МАТЕМАТИЧЕСКАЯ БИОЛОГИЯ, БИОИНФОРМАТИКА / MATHEMATICAL BIOLOGY, BIOINFORMATICS

DOI: https://doi.org/10.60797/jbg.2024.26.5

MACHINE LEARNING USING MULTIPLE LOGISTIC REGRESSION FOR ANTIMICROBIAL AND HEMOLYTIC PEPTIDES PREDICTION AND THEIR IDENTIFICATION IN LARGE PROTEINS

Research article

Krenev I.A.[1, *]
[1] ORCID : 0000-0001-7970-3291;
[1] Institute of Experimental Medicine, Saint-Petersburg, Russian Federation

* Corresponding author (il.krenevv13[at]yandex.ru)

Abstract

Antimicrobial peptides (AMPs) are considered as a promising pool of alternative antimicrobial agents in the post-antibiotic era. Since a number of limitations, especially cytotoxicity, restrict their implementation into clinic, search for novel non-toxic AMPs is of high relevance. In the present study, we used multiple logistic regression for prediction of both antimicrobial and hemolytic capacities of peptides. The two constructed models demonstrated acceptable predictive power (at estimated optimal cut-offs, accuracy, sensitivity, specificity, F-measure ≥ 0.82, ROC AUC > 0.91). The model for antimicrobial activity prediction was further applied for identification of possible AMPs in large protein sequences. The validation of the method was performed on precursors of well-known AMPs from different structural classes – human neutrophil peptide 1 (HNP1), LL-37 cathelicidin as well as of tachyplesin I. In all cases, the mature AMPs localization was predicted correctly, i.e. at the C-terminus (HNP1, LL-37) or in the middle of the precursor sequence (tachyplesin I). The study provides the easy-for-interpretation method for prediction of antimicrobial and hemolytic peptides and their identification in large proteins.

Keywords: antimicrobial peptides, alternative antibiotics, peptides, hemolytic peptides, machine learning.

МАШИННОЕ ОБУЧЕНИЕ С ПРИМЕНЕНИЕМ МНОЖЕСТВЕННОЙ ЛОГИСТИЧЕСКОЙ РЕГРЕССИИ ДЛЯ ПРЕДСКАЗАНИЯ АНТИМИКРОБНЫХ И ГЕМОЛИТИЧЕСКИХ ПЕПТИДОВ И ИХ ОБНАРУЖЕНИЯ В КРУПНЫХ БЕЛКАХ

Научная статья

Кренев И.А.[1, *]
[1] ORCID : 0000-0001-7970-3291;
[1] Институт экспериментальной медицины, Санкт-Петербург, Российская Федерация

* Корреспондирующий автор (il.krenevv13[at]yandex.ru)

Аннотация

Антимикробные пептиды (АМП) рассматриваются в качестве перспективного пула альтернативных антимикробных агентов в пост-антибиотическую эру. Поскольку ряд ограничений, в особенности цитотоксичность, лимитирует их имплементацию в клинику, поиск новых нетоксичных АМП имеет высокую актуальность. В настоящем исследовании мы применили множественную логистическую регрессию для предсказания антимикробной и гемолитической способности пептидов. Две построенные модели продемонстрировали приемлемую предсказательную силу (аккуратность, чувствительность, специфичность, точность, F-мера ≥ 0,82, ROC AUC > 0,91 при оцененных оптимальных пороговых уровнях). Модель для предсказания антимикробной активности была далее применена для обнаружения возможных АМП в последовательностях крупных белков. Валидация метода была осуществлена на предшественниках известных АМП различных структурных классов – нейтрофильный пептид человека 1 (HNP1), кателицидин LL-37, а также тахиплезин I. Во всех случаях локализация зрелых АМП была предсказана верно, т.е. на С-конце (HNP1, LL-37) или в середине последовательности предшественника (тахиплезин I). Исследование обеспечивает простой для интерпретации метод предсказания антимикробной и гемолитической активности пептидов и их идентификации в крупных белках.

Ключевые слова: антимикробные пептиды, альтернативные антибиотики, пептиды, гемолитические пептиды, машинное обучение.

Introduction

The antibiotic resistance crisis is considered as a major challenge for modern medicine and raises questions of obtaining alternative antimicrobials [1], [2]. Antimicrobial peptides (AMPs) represent a group of promising candidates due to their high anti-pathogen activity in the low micromolar range of concentrations, a broad spectrum of targets and a low potential to provoke bacterial resistance [3], [4]. AMPs act as effector molecules of innate immunity of virtually all living organisms. Although AMPs are a highly diverse and difficult-for-classification group of biologically active peptides, most of them possess a common set of physicochemical properties. Typically, they are amphipathic peptides with spaced clusters of cationic and hydrophobic amino acid residues (a.a.). The primary mechanism of their action includes adsorption on negatively charged molecules of pathogenic surface due to electrostatic interactions and insertion into the lipid bilayer of the cell. The pore formation results in efflux of vital ions and metabolites and eventually to the target cell death, although alternative mechanisms of AMPs action have been described [5], [6].

The main and well-described AMPs of human innate immunity are cathelicidin LL-37 [7] and human neutrophil peptides (HNPs) [8].

In spite of indisputable advantages, AMPs have not yet become a commonly prescribed antimicrobial drugs. One of the recognized limitations of AMPs implementation is cytotoxicity, i.e. the action on host cell membranes. In most pipelines of AMPs elaboration, cytotoxicity is evaluated in hemolytic assays [9]. Clinically prospective AMPs are expected to have high selectivity of action.

In the context of the AMPs action as a double-edged sword, databases of antimicrobial and hemolytic peptides (HPs) have been created. The databases and existing predictive tools are reviewed in [10], [11], [12]; a more detailed review passage on tools for cytotoxic peptides' prediction can be found in a recent research paper [13]. Experimentally validated peptides' antimicrobial and hemolytic action is a tempting pool of data for machine learning (ML). Methods of AMPs prediction using ML include support-vector machine, hidden Markov models, random forest, artificial neuronal networks, decision tree, K-nearest neighbor, linear discriminant analysis, in addition to others. While quantitative models intend to predict minimum inhibitory concentration (MIC) or analogous numeric characteristics of a peptide's activity, qualitative models are binary-type tools distinguishing between AMPs and non-AMPs or HPs and non-HPs. The qualitative approach seems to have advantages since protocols of antimicrobial and hemolytic activity measurement are not strictly standardized, and exact active concentrations can vary greatly depending on experimental conditions, a specific target, i.e. microbial strain or a source of red blood cells, etc. Logistic regression (LR) is not a widely spread method of prediction among available servers, but it is easy for use and interpretation and has been used for AMPs prediction [14].

Search of novel AMPs may build on known sequences of large proteins. Some of them are precursors of well-described AMPs typically containing a signal peptide at the N-terminus and a fragment corresponding to a mature AMP (in addition, other domains can be present). Nevertheless, peptides with antimicrobial activity may be derived from large proteins without described direct antimicrobial action. This may occur under physiological proteolysis, as it was demonstrated for anaphylatoxins, i.e. the derivatives of complement proteins C3, C4, C5 [15], [16], [17], [18], and α-melanocyte stimulating hormone, i.e. the proopiomelanocortin derivative [19]. Identification of AMPs in large proteins utilizes the concept of a sliding window to screen the whole input sequence. Such type of peptides search is available via some web-servers. However, they have some disadvantages. The AMPA server [20] is adapted for narrow sliding windows and, in our experience, does not predict the location of actual active fragments even in such classic AMPs as human defensins and cathelicidin LL-37-related peptides. Another server, Antifp [21] is restricted to the prediction of antifungal peptides. Moreover, to our knowledge, LR has never been applied for revealing promising fragments in proteins sequences.

Since we conclude that there is a lack of easy-for-interpretation and validated for a broad spectrum of peptides models for AMPs and HPs prediction, including that from long sequences, we aimed to elaborate a new approach based on LR.

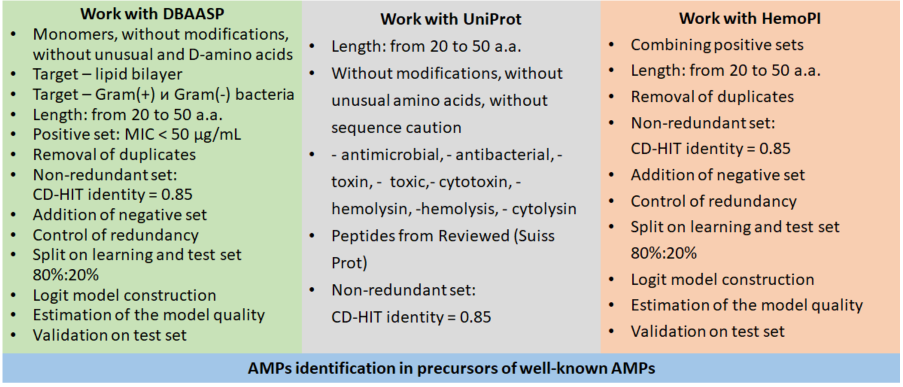
1.1. Abbreviations

a.a. – Amino acid residue; AAC – mino-acid composition; AMP – antimicrobial peptide; AUC – area under curve; df – degrees of freedom; HP – hemolytic peptide; LR – logistic regression; MIC – minimum inhibitory concentration; MCC – Matthews correlation coefficient; ML – machine learning; ROC – receiver operating characteristic; VIF – variance inflation factor

Research methods and principles

2.1. The general algorithm of the study

The general algorithm of the work is represented in the Fig. 1. The study design includes two consecutive parts: ML and AMPs identification in large protein sequences. ML was performed to build models for both antimicrobial and hemolytic activity prediction. Positive and negative datasets were filtered, used in training and validation in order to construct predictive models. The model for AMPs prediction was applied for identification of promising sequences of well-known AMPs precursors.



Work with DBAASP

- Monomers, without modifications, without unusual and D-amino acids
- Target – lipid bilayer
- Target – Gram(+) и Gram(-) bacteria
- Length: from 20 to 50 a.a.
- Positive set: MIC < 50 µg/mL
- Removal of duplicates
- Non-redundant set: CD-HIT identity = 0.85
- Addition of negative set
- Control of redundancy
- Split on learning and test set 80%:20%
- Logit model construction
- Estimation of the model quality
- Validation on test set

Work with UniProt

- Length: from 20 to 50 a.a.
- Without modifications, without unusual amino acids, without sequence caution
- - antimicrobial, - antibacterial, - toxin, - toxic,- cytotoxin, - hemolysin, -hemolysis, - cytolysin
- Peptides from Reviewed (Suiss Prot)
- Non-redundant set: CD-HIT identity = 0.85

Work with HemoPI

- Combining positive sets
- Length: from 20 to 50 a.a.
- Removal of duplicates
- Non-redundant set: CD-HIT identity = 0.85
- Addition of negative set
- Control of redundancy
- Split on learning and test set 80%:20%
- Logit model construction
- Estimation of the model quality
- Validation on test set

AMPs identification in precursors of well-known AMPs

Figure 1 - The general algorithm of the study
DOI: https://doi.org/10.60797/jbg.2024.26.5.1

2.2. Machine learning

2.2.1. Antimicrobial activity: positive dataset

DBAASP (the Database of Antimicrobial Activity and Structure of Peptides) was used as a source of positive dataset containing naturally occurring and artificial AMPs [22]. The database was filtered as follows: only monomeric peptides without modifications, unusual and D-amino acids, without "X" as an a.a., having MIC as described target activity value and

MIC < 50 μg/mL, acting on lipid bilayer of gram-positive and/or gram-negative bacteria. The length from 20 to 50 was considered optimal. Too short and too long peptides were avoided for some reason. For example, amino-acid composition (AAC) value was intended to be used in ML (see Section 2.2.4) but a single a.a. in a very short peptide would give a very high AAC value and, at the same time, too many a.a. would have AAC value equal to 0. On the other hand, too long sequences are not prospective for synthesis and implementation. Besides, the use of sequences in a very wide window of lengths would make the dataset less uniform. Among duplicates (i.e., rows corresponding to one peptide with numerous target species) only one representative was retained while other representatives were excluded. Importantly, they were removed after MIC filtration was performed, so all active peptides were preserved in the positive set.

The data redundancy problem was solved by using CD-HIT, which is a widely used algorithm for sequences clustering and comparison [23], [24], [25]. Instead of a standard program, the analogous cdhit() function from the CellaRepertorium R package was used. Before that, sequences were transformed to the AAString object format using the Biostrings package. Identity level 0.85 was used to generate a non-redundant dataset.

2.2.2. Hemolytic activity: positive dataset

The positive set of HPs was the assembled by combining positive datasets from the HAPPENN database [26]: HemoPI-1 main positive, HemoPI-1 validation positive, HemoPI-2 main positive, HemoPI-2 validation positive, HemoPI-3 main positive, HemoPI-3 validation positive. Peptides of the length 20-50 were retained. CH-HIT was used to remove similar peptides (identity 0.85).

2.2.3. Negative dataset

The negative dataset was extracted from UniProt, which is a comprehensive database containing polypeptide sequences with functional information [27]. Similarly to the positive dataset of AMPs, only 20-50-a.a peptides without modifications, unusual a.a., sequence caution were selected. Entries with keywords “antimicrobial”, “antibacterial”, “toxin”, “toxic”, “cytotoxin”, “hemolysin”, “hemolysis”, “cytolysin” were excluded. Then the reviewed database (SwissProt) was used for the final negative set formation. Duplicates and redundant peptides were excluded as described above. The negative dataset was used in the both antimicrobial and hemolytic ML models.

2.2.4. Performance of machine learning

In the both models, sizes of positive and negative datasets were equal, i.e. the datasets were balanced. This was reached by random exclusion of sequences from the negative dataset. For the both ML pipelines, CD-HIT was used to control possible similarities between positive and negative sets to ensure that they were different enough with the identity cut-off 0.85. Before the performance of ML, all datasets were randomized. All datasets were split to obtain a training set (80%) and test set (20%).

For all peptides, a number of properties was calculated. Peptides R package was used to calculate net charge at pH 7.0, hydrophobicity (scale Fasman was selected in preliminary models as leading to the best model performance), hydrophobic moment at the angle $\vartheta = 96°$ [28]. Amino-acid composition (AAC) was calculated using protr package.

LR was used to estimate probability of antimicrobial or hemolytic capacity. LR is a statistical method applying sigmoid function operating with probabilities to linearize the data. Logit function of probability of a positive outcome $p = Pr(Y = 1|x_1,\ x_2, \ldots,\ x_n)$ is calculated as intercept plus linear combination of predictors multiplied by their coefficients β_i (weights):

$$z = logit(p) = ln\frac{p}{1-p} = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$$

Therefore,

$$p = \frac{1}{1+e^{-z}}$$

The optimal cut-off can be used to predict the outcome in a binary manner.

Features selection was a critical step in the modelling. It can be said that no rigorous algorithms were used to make a final list of predictors. However, several rules were used to construct models of acceptable quality. Since LR belongs to the class of generalized linear models, it is sensitive to multicollinearity. Correlated features were excluded in such a way to preserve as more as possible features. Methionine residues were excluded from the final list of the features since methionine represents the first a.a. in immature peptides from the negative sets downloaded from UniProt; inclusion of methionine residues would produce bias towards prediction of polypeptides with preserved signal peptides regardless their actual biological activity. ML using LR was performed using R function glm(). The most significant predictors were selected. After features selection, variance inflation factor (VIF) was calculated to finally make sure that no relations are present in the list of predictors. In the both models, the number of predictors was considerably less than the number of observations in both positive and negative groups. Besides, the models' fitness was estimated by χ^2:

$$\chi^2 \ = \ Null\ deviance \ - \ Residual\ deviance,$$

and p-value was calculated for degrees of freedom (df) equal to the difference between df in the null model (intercept-only model) and the model with predictors (in other words, to the number of predictors). Density plots were built to visualize the distribution of predictors between the positive and the negative group in training sets.

As the output of each of the models, formulae for calculation of LR probabilities (i.e., $p_{\text{Antimicrobial}}$ or $p_{\text{Hemolytic}}$) were obtained.

One way or another, predictive power rather than formal correctness was the main criterion of the models' quality. Models validation was performed on test sets, and standard parameters were calculated for estimation of the models quality using confusion matrices (TP – true positive, TN – true negative, FP – false positive, FN – false negative; MCC – Matthews correlation coefficient):

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

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

$$Specificity = \frac{TN}{TN+FP}$$

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

$$F - measure = \frac{2TP}{2TP+FP+FN}$$

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

Interpretation of these parameters can be found elsewhere. The parameters were calculated for cut-offs 0.35, 0.4, …, 0.85.

ROC (receiver operating characteristic) analysis was performed using pROC R package. Density plots were built to visualize distribution of predicted $p_{\mathrm{Antimicrobial}}$ or $p_{\mathrm{Hemolytic}}$ values by actual positive and negative groups.

2.3. Prediction of AMPs from precursor proteins

The model for AMPs prediction was applied for recognition of well-known AMPs in their precursor proteins. Mature peptides cathelicidin LL-37 (UniProt ID 49913), α-defensin HNP1 (UniProt ID P59665) comprise C-terminal parts of their precursor molecules. In contrast, tachyplesin I (UniProt ID P14213) is located in the middle part of its unprocessed precursor. The three peptides belong to different structural classes: LL-37 is α-helical, HNP1 contains 3 antiparallel β-strands, a disordered linker and is stabilized by three disulfide bonds; tachyplesin I is a relatively short β-hairpin. They also belong to different taxa – human and invertebrates, namely horseshoe crabs [29].

The concept of the sliding window was applied to screen a whole sequence (Fig. 2). Let L be the length of the peptide, L' is the length of the sliding window, l is the position of the rightmost (i.e., C-terminal) a.a. within the window at any step of the sequence screening. Therefore, the position of the leftmost (i.e., N-terminal) a.a. within the window is $l\text{-}L'+1$. Obviously, $L' \leqslant l \leqslant L$ and $l = Step + L' - 1$. Starting from $l=L'$ (step 1) the window is shifted by one position at every other step, and $L\text{-}L'+1$ steps are performed for each peptide. At any step, $p_{\mathrm{Antimicrobial}}$ was calculated, and $p_{Antimicrobial}$ values were plotted against $p_{Antimicrobial}$ values.

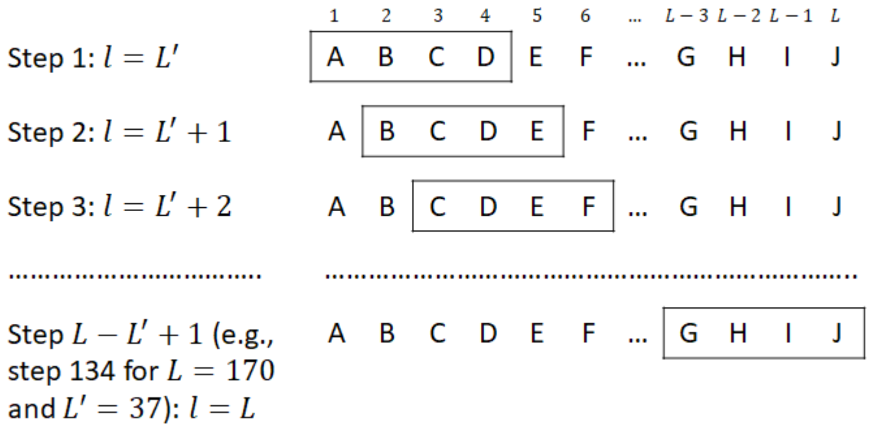


Figure 2 - The sliding window applied to the AMPs recognition in proteins
DOI: https://doi.org/10.60797/jbg.2024.26.5.2

Note: L is the length of the peptide, L' is the length of the sliding window, l is the position of the C-terminal a.a. within the window

2.4. Software

The study was performed using R language (v4.3.0) in the RStudio integrated development environment (2023.03.0) (R Core Team (2023)) [30]. The following R packages were used: Biostrings v2.68.1, car v3.1-2 [31], CellaRepertorium v1.10.0, dplyr v1.1.2 [32], ggplot2 v3.4.2 [33], ggpubr v0.6.0 [34], Peptides v2.4.5 [35], pROC v1.18.2 [36], protr v1.6-3 [37].

Main results

3.1. Results of machine learning

Data redundancy and distribution of cluster sizes is demonstrated in the Fig. 3. Although many peptides were quite distinct from others in each of the three datasets, numerous clusters of different sizes (up to 46 peptides) were found.

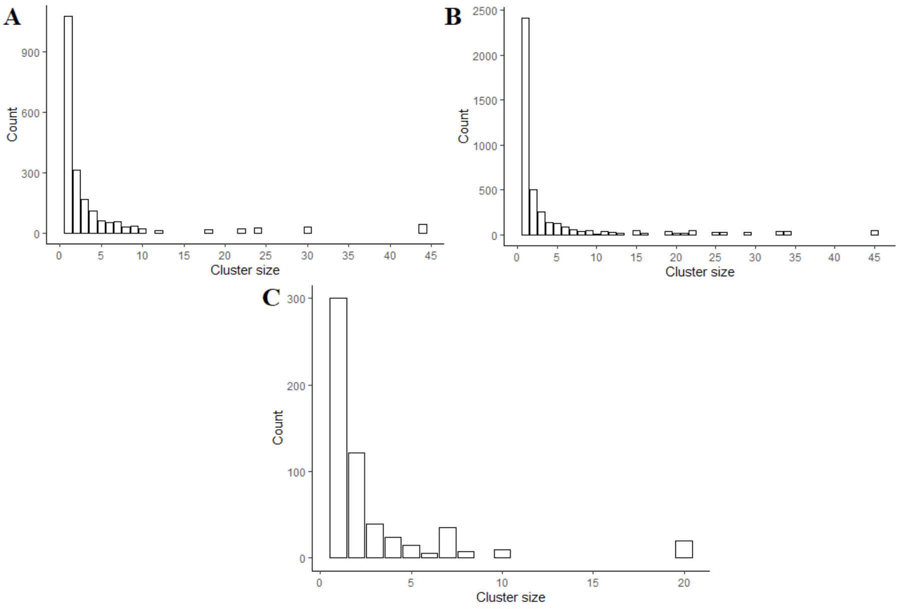


Figure 3 - Distribution of cluster sizes in redundant datasets:
A – Antimicrobial positive; *B* – Negative; *C* – Hemolytic positive
DOI: https://doi.org/10.60797/jbg.2024.26.5.3

Non-redundant datasets were obtained and used for ML. All used datasets can be found in supplementary files: test_set_ANTIMICROBIAL.xlsx, test_set_HEMOLYTIC. xlsx, training_set_ANTIMICROBIAL. xlsx, training_set_HEMOLYTIC. xlsx; all calculated features are represented in the files. The sizes of the sets are shown in Tables 1 and 2. 2670 peptides were present in the set for antimicrobial activity prediction and 780 peptides were used for hemolytic activity prediction. The datasets were balanced 1:1 (positive : negative) and split 80%:20% (training set : test set).

Table 1 - Sizes of sets used in ML for antimicrobial activity prediction

DOI: https://doi.org/10.60797/jbg.2024.26.5.4

Set	Split	Number of peptides
AMP (1335)	Training set: 80%	1068
	Test set: 20%	267
Non-AMP (1335)	Training set: 80%	1068
	Test set: 20%	267
	Training set: 80% (2136) Test set: 20% (534)	Sum: 2670

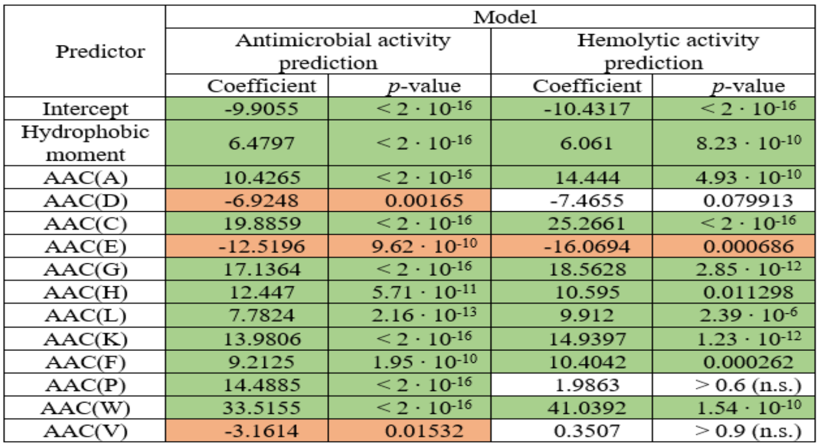
Table 2 - Sizes of sets used in ML for hemolytic activity prediction

DOI: https://doi.org/10.60797/jbg.2024.26.5.5

Set	Split	Number of peptides
Hemolytic (390)	Training set: 80%	312
	Test set: 20%	78
Non-hemolytic (390)	Training set: 80%	312
	Test set: 20%	78
	Training set: 80% (624) Test set: 20% (156)	Sum: 780

The results of prediction of both antimicrobial and hemolytic activity are represented in Table 3. Hydrophobic moment, proportions of alanine, cysteine, glycine, histidine, leucine, lysine, phenylalanine, proline, tryptophan produce positive effect on logit($p_{Antimicrobial}$); in contrast, aspartate, glutamate and valine produce negative effect. Essentially the same results were obtained in the hemolytic prediction model. In all cases, VIF was about 1 indicating the absence of significant multicollinearity. The expected results clearly demonstrate that the features predisposing peptides to antimicrobial and hemolytic activity are co-directed. The coefficients in the equations are in accordance with the distribution of the predictors by positive and negative groups (Fig. 4 and Fig. 5) suggesting that the equations are biologically relevant

Table 3 - Predictors used in the two models
DOI: https://doi.org/10.60797/jbg.2024.26.5.6



Predictor	Model			
	Antimicrobial activity prediction		Hemolytic activity prediction	
	Coefficient	*p*-value	Coefficient	*p*-value
Intercept	-9.9055	$< 2 \cdot 10^{-16}$	-10.4317	$< 2 \cdot 10^{-16}$
Hydrophobic moment	6.4797	$< 2 \cdot 10^{-16}$	6.061	$8.23 \cdot 10^{-10}$
AAC(A)	10.4265	$< 2 \cdot 10^{-16}$	14.444	$4.93 \cdot 10^{-10}$
AAC(D)	-6.9248	0.00165	-7.4655	0.079913
AAC(C)	19.8859	$< 2 \cdot 10^{-16}$	25.2661	$< 2 \cdot 10^{-16}$
AAC(E)	-12.5196	$9.62 \cdot 10^{-10}$	-16.0694	0.000686
AAC(G)	17.1364	$< 2 \cdot 10^{-16}$	18.5628	$2.85 \cdot 10^{-12}$
AAC(H)	12.447	$5.71 \cdot 10^{-11}$	10.595	0.011298
AAC(L)	7.7824	$2.16 \cdot 10^{-13}$	9.912	$2.39 \cdot 10^{-6}$
AAC(K)	13.9806	$< 2 \cdot 10^{-16}$	14.9397	$1.23 \cdot 10^{-12}$
AAC(F)	9.2125	$1.95 \cdot 10^{-10}$	10.4042	0.000262
AAC(P)	14.4885	$< 2 \cdot 10^{-16}$	1.9863	> 0.6 (n.s.)
AAC(W)	33.5155	$< 2 \cdot 10^{-16}$	41.0392	$1.54 \cdot 10^{-10}$
AAC(V)	-3.1614	0.01532	0.3507	> 0.9 (n.s.)

Note: green and red cells correspond to predictors with positive or negative effect on the dependent variable, respectively; AAC(X) – Amino-acid composition value of a residue X (name of an amino acid in standard one-letter code); n.s. – non-significant

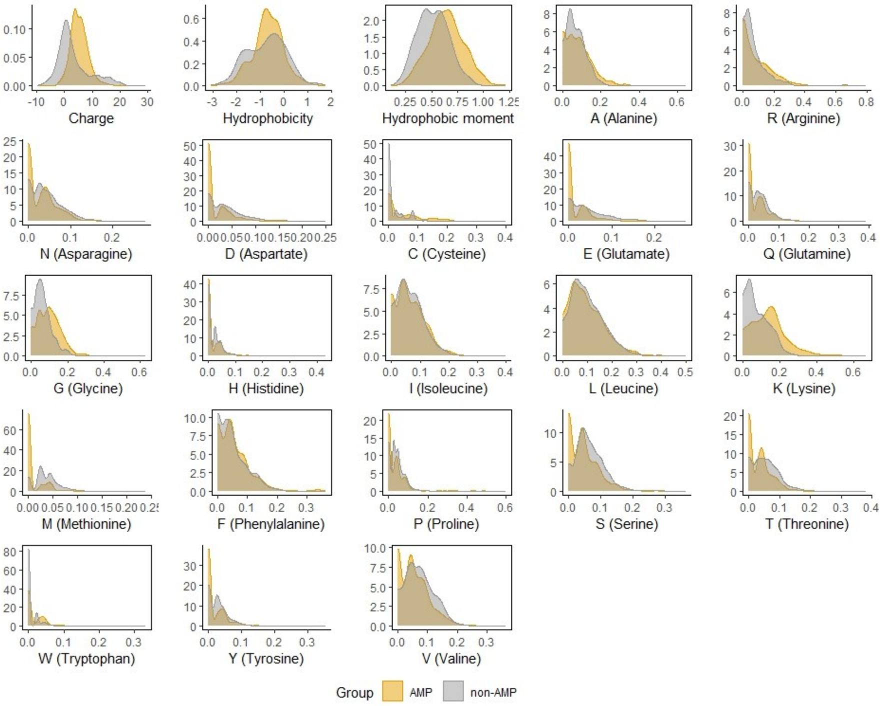


Figure 4 - Distribution of all predictors (features) by the groups of AMPs and non-AMPs in the training set

DOI: https://doi.org/10.60797/jbg.2024.26.5.7

Note: plots of the selected predictors are bordered

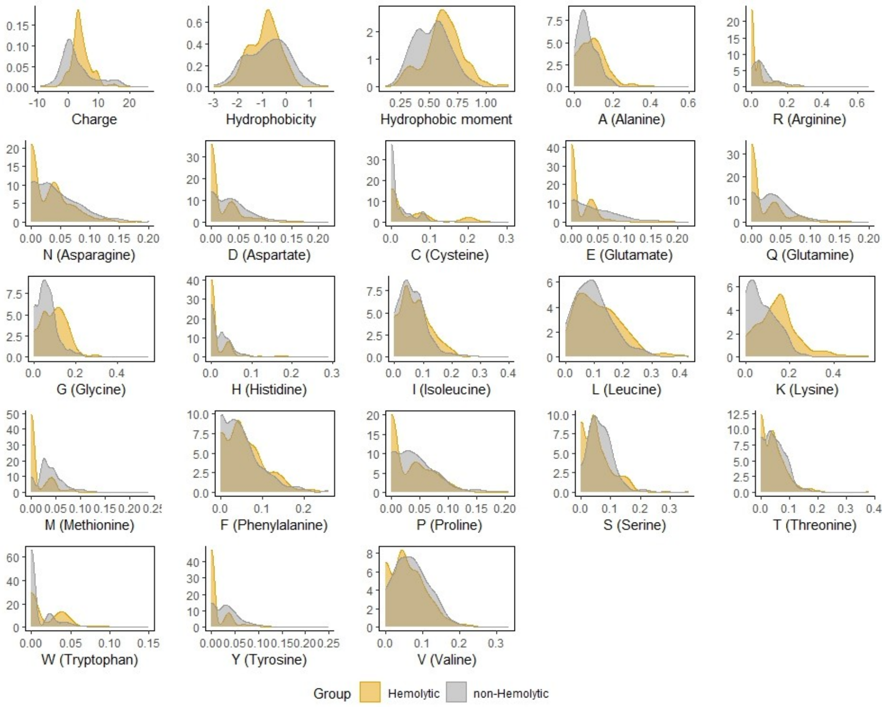


Figure 5 - Distribution of the all predictors (features) by the groups of hemolytic and non-peptides in the training set

DOI: https://doi.org/10.60797/jbg.2024.26.5.8

Note: plots of the selected predictors are bordered

Essential models parameters are represented in Table 4. It can be concluded the addition of 13 predictors to the intercept-only models significantly improves their fitness.

Table 4 - Essential parameters of the models

DOI: https://doi.org/10.60797/jbg.2024.26.5.9

Parameter	Model	
	Antimicrobial activity prediction	Hemolytic activity prediction
Null deviance (null model)	2961.1	865.05
df (null model)	2135	623
Residual deviance (model with predictors)	1610.6	418.09
df (model with predictors)	2122	610
χ^2	1350.5	449.96
df (number of predictors, i.e. df in the model with predictors – df in the null model)	13	13
p-value	<< 0.001	<< 0.001

Validation on test sets was performed. The estimated optimal (maximum accuracy) probability cut-off for antimicrobial activity prediction was 0.6; for hemolytic activity – 0.55. For antimicrobial activity prediction, accuracy = 0.86, sensitivity = 0.82, specificity = 0.90, precision = 0.91, F-measure = 0.86, MCC = 0.72. For hemolytic activity prediction, accuracy = 0.88, sensitivity = 0.87, specificity = 0.91, precision = 0.91, F-measure = 0.89, MCC = 0.77. Confusion matrices and the calculated parameters for cut-offs 0.35, 0.4, 0.45, …, 0.85 are represented in supplementary file Quality_of_prediction.xlsx (Sheets “Antimicrobial” and “Hemolytic”), and optimal cut-offs are colored in yellow for each of the models. ROC analysis was performed for estimation of the predictive quality of the two models (Fig. 6). ROC AUC for the first model was 0.914, and for the second model – 0.933. Briefly, the both models can be considered as adequately describing both antimicrobial and hemolytic activity.

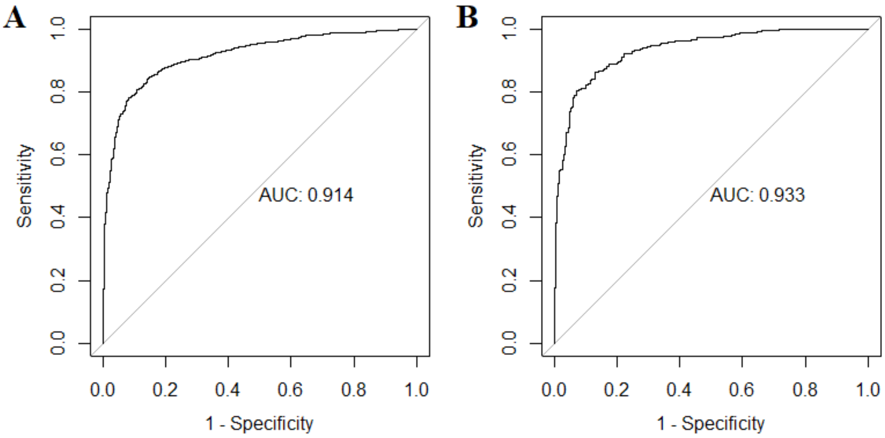


Figure 6 - Results of ROC analysis of antimicrobial (*A*) and hemolytic (*B*) activity prediction
DOI: https://doi.org/10.60797/jbg.2024.26.5.10

Further, density plots were constructed for the both predictive models: predicted $P_{Antimicrobial}$ (Fig. 7, **A**) and $P_{Hemolytic}$ (Fig. 7, **B**) are distributed with respect to the actual groups.

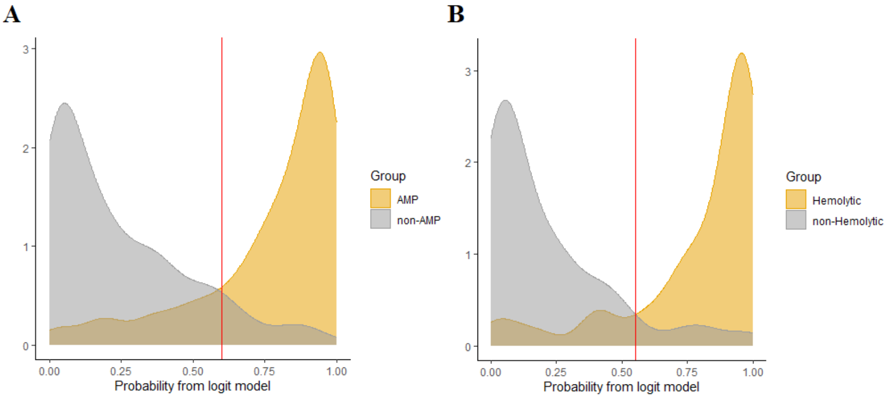


Figure 7 - Results of antimicrobial and hemolytic activity prediction (validation on test sets):
A – Performance of the model for antimicrobial activity prediction; *B* – Performance of the model for hemolytic activity prediction
DOI: https://doi.org/10.60797/jbg.2024.26.5.11

Note: orange density plots correspond to positive groups and grey plots correspond to negative groups; red vertical lines represent optimal cut-offs for distinguishing between positive and negative groups (0.6 for the AMPs model and 0.55 for the HPs model)

To sum up, two models for prediction of antimicrobial and hemolytic activity of peptides were constructed and estimated to have promising predictive potency. As such, these models can be used can be used for prediction of AMPs and HPs in large sequences.

3.2. Results of mature AMPs identification in their precursor proteins

The formula extracted from the AMPs predictive model was applied to reveal active peptides in their precursors. First, LL-37 was used for the algorithm validation. LL-37 precursor includes 170 a.a., and the mature peptide comprises a.a. 134–170 (the length is 37 a.a.). Indeed, a clear increase in $p_{Antimicrobial}$ can be seen at the C-terminus of the precursor after its screening by the sliding windows of lengths 20, 37 and 50 (Fig. 10). Alongside with this, N-terminal peptides also possessed high $p_{Antimicrobial}$ values; for 20-a.a. sliding window, active peptides were revealed in the center of the sequence (Fig 10, **A**), this was less pronounced for 37- and 50-a.a. windows (Fig. 10, **B**, **C**).

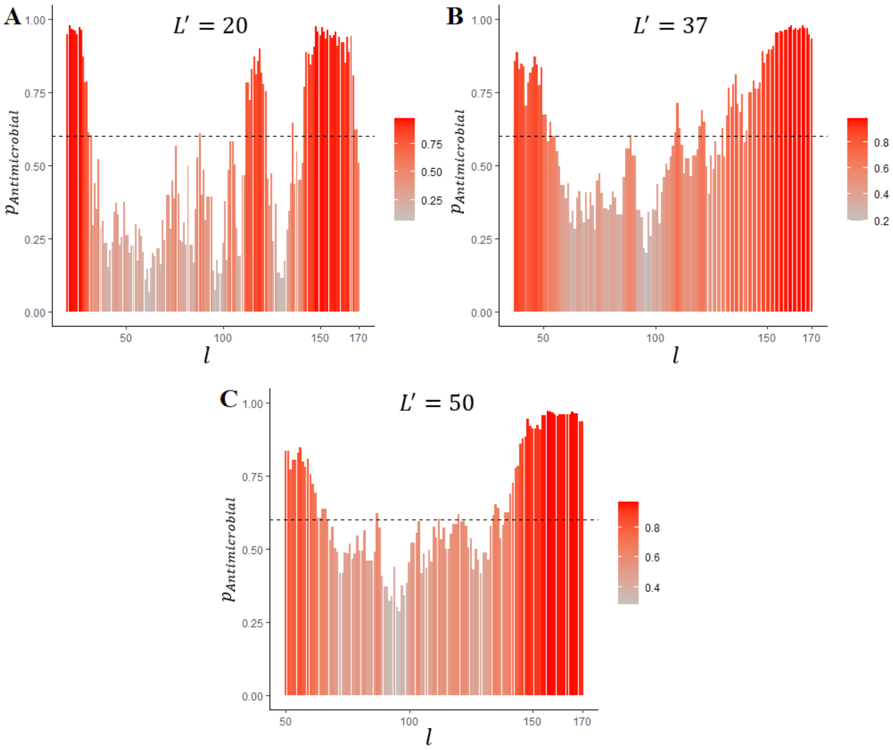


Figure 8 - Prediction of LL-37 localization in its precursor protein:
l is the rightmost a.a. in the sliding window, $p_{Antimicrobial}$ is probability of antimicrobial activity, *L'* is the sliding window length (*A* – 20, *B* – 37 and *C* – 50 amino acid a.a.)
DOI: https://doi.org/10.60797/jbg.2024.26.5.12

Note: the actual mature LL-37 length is 37 a.a. The dashed horizontal line represents the estimated optimal cut-off for AMPs prediction (0.6)

HNP1 precursor includes 94 a.a., and the mature peptide comprises a.a. 65–94 (the length is 30 a.a.). As for LL-37 prediction, $p_{Antimicribial}$ was undoubtedly increased at the C-terminus and some active peptides were recognized in the middle of the sequence (Fig. 9).

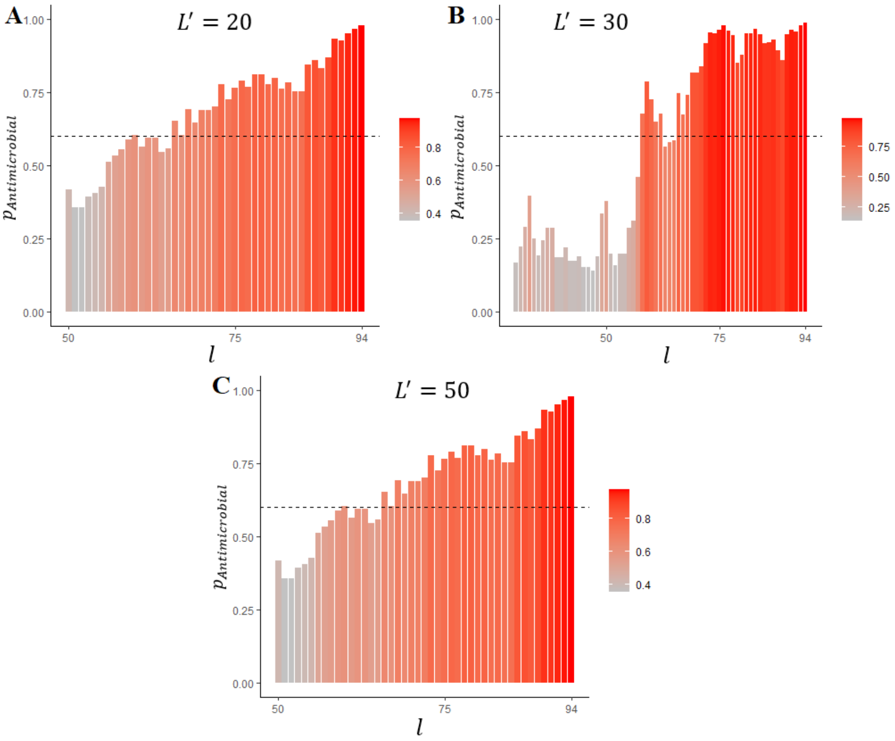


Figure 9 - Prediction of HNP1 localization in its precursor protein:
l is the rightmost a.a. in the sliding window, $p_{Antimicrobial}$ is probability of antimicrobial activity, *L'* is the sliding window length (*A* – 20 a.a.; *B* – 30 a.a. and *C* – 50 a.a.)
DOI: https://doi.org/10.60797/jbg.2024.26.5.13

Note: the actual mature HNP1 length is 30 a.a.; the dashed horizontal line represents the estimated optimal cut-off for AMPs prediction (0.6)

Tachyplesin I was used to illustrate recognition of AMPs in precursors containing mature peptides located not at their C-termini but in the middle part of the molecule. Tachyplesin precursor contains 77 a.a., and the AMP comprises positions 24–40 (the length is 17 a.a.). Fig. 10 demonstrates a great prediction of tachyplesin I in the protein with the sole $p_{\text{Antimicrobial}}$ maximum around the position , i.e. the rightmost a.a. of the AMP sequence.

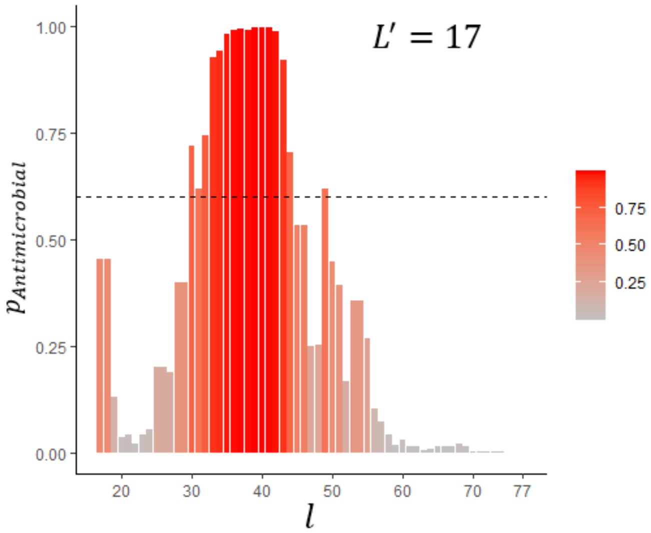


Figure 10 - Prediction of Tachyplesin I localization in its precursor protein:
l is the rightmost a.a. in the sliding window, $p_{Antimicrobial}$ is probability of antimicrobial activity, *L'* is the sliding window length (the only variant corresponding to the actual tacheplesin I length, i.e. 17 a.a.)
DOI: https://doi.org/10.60797/jbg.2024.26.5.14

Note: the dashed horizontal line represents the estimated optimal cut-off for AMPs prediction (0.6)

Discussion

The study includes two successive parts: ML and AMPs prediction from large proteins. The main idea used for prediction of AMPs and HPs was to reveal key structural and physicochemical parameters which are different between positive and negative peptides and to use them as predictors in logistic regression. As expected, predictors in the two models appeared to be co-directed (see Section 3.1). The effects of the selected features are consistent with typical parameters of AMPs widely described in literature (see Introduction) and are in accordance with the distribution of the peptides characteristics illustrated in Fig. 4 and Fig 5. Ideally, prediction of hemolytic activity should be based on direct comparison of hemolytic and non-hemolytic AMPs, which would require a special database.

In the present study, only antibacterial peptides were selected from the original database, but both gram-positive and gram-negative bacteria were included as targets. Indeed, the vast majority of the peptides in the original database were active against these two groups of microorganisms simultaneously. A negligible portion of AMPs were active strictly against one target group other than bacteria (this was the case for fungi, viruses but was not observed for cancers, mammalian cells and parasites). Most predictive tools deal with general antimicrobial activity, although some algorithms attempt to distinguish between AMPs activities by performing multi-label classification [38], [39]. Nevertheless, this approach was considered as redundant in the present study since one of the main traits of most AMPs is their wide spectrum of activity – antiviral, antiparasitic, antifungal, antitumor, often simultaneously [6]. As such, the built dataset may be relevant for description of AMPs with a wide spectrum of taxonomic targets.

In this study, cytotoxic activity was identified with hemolytic activity, which is not completely correct. Actually, different cell lines and red blood cells from different species demonstrate different susceptibility to the peptides' membranolytic action depending on their membrane lipid content [40]. However, it is difficult to take into account all possible side targets of AMPs due to their enormous variety, and hemolytic activity is a standard measure of AMPs toxic capacity. It should also be noted that cytotoxicity is not the sole limitation of AMPs implementation into the clinic. Stability *in vivo*, action on humoral targets [41] and a high cost of synthesis should be controlled during elaboration of novel AMPs.

We intended to elaborate an approach to reveal AMPs in long protein sequences. To validate the proposed model, we used precursor sequences of well-described AMPs because such sequences undoubtedly contain either non-antimicrobial or antimicrobial fragments corresponding to mature AMPs. The lengths of LL-37 and HNPs lie within the range 20–50 used for ML, so the sliding windows lengths applied for the sequences' analysis were 20 (the minimum possible size), 50 (the maximum possible size) and exactly equal to the AMP's length. In both cases, the C-terminal localization of the AMPs was correctly predicted (Fig. 8–10). Tachiplesin I was used as an example of AMP which is located in the middle but not at the C-terminus of its precursor. This AMP includes 17 a.a. and is shorter than peptides participating in ML. Nevertheless, an excellent prediction was made by the algorithm (Fig. 10). To sum up the results of the prediction, the proposed algorithm is effective at search for AMPs in proteins but many false positive discoveries were detected, especially for LL-37 and HNP1 prediction. Adjustment of the sliding window lengths and the cut-off levels may be beneficial for a particular research task solution. It should be also noted that antimicrobial activity can be possessed not only by the natural AMPs of the given lengths but also by slightly longer or shorter peptides. The preservation of activity of shorter AMPs derived from their longer parents is a desirable strategy but also a common challenge in AMPs elaboration pipelines [42], [43].

In principle, the suggested algorithm may be used for a) discovery of natural AMPs primarily serving as defensive molecules of immune systems in different species; b) revealing endogenous peptides with possible antimicrobial function releasing from proteins which are known to undergo proteolytic cleavage, including that related to immune response (complement cascade [44], blood clotting [45], etc.); c) search for antimicrobial domains in large proteins with miscellaneous functions. In all cases, the revealed peptides can be used as templates for further design of AMP-based drugs.

Conclusion

In this study, we proposed an approach combining ML by multiple LR and AMPs recognition in large proteins, namely in their precursors. Firstly, two predictive models of acceptable quality were constructed. Secondly, the use of the equations coefficients allowed to identify regions corresponding to mature AMPs correctly, although further improvements may be beneficial for enhancement of prediction quality.

Дополнительные материалы

Дополнительные материалы доступны на онлайн-странице статьи.

Финансирование

State assignment 122020300189-6.

Благодарности

Автор выражает благодарность к.б.н. М.Н. Берлову и к.х.н. М.М. Хайдуковой за их ценные комментарии.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Supplementary materials

Supplementary materials are available online on the article’s webpage.

Funding

Государственное задание №122020300189-6.

Acknowledgement

The author expresses his acknowledgements to Dr. M.N.Berlov and Dr. M.M.Khaydukova for their valuable comments.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы на английском языке / References in English

1. Ventola C.L. The antibiotic resistance crisis: part 1: causes and threats / C.L. Ventola // Pharmacy and Therapeutics. — 2015. — Vol. 40. — № 4. — P. 277.

2. Ventola C.L. The antibiotic resistance crisis: part 2: management strategies and new agents / C.L. Ventola // Pharmacy and Therapeutics. — 2015. — Vol. 40. — № 5. — P. 344.

3. Ghosh C. Alternatives to conventional antibiotics in the era of antimicrobial resistance / C. Ghosh [et al.] // Trends in Microbiology. — 2019. — Vol. 27. — № 4. — P. 323–338.

4. Rončević T. Antimicrobial peptides as anti-infective agents in pre-post-antibiotic era? / T. Rončević, J. Puizina, A. Tossi // International Journal of Molecular Sciences. — 2019. — Vol. 20. — № 22. — P. 5713.

5. Hale J.D.F. Alternative mechanisms of action of cationic antimicrobial peptides on bacteria / J.D.F. Hale, R.E.W. Hancock // Expert Review of Anti-infective Therapy. — 2007. — Vol. 5. — № 6. — P. 951–959.

6. Zhang Q.Y. Antimicrobial peptides: mechanism of action, activity and clinical potential / Q.Y. Zhang [et al.] // Military Medical Research. — 2021. — Vol. 8. — P. 1–25.

7. Ridyard K.E. The potential of human peptide LL-37 as an antimicrobial and anti-biofilm agent / K.E. Ridyard, J. Overhage // Antibiotics. — 2021. — Vol. 10. — № 6. — P. 650.

8. Lehrer R.I. α-Defensins in human innate immunity / R.I. Lehrer, W. Lu // Immunological Reviews. — 2012. — Vol. 245. — № 1. — P. 84–112.

9. Panteleev P.V. Design of antimicrobial peptide arenicin analogs with improved therapeutic indices / P.V. Panteleev [et al.] // Journal of Peptide Science. — 2015. — Vol. 21. — № 2. — P. 105–113.

10. Ramazi S. A review on antimicrobial peptides databases and the computational tools / S. Ramazi [et al.] // Database. — 2022. — Vol. 2022. — P. baac011.

11. Agüero-Chapin G. Emerging computational approaches for antimicrobial peptide discovery / G. Agüero-Chapin [et al.] // Antibiotics. — 2022. — Vol. 11. — № 7. — P. 936.

12. Wang G. Machine learning prediction of antimicrobial peptides / G. Wang, I.I. Vaisman, M.L. van Hoek // Computational Peptide Science: Methods and protocols. — New York : Springer US, 2022. — P. 1–37.

13. Ebrahimikondori H. Structure-aware deep learning model for peptide toxicity prediction / H. Ebrahimikondori [et al.] // Protein Science. — 2024. — Vol. 33. — № 7. — P. e5076.

14. Randou E.G. Binary response models for recognition of antimicrobial peptides / E.G. Randou, D. Veltri, A. Shehu // Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. — 2013. — P. 76–85.

15. Nordahl E.A. Activation of the complement system generates antibacterial peptides / E.A. Nordahl [et al.] // Proceedings of the National Academy of Sciences. — 2004. — Vol. 101. — № 48. — P. 16879–16884.

16. Pasupuleti M. Preservation of antimicrobial properties of complement peptide C3a, from invertebrates to humans / M. Pasupuleti [et al.] // Journal of Biological Chemistry. — 2007. — Vol. 282. — № 4. — P. 2520–2528.

17. Sonesson A. Antifungal activity of C3a and C3a-derived peptides against Candida / A. Sonesson [et al.] // Biochimica et Biophysica Acta (BBA)-Biomembranes. — 2007. — Vol. 1768. — № 2. — P. 346–353.

18. Zhang X.J. Insights into the antibacterial properties of complement peptides C3a, C4a, and C5a across vertebrates / X.J. Zhang [et al.] // The Journal of Immunology. — 2022. — Vol. 209. — № 12. — P. 2330–2340.

19. Singh M. Alpha-melanocyte stimulating hormone: an emerging anti-inflammatory antimicrobial peptide / M. Singh, K. Mukhopadhyay // BioMed Research International. — 2014. — Vol. 2014. — № 1. — P. 874610.

20. Torrent M. AMPA: an automated web server for prediction of protein antimicrobial regions / M. Torrent [et al.] // Bioinformatics. — 2012. — Vol. 28. — № 1. — P. 130–131.

21. Agrawal P. In silico approach for prediction of antifungal peptides / P. Agrawal [et al.] // Frontiers in Microbiology. — 2018. — Vol. 9. — P. 323.

22. Pirtskhalava M. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics / M. Pirtskhalava [et al.] // Nucleic Acids Research. — 2021. — Vol. 49. — № D1. — P. D288–D297.

23. Li W. Clustering of highly homologous sequences to reduce the size of large protein databases / W. Li, L. Jaroszewski, A. Godzi // Bioinformatics. — 2001. — Vol. 17. — № 3. — P. 282–283.

24. Li W. Tolerating some redundancy significantly speeds up clustering of large protein databases / W. Li, L. Jaroszewski, A. Godzik // Bioinformatics. — 2002. — Vol. 18. — № 1. — P. 77–82.

25. Li W. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences / W. Li, A. Godzik // Bioinformatics. — 2006. — Vol. 22. — № 13. — P. 1658–1659.

26. Timmons P.B. HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks / P.B. Timmons, C.M. Hewage // Scientific Reports. — 2020. — Vol. 10. — № 1. — P. 10869.

27. UniProt: the universal protein knowledgebase in 2023 // Nucleic Acids Research. — 2023. — Vol. 51. — № D1. — P. D523–D531.

28. Vishnepolsky B. Prediction of linear cationic antimicrobial peptides based on characteristics responsible for their interaction with the membranes / B. Vishnepolsky, M. Pirtskhalava // Journal of Chemical Information and Modeling. — 2014. — Vol. 54. — № 5. — P. 1512–1523.

29. Nakamura T. Tachyplesin, a class of antimicrobial peptide from the hemocytes of the horseshoe crab (Tachypleus tridentatus). Isolation and chemical structure / T. Nakamura [et al.] // Journal of Biological Chemistry. — 1988. — Vol. 263. — № 32. — P. 16709–16713.

30. R: A Language and Environment for Statistical Computing / R Foundation for Statistical Computing. — Vienna.

31. Fox J. An R companion to applied regression / J. Fox, S. Weisberg. — Sage publications, 2018.

32. Wickham H. dplyr: a grammar of data manipulation. R package version 1.1. 2 / H. Wickham [et al.] // Computer Software. — 2023.

33. Wickham H. Data analysis / H. Wickham. — Springer International Publishing, 2016. — P. 189–201.

34. Kassambara A. ggpubr:'ggplot2'based publication ready plots / A. Kassambara // R package version. — 2018. — P. 2.

35. Osorio D. Peptides: a package for data mining of antimicrobial peptides / D. Osorio, P. Rondón-Villarreal, R. Torres // Small. — 2015. — Vol. 12. — P. 44–444.

36. Robin X. pROC: an open-source package for R and S+ to analyze and compare ROC curves / X. Robin [et al.] // BMC Bioinformatics. — 2011. — Vol. 12. — P. 1–8.

37. Xiao N. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences / N. Xiao [et al.] // Bioinformatics. — 2015. — Vol. 31. — № 11. — P. 1857–1859.

38. Chung C.R. Characterization and identification of antimicrobial peptides with different functional activities / C.R. Chung [et al.] // Briefings in Bioinformatics. — 2020. — Vol. 21. — № 3. — P. 1098–1114.

39. Lin W. Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types / W. Lin, D. Xu // Bioinformatics. — 2016. — Vol. 32. — № 24. — P. 3745–3752.

40. Belokoneva O.S. The hemolytic activity of six arachnid cationic peptides is affected by the phosphatidylcholine-to-sphingomyelin ratio in lipid bilayers / O.S. Belokoneva [et al.] // Biochimica et Biophysica Acta (BBA)-Biomembranes. — 2003. — Vol. 1617. — № 1-2. — P. 22–30.

41. Krenev I.A. In vitro modulation of complement activation by therapeutically prospective analogues of the marine polychaeta arenicin peptides / I.A. Krenev [et al.] // Marine Drugs. — 2022. — Vol. 20. — № 10. — P. 612.

42. Mazurkiewicz-Pisarek A. Antimicrobial peptides: challenging journey to the pharmaceutical, biomedical, and cosmeceutical use / A. Mazurkiewicz-Pisarek, J. Baran, T. Ciach // International Journal of Molecular Sciences. — 2023. — Vol. 24. — № 10. — P. 9031.

43. Bucataru C. Antimicrobial peptides: Opportunities and challenges in overcoming resistance / C. Bucataru, C. Ciobanasu // Microbiological Research. — 2024. — P. 127822.

44. Egorova E.V. Antimicrobial activity of the complement system / E.V. Egorova [et al.] // Medical Academic Journal. — 2023. — Vol. 23. — № 2. — P. 31–45.

45. Wilhelm G. The crossroads of the coagulation system and the immune system: Interactions and connections / G. Wilhelm [et al.] // International Journal of Molecular Sciences. — 2023. — Vol. 24. — № 16. — P. 12563.