already been used for DNA-binding propensity prediction of proteins [4] following earlier work of Nassif et al. [5] in a similar context. Whereas our method has been only used for classification problems so far, here we use it for regression (prediction of antimicrobial activity, which is a continuous variable). We show that this relational learning method for regression improves on a state-of-the-art approach to antimicrobial activity prediction in terms of predictive accuracy. Another positive aspect of our method is that it provides us with interpretable features. Finally, our method is not bound to prediction of antimicrobial activity, but it can be used to predict also other properties of peptides like their hemolytic activity - also examined in this paper.

Data, source codes and executables can be downloaded from *http://ida.felk.cvut.cz/peptides/BIBM2012.zip*.

II. Antimicrobial Peptides

Antimicrobial peptides (AMPs) have been actively researched for their potential therapeutic application against infectious diseases. AMPs are amino acid sequences of length typically from 6 to 100. They are produced by living organisms of various types as part of their innate immune system [6]. They express potent antimicrobial activity and are able to kill a wide range of microbes. In contrast to conventional antibiotics, AMPs are bacteriocidal (i.e. bacteria killer) instead of bacteriostatic (i.e. bacteria growth inhibitor). Most AMPs work directly against microbes through a mechanism which starts with membrane disruption and subsequent pore formation, allowing efflux of essential ions and nutrients. According to current view this mechanism works as follows: AMPs bind to the cytoplasmic membrane and create micelle-like aggregates, which leads to disruption of the membrane. In addition, there may be complementary mechanisms such as intracellular targeting of cytoplasmic components crucial to proper cellular physiology. Thus, the initial interaction between the peptides and the microbial cell membrane allows the peptides to penetrate into the cell to disrupt vital processes, such as cell wall biosynthesis and DNA, RNA, and protein synthesis. A convenient property of AMPs is their selective toxicity to microbial targets, which makes them non-toxic to mammalian cells. This specificity is based on the significant distinctions between mammalian and microbial cells, such as composition, transmembrane potential, polarization and structural features.

Antimicrobial peptides are small, positively charged, amphipathic molecules. They include two or more positively charged residues and a large proportion of hydrophobic residues. Many AMPs exist in relatively unstructured conformations prior to interaction with target cells. Upon binding to pathogen membranes, peptides may undergo significant conformational changes to helical or other structures. These conformations of antimicrobial peptides may impact their selective toxicity [7]. The three-dimensional folding of the peptides results in the hydrophilic or charged amino acids segregating in space from the hydrophobic residues, leading to either an amphipathic structure, or a structure with two charged regions spatially separated by a hydrophobic segment. Such a structure can interact with the membrane [8].

The amphipathicity of the AMPs enables insertion into the membrane lipid bilayer.

III. Existing Approaches to Activity Prediction

Several methods have been developed to predict antimicrobial activity of AMPs with potential therapeutic application. Some algorithms take advantage of data mining and high-throughput screening techniques and apply attribute-value approach to scan protein and peptide sequences [9], [10]. Similar strategies were proposed based on supervised learning techniques, such as artificial neural networks or support vector machines, in order to evaluate amounts of complex data [11]. Most attempts have been focused to the prediction of peptide's activity using quantitative structure-activity relationships (QSAR) descriptors together with artificial neural networks [12], [13], [1], linear discriminant [14] or principal component analysis [15]. A QSAR-based artificial neural network system was experimentally validated using SPOT high-throughput peptide synthesis, demonstrating that this methodology can accomplish a reliable prediction [16]. Recently, an artificial neural network approach based on the peptide's physicochemical properties has been introduced [2]. These properties were derived from the peptide sequence and were suggested to comprise a complete set of parameters accurately describing antimicrobial peptides.

The approach that we propose in this paper differs from the above mentioned approaches mainly in the following. Rather than using an ad-hoc set of physicochemical properties of the peptides, we use an automatic feature construction method based on relational machine learning to discover structural patterns capturing spatial configuration of amino acids in peptides. A positive aspect of our method (besides improving predictive accuracy) is that it provides us with interpretable features involving spatial configurations of selected amino acids. Moreover, it is not limited to the prediction of antimicrobial activities as it can easily be used also for prediction of other numeric properties of peptides.

IV. Antimicrobial Activity Prediction

Our approach exploits structure prediction methods and techniques of relational machine learning in conjunction with state-of-the-art attribute-value learning algorithms. Very briefly, our method can be imagined as proceeding in four steps. It starts with AMP sequences, for which we obtain spatial models using LOMETS structure prediction software [17] (*step 1*). This gives us 3D information in PDB files. Then we create a relational representation of the peptides (*step 2*). After that we use our relational learning algorithm RelF [3] to extract meaningful relational patterns from the relational structures describing peptides and convert them to an approximate attribute-value representation of the peptides (*step 3*). The output of RelF - *.arff* file readable by WEKA [18] is then used for learning regression models (*step 4*).

In the first step, 3D structures of peptides are computed using LOMETS software. LOMETS combines results of several threading-based structure prediction algorithms and returns several models with predicted coordinates of alpha carbon atoms. We use only the best full-length model according to ordering given by LOMETS for each sequence.

In the second step, we create a representation of peptides' spatial structures suitable for relational learning. A *literal* is an expression of the form $literalName(A_1, \ldots, A_k)$ where $A_1, \ldots, A_2$ are variables or constants. We use the convention that variables start with an upper-case letter. For example $residue(A, his)$ or $dist(A, B, 10)$ are literals and A, B are variables whereas his and 10 are constants. An *example* is simply a set of literals none of which contains a variable. For instance

$$e_1 = \begin{array}{l} residue(a, glu), residue(b, cys), \\ dist(a, b, 4), dist(b, a, 4) \end{array}$$

is an example describing a peptide.

Besides examples, we also need *patterns*. A pattern is a set of literals which, unlike examples, may contain variables. An example of a pattern is

$$p_1 = residue(A, X), dist(A, B, 10), residue(B, glu)$$

A pattern p is said to *cover* an example e when we are able to find a substitution θ to variables of p such that $p\theta \subseteq e$. For example the pattern p_1 covers the example e_1 because $p_1\theta \subseteq e$ for substitution $\theta = \{A/b, B/a\}$. We are not interested only whether a pattern p covers a given example e but also how many *covering substitutions* there are, i.e. how many substitutions θ such that $p\theta \subseteq e$ there are. We call the number of covering substitutions of a pattern p its *value*.

We use a representation of peptides that consists of literals representing types of the residues and literals representing pair-wise distances between the residues up to 10 Å. These distances are computed from alpha carbon atom coordinates obtained from PDB files generated by LOMETS. The data transformation process is shown in Figure 1.

In the third step, we use RelF to construct a set of meaningful structural patterns. RelF restricts the shape of possible patterns to be tree-like, because the complexity of pattern-set construction is generally lower for tree-like patterns than for general patterns and it has been shown that tree-like patterns are sufficiently rich for proteomics problems [4]. We customized the pattern search algorithm RelF. The original pattern search algorithm prunes pattern space using two measures: *redundancy* (described in [3]) and *minimum frequency* which is a minimum number of examples that must be covered by a pattern. Since RelF had been designed for classification problems, we had to find a way to use it for regression problems. We decided to follow

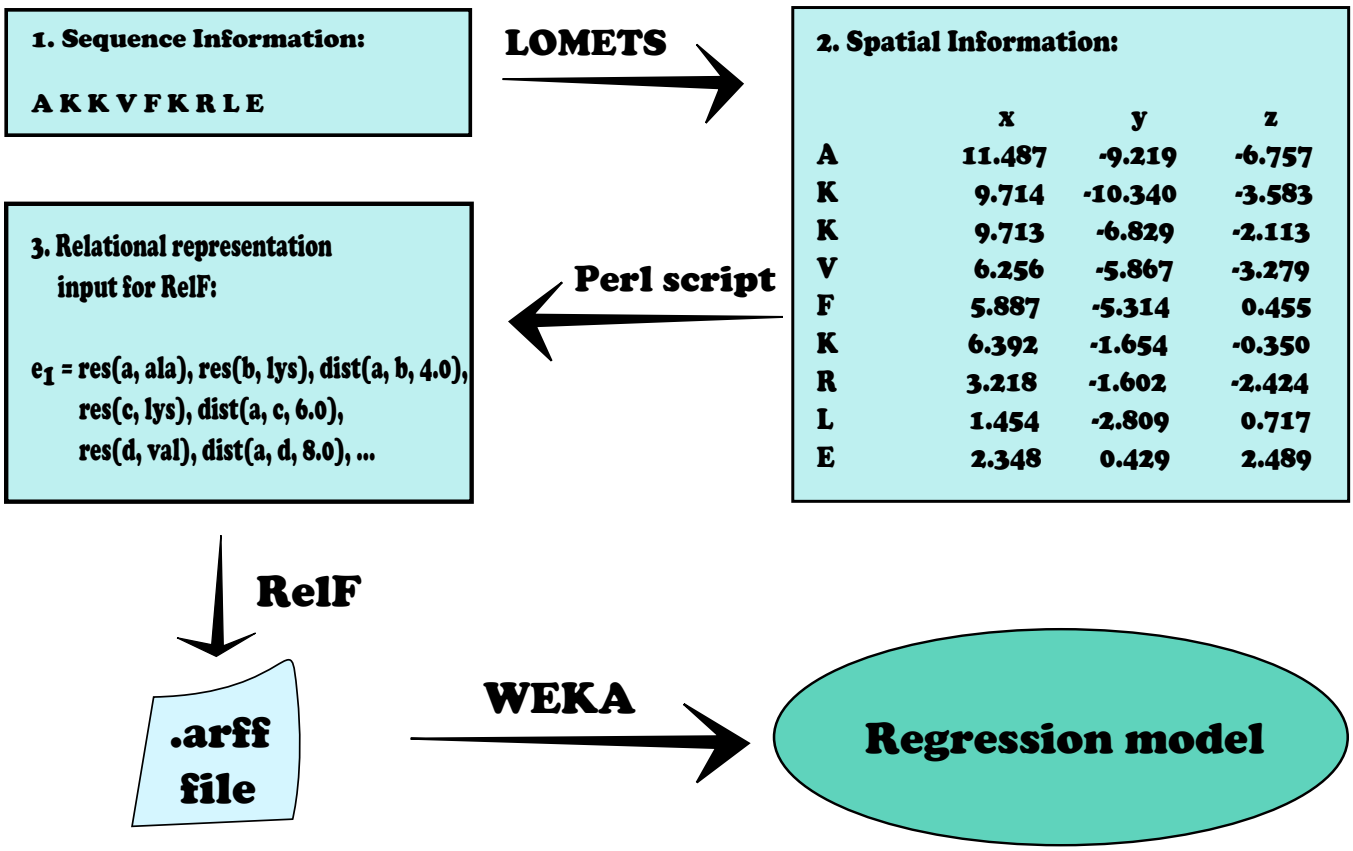

Figure 1. The main steps of the method.

a straightforward approach. We enriched RelF with preprocessing in which the training data are split into two sets[1] according to antimicrobial activity - the first set containing peptides with lower-than-median activities, the second set containing peptides with higher-than-median activities. As soon as we have a data set with at least two classes, RelF can be used for construction of discriminative features. The output of RelF is an attribute-value representation in WEKA format. We also added to these files additional information about dipole moment, proportions of amino acid types and their spatial asymmetries [19] which proved to be useful when added to relational patterns [4]. Once we had these WEKA files, we could easily exploit implementations of machine learning algorithms present in WEKA.

In the last step, we use implementation of SVM with RBF kernel present in the WEKA open-source machine learning software to train a regression model using the files generated in *step 3*. Parameters of the regression model are tuned using internal cross-validation. When performing cross-validation, the set of patterns is created separately for each train-test split corresponding to iterations of the cross-validation procedure.

[1]When performing cross-validation, we always split the data taking into account only the training set to avoid information leakage into the independent test set.

V. Data

We used three data sets to evaluate our novel method. The first data set named CAMEL was described by Cherkasov et al. [1]. It is composed of 101 antimicrobial peptides with experimentally tested antimicrobial potency. These peptides are rich in leucine and it has been demonstrated that they exhibit high activity against various strains of bacteria. The minimal inhibitory concentrations for these peptides have been averaged over 13 microorganisms. The average minimal inhibitory concentrations (MIC) were used to calculate average potencies according to formula from [20]

$$Potency = \log_2 \frac{1066}{MIC}.$$

The second data set named RANDOM was presented by Fjell et al. [16]. It contains 200 peptides with fixed length which are composed of a few amino acids (TRP, ARG and LYS and, more limitedly, LEU, VAL and ILE). Although antimicrobial peptides are actually enriched in these residues, a wide diversity in the amino acid content can be found in natural antimicrobial peptides [21]. The peptides were assayed for antimicrobial activity using a strain of

Pseudomonas aeruginosa. Fjell et al. did not report absolute MIC values, but only MIC values divided by MIC of Bac2A peptide (to simplify the measurements). Using relative MIC values poses no problem, because it manifests itself only through addition of a constant to the potency values (due to the logarithm).

We named the last data set BEE. We compiled it from three different sources: peptides from the venom of the eusocial bee Halictus sexcinctus and their analogs [22], peptides from the venom of the eusocial bee Lasioglossum laticeps [23] and peptides from the venom of the cleptoparasitic bee Melecta albifrons [24]. They contain peptides of length ca. 5 - 15 amino acids. The minimal inhibitory concentrations for these peptides were obtained for *Bacillus subtilis, Escherichia coli, Staphylococcus aureus, Pseudomonas aeruginosa*. We used the average of these values following the methodology of previous works [1], [16]. In some cases, when only lower bounds on MIC were available, we used these values.

VI. Results

In this section we present experiments performed on real-life data described in Section V. We used a representation of peptides that consisted of literals representing types of the amino acids and literals representing pair-wise distances between the amino acids up to 10 Å. These distances were computed from alpha-carbon coordinates obtained from PDB files computed by LOMETS. We used discretisation of distances with discretisation step 2 Å. We trained support vector machine [25] regression models with RBF kernel selecting optimal C (complexity constant) and $gamma$ (determines the kernel width parameter) for each fold by internal cross-validation. The estimated results are shown in Table I.

We performed experiments on three data sets (CAMEL, RANDOM and BEE). We compared the results of our relational learning method for regression with the results reported by Torrent et al. which is a state-of-the-art method. In [2] by Torrent et al., only cross-validated coefficients of determination (q^2 - see Appendix for definition) were given. Coefficient of determination can be regarded as the proportion of variability in a data set that is accounted for by the statistical model. In addition, we also report correlation coefficient (q) and root-mean-square error ($RMSE$) for our regression method. On data set CAMEL we achieved the same results as Torrent et al. On data set RANDOM we improved upon the results of Torrent et al. in terms of coefficients of determination. Since data set BEE is a newly compiled data set, there are no results to compare our approach with. It is also a harder data set, than the other two, because it is composed of three different sources. Each of these sources is homogeneous on their own, but heterogeneous when joined into one big data set. Also the variance of antimicrobial activity is lower in this data set than in the other two. This explains why the coefficient of determination is so small as compared to the coefficients of determination obtained for the other data sets.

A problem of antimicrobial peptides as antibiotics is that they often have the ability to lyse eukaryotic cells, which is commonly expressed as hemolytic activity or toxicity to red blood cells. Unlike the other methods which use a pre-fixed set of physicochemical features our method is not limited to one particular task. Since the sources from which we compiled the data set BEE contained also information about the hemolytic activity, we decided to assess the potential of our method also for prediction of hemolytic activity. Because more than half of the reported hemolytic activities were given only by an lower-bound (200 μM) (i.e. they were not capable to measure the exact value), we decided to transform the problem to a two-class classification problem - the first class corresponding to peptides with activities below the lower-bound, the second class corresponding to peptides with activities higher than the lower-bound. We performed experiments following the same steps as in the prediction of antimicrobial activity, but with a random forest classifier instead of support vector machine classifier for regression. We obtained accuracy 60.83% and AUC (area under ROC curve) 0.725.

In addition, we can analyse the structural patterns used in the regression model which can give us insights about the process by which the antimicrobial peptides kill bacteria. We used the following methodology. First, we discretized the antimicrobial activity attribute, so that we could apply χ^2 criterion for ranking of patterns. Then, for each split of the datasets (CAMEL, RANDOM and BEE) induced by 10-fold cross-validation we selected the three most informative structural patterns according to the χ^2 criterion. We chose one pattern which was selected most often among the folds for each data set. These patterns are shown in Figure 2.

The selected pattern for the data set CAMEL assumes presence of five amino acids: ILE, LEU, 2×LYS, VAL with distances between them as depicted in Figure 2. The positively charged Lysines are known to correlate with antimicrobial activity and the presence of Leucine can be explained by the fact that the data set CAMEL contains mostly Leucine-rich peptides. Interestingly, the remaining two amino acids - Isoleucine and Valine - and Leucine are the only proteinogenic branched-chain amino acids - they each have a carbon chain that branches off from the amino acid's main chain, or backbone.

The selected pattern for the data set RANDOM is very simple. It assumes presence of Tryptophan. Since the patterns count the number of occurrences, it corresponds to proportion of TRP in peptides. This is not surprising, given that the peptides of the data set RANDOM are composed mostly of TRP and some other amino acids.

Finally, the selected pattern for the data set BEE assumes presence of two amino acids: LEU and LYS in the distance

	Torrent et al. [2]	Our Regression Model		
	q^2	q^2	q	$RMSE$
CAMEL	**0.65**	**0.65**	0.81	1.23
RANDOM	0.72	**0.74**	0.87	1.23
BEE	-	0.3	0.61	1.04

Table I
EXPERIMENTAL RESULTS OBTAINED BY CROSS-VALIDATION, WHERE q^2 IS COEFFICIENT OF DETERMINATION, q IS CORRELATION COEFFICIENT AND $RMSE$ IS ROOT-MEAN-SQUARE ERROR.

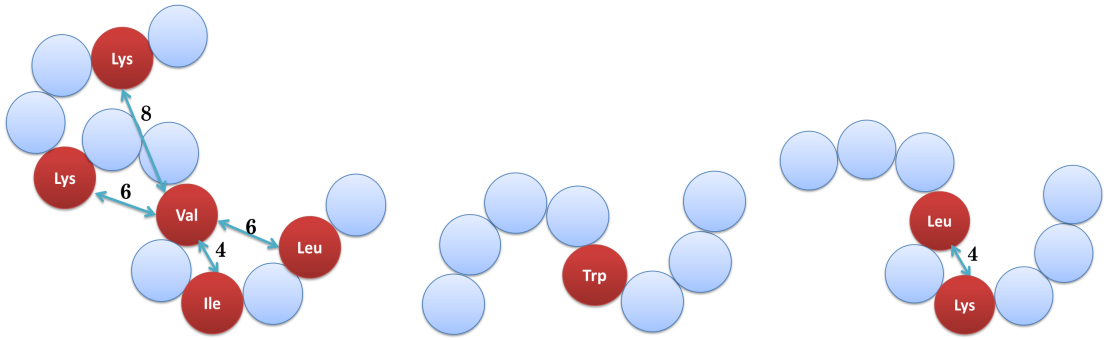

Figure 2. Most informative structural patterns according to the χ^2 criterion for the data set of CAMEL (A), RANDOM (B) and BEE (C)(edges not to scale).

$4\mathring{A}$ from each other. Again, the positively charged amino acid - Lysine is known to correlate well with antimicrobial activity. Both Leucine and Lysine appeared also in the selected pattern for the data set CAMEL.

VII. CONCLUSIONS

We applied relational machine learning techniques to predict antimicrobial activity of peptides. To our best knowledge this study is the first attempt to automatically discover common structural patterns present in antimicrobial peptides and to use them for prediction of antimicrobial activity. We utilized our relational learning method [3] which has already been used for DNA-binding propensity prediction of proteins [4]. There are two main differences between the work presented in this paper and our earlier work. First, the problem that we tackled in [4] dealt with classification, whereas here we built a regression model. Second, here only primary structures of peptides are available (therefore we had to rely on structure prediction), whereas we could use spatial structures obtained by X-ray crystallography in our previous study with DNA-binding proteins. We have shown that our relational learning approach for regression improves on a state-of-the-art approach to antimicrobial activity prediction in terms of predictive accuracy. Moreover, we have illustrated that our method is capable to also provide interpretable patterns describing spatial configurations of amino acids in peptide structures.

ACKNOWLEDGMENT

This work has been supported by the Czech Science Foundation through project Predicting Protein Properties with Spatial Statistical Relational Machine Learning (P202/12/2032).

APPENDIX

Coefficient of determination

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}$$

Here, SS_{err} is the sum of squares of residuals and SS_{tot} is the total sum of squares.

$$SS_{err} = \sum_i (y_i - f_i)^2,$$

where y_i is the true value and f_i is the predicted value.

$$SS_{tot} = \sum_i (y_i - \overline{y})^2,$$

where

$$\overline{y} = \frac{1}{n} \sum_i y_i.$$

The coefficient of determination q^2 is an estimate of R^2 obtained by cross-validation.

Correlation coefficient

$$R = \frac{1}{n-1} \sum_{i}^{n} \left(\frac{y_i - \overline{y}}{s_y} \right) \left(\frac{f_i - \overline{f}}{s_f} \right)$$

Here, s_y is standard deviation of true values and s_f is standard deviation of predicted values. The correlation coefficient q is an estimate of R obtained by cross-validation.

REFERENCES

[1] A. Cherkasov and B. Jankovic, "Application of 'inductive' qsar descriptors for quantification of antibacterial activity of cationic polypeptides," *Molecules*, vol. 9, no. 12, pp. 1034–1052, 2004.

[2] M. Torrent, D. Andreu, V. Nogués, and E. Boix, "Connecting peptide physicochemical and antimicrobial properties by a rational prediction model," *PLoS ONE*, vol. 6, 02 2011.

[3] O. Kuželka and F. Železný, "Block-wise construction of tree-like relational features with monotone reducibility and redundancy," *Machine Learning*, vol. 83, pp. 163–192, 2011.

[4] A. Szabóová, O. Kuželka, F. Železný, and J. Tolar, "Prediction of dna-binding proteins from structural features," in *MLSB 2010: 4th International Workshop on Machine Learning in Systems Biology*, 2010, pp. 71–74.

[5] H. Nassif, H. Al-Ali, S. Khuri, W. Keirouz, and D. Page, "An Inductive Logic Programming approach to validate hexose biochemical knowledge," in *Proceedings of the 19th International Conference on ILP*, Leuven, Belgium, 2009, pp. 149–165.

[6] B. Peters, M. Shirtliff, and M. Jabra-Rizk, "Antimicrobial peptides: Primeval molecules or future drugs?" *PLoS Pathogens*, vol. 6, 10 2010.

[7] M. Yeaman and N. Yount, "Mechanisms of antimicrobial peptide action and resistance," *Pharmacological Reviews*, vol. 55, no. 1, pp. 27–55, 2003.

[8] R. Hancock and A. Rozek, "Role of membranes in the activities of antimicrobial cationic peptides," *FEMS Microbiology Letters*, vol. 206, no. 2, pp. 143–149, 2002.

[9] S. Lata, N. Mishra, and G. Raghava, "Antibp2: improved version of antibacterial peptide prediction," *BMC Bioinformatics*, vol. 11, 2010.

[10] M. Torrent, V. Nogués, and E. Boix, "A theoretical approach to spot active regions in antimicrobial proteins," *BMC Bioinformatics*, vol. 10, no. 1, 2009.

[11] H. Jenssen, T. Lejon, K. Hilpert, C. Fjell, A. Cherkasov, and R. Hancock, "Evaluating different descriptors for model design of antimicrobial peptides with enhanced activity toward p. aeruginosa," *Chemical Biology & Drug Design*, vol. 70, no. 2, 2007.

[12] H. Jenssen, C. Fjell, A. Cherkasov, and R. Hancock, "Qsar modeling and computer-aided design of antimicrobial peptides," *Journal of Peptide Science*, vol. 14, no. 1, 2008.

[13] V. Frecer, "Qsar analysis of antimicrobial and haemolytic effects of cyclic cationic antimicrobial peptides derived from protegrin-1," *Bioorganic & Medicinal Chemistry*, vol. 14, no. 17, pp. 6065 – 6074, 2006.

[14] S. Thomas, S. Karnik, R. Barai, V. Jayaraman, and S. Idicula-Thomas, "Camp: a useful resource for research on antimicrobial peptides," *Nucleic Acids Research*, vol. 38, pp. D774–D780, 2010.

[15] O. Taboureau, O. Olsen, J. Nielsen, D. Raventos, P. Mygind, and H. Kristensen, "Design of novispirin antimicrobial peptides by quantitative structureactivity relationship," *Chemical Biology & Drug Design*, vol. 68, no. 1, pp. 48–57, 2006.

[16] C. Fjell, H. Jenssen, K. Hilpert, W. Cheung, N. Pante, R. Hancock, and A. Cherkasov, "Identification of novel antibacterial peptides by chemoinformatics and machine learning," *Journal of Medicinal Chemistry*, vol. 52, no. 7, pp. 2006–2015, 2009.

[17] S. Wu and Y. Zhang, "Lomets: A local meta-threading-server for protein structure prediction," *Nucleic Acids Research*, vol. 35, no. 10, pp. 3375–3382, 2007.

[18] I. Witten and E. Frank, *Data Mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, San Francisco, 2005.

[19] A. Szilágyi and J. Skolnick, "Efficient prediction of nucleic acid binding function from low-resolution protein structures," *Journal of Molecular Biology*, vol. 358, no. 3, pp. 922–933, 2006.

[20] R. Mee, T. Auton, and P. Morgan, "Design of active analogues of a 15-residue peptide using d-optimal design, qsar and a combinatorial search algorithm," *The Journal of Peptide Research*, vol. 49, no. 1, 1997.

[21] P. Bulet, R. Stöcklin, and L. Menin, "Anti-microbial peptides: from invertebrates to vertebrates," *Immunological Reviews*, vol. 198, no. 1, 2004.

[22] L. Monincová, M. Buděšínský, J. Slaninová, O. Hovorka, J. Cvačka, Z. Voburka, V. Fučík, L. Borovičková, L. Bednárová, J. Straka, and V. Čeřovský, "Novel antimicrobial peptides from the venom of the eusocial bee halictus sexcinctus (hymenoptera: Halictidae) and their analogs," *Amino Acids*, vol. 39, pp. 763–775, 2010.

[23] V. Čeřovský, M. Buděšínský, O. Hovorka, J. Cvačka, Z. Voburka, J. Slaninová, L. Borovičková, V. Fučík, L. Bednárová, I. Votruba, and J. Straka, "Lasioglossins: Three novel antimicrobial peptides from the venom of the eusocial bee lasioglossum laticeps (hymenoptera: Halictidae)," *ChemBioChem*, vol. 10, no. 12, pp. 2089–2099, 2009.

[24] V. Čeřovský, O. Hovorka, J. Cvačka, Z. Voburka, L. Bednárová, L. Borovičková, J. Slaninová, and V. Fučík, "Melectin: A novel antimicrobial peptide from the venom of the cleptoparasitic bee melecta albifrons," *ChemBioChem*, vol. 9, no. 17, pp. 2815–2821, 2008.

[25] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.