Sense the Moment: a highly sensitive antimicrobial activity predictor based on hydrophobic moment

William F. Porto[1,2]*, Karla C. V. Ferreira[1,3], Suzana M. Ribeiro[4], and Octavio L. Franco[1,3,5]*

[1] Pós-Graduação em Ciências Genômicas e Biotecnologia Universidade Católica de Brasília, Brasília-DF, Brazil.

[2] Porto Reports, Brasília-DF, Brazil.

[3] Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia Universidade Católica de Brasília, Brasília-DF, Brazil.

[4] Programa de Pós-Graduação em Ciências da Saúde, Universidade Federal da Grande Dourados, Dourados, MS, Brazil

[5] S-Inova Biotech, Pós-Graduação em Biotecnologia, Universidade Católica Dom Bosco, Campo Grande, MS, Brazil.

*Corresponding authors: William Farias Porto – williamfp7@yahoo.com.br, website: http://portoreports.com; Octavio Luiz Franco - ocfranco@gmail.com, Phone number: +55 61 34487167/ +55 61 34487220 Fax number: + 55 61 33474797

Abstract

The antibiotic resistance crisis has led to the need for new antimicrobial compounds. Computer-aided identification and design tools are indispensable for developing these antimicrobial agents. Antimicrobial peptides (AMPs) have aroused intense interest, since they have a broad spectrum of activity. Several systems for predicting antimicrobial peptides have been developed, using scalar physicochemical properties; however, regardless of the machine learning algorithm, these systems often fail in discriminating AMPs from their shuffled versions, leading to the need for new training methods to overcome this bias. Here we present "Sense the Moment", a prediction system capable of discriminating AMPs and shuffled versions. The system was trained using 684 entries: 342 from known AMPs and another 342 based on shuffled versions of known AMPs. Each entry contained the geometric average of three hydrophobic moments measured with different scales. The model showed good accuracy (>80 %) and excellent sensitivity (>90 %) for AMP prediction. Together with data from a direct screening from the NCBI non-redundant protein data base, our results demonstrate the system's applicability in a real-world scenario, aiding in identifying and discarding non-AMPs, since the number of false negatives is lower than false positives. The application of this model in virtual screening protocols for identifying and/or creating antimicrobial agents could aid in the identification of potential drugs to control pathogenic microorganisms and in solving the antibiotic resistance crisis. The system was implemented as a web application, available at <http://portoreports.com/stm/>.

Keywords: Shuffled Peptides, Antimicrobial Peptides, Hydrophobic Moment, Peptide screening, α-Helical Peptides.

1 Introduction

In the last few decades, the activity of conventional antibiotics has been reduced due to bacterial resistance development. In this context, antimicrobial peptides (AMPs) have been proposed as an alternative strategy to control antibiotic-resistant microorganisms.[1–3] AMPs are evolutionarily ancient molecules that have been identified from diverse sources, such as microorganisms, plants and animals. They play an important role in the innate immune system and are the first line of defence to protect internal and external surfaces of the host.[4] Overall, AMPs have the ability to adopt amphipathic structures, with a positive net charge between +3 and +9, and they are composed of dozens of amino acid residues. AMPs comprise a highly diverse group of antimicrobial compounds, split into two major groups, according to the presence or absence of disulphide bonds: cysteine-stabilized and disulphide-free peptides, respectively.[5] In addition, AMPs can be classified according to their origins: naturally occurring AMPs, which can be found in members of all kingdoms of life, and which are a product of specific genes;[4,6–8] cryptic AMPs, in larger proteins, where proteolytic cleavage under specific environmental conditions releases them;[9–11] and designed AMPs, developed with recent technologies of rational design and peptide synthesis.[1,12,13]

Conducting a search for AMPs in databases and undertaking their rational design remains a promising method for identifying and/or creating novel antimicrobial agents against pathogenic microorganisms. However, the identification of an active peptide sequence prior to chemical synthesis and susceptibility testing represents a challenge. Indeed, the identification of naturally occurring AMPs can be done by domain identification using regular expressions associated with programs like InterPro Scan,[14] as described by Porto et al.[15,16] However, for cryptic or even designed synthetic AMPs,

identification prior to synthesis still requires more sophisticated techniques, since tens[12] to hundreds[13] of peptides still need to be synthesized and assayed for their antimicrobial potency. Therefore, prediction systems could help to reduce the number of peptides to be tested, especially those systems implemented following Chou's 5-step rule.[17]

Most AMP prediction systems have been based on physicochemical/structural properties, such as charge, hydrophobicity and secondary structure formation[5,18,19] and, regardless of machine learning algorithm (e.g. neural network, support vector machine), the systems have the same or similar prediction assessments. However, our group previously demonstrated that these prediction tools often fail when challenged with shuffled sequences,[20] because some physicochemical properties (e.g. charge and hydrophobicity) keep the same values after the sequence is shuffled, in contrast to antimicrobial activity.[12]

Thus, exploring genomes and or sequence databases with current antimicrobial activity prediction tools is not feasible, because sequences with a similar composition could generate false positive hits. Instead of using scalar properties, the application of vector properties, such as hydrophobic moment, as developed by Eisenberg et al.,[21] could allow more accurate predictions since these types of properties would change if sequence is shuffled.[22] Then, the use of vector properties could help in the process of designing and selecting novel antimicrobial peptides. Here we report the development of a highly sensitive model for the prediction of AMPs, named "Sense the Moment", and describe the system's application for direct AMP screening.

2 Results

2.1 Hydrophobic moment discriminates AMPs and their shuffled versions

Initially, for system implementation, positive and negative data sets were constructed (PS and NS), containing the geometric hydrophobic moments, calculated by the geometric average of three hydrophobic moments in different scales. NS was derived from PS by shuffling the sequences in PS, resulting in reduced geometric hydrophobic moments in NS (Figure 1). According to receiver-operating characteristic (ROC) curve (Figure 1), the system presented a very good performance, showing an area under the curve of 0.899, where the threshold of 0.577 showed an accuracy of 89.2 %.

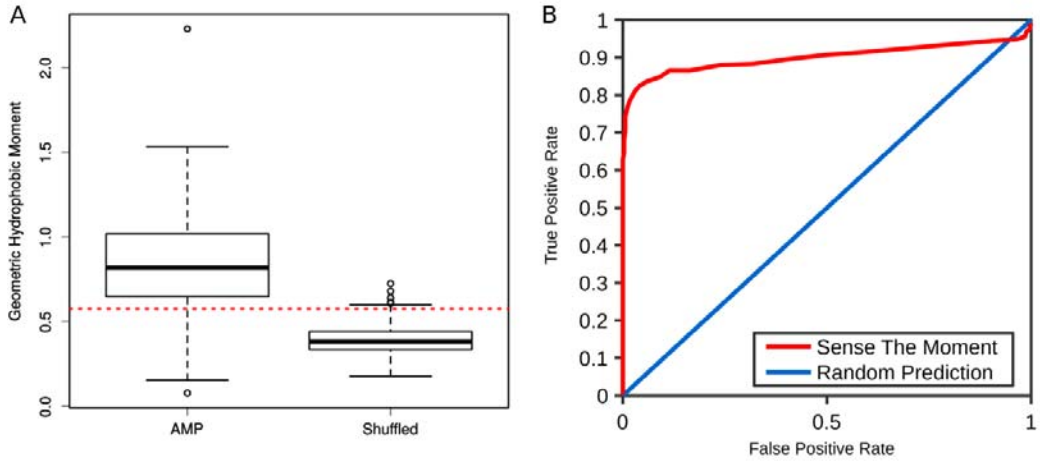

Figure 1. Discriminating AMPs and their shuffled versions using the geometric hydrophobic moment. (A) Geometric hydrophobic moment distribution within PS and NS. Despite the overlap between the sets, the geometric hydrophobic moment values calculated from PS were statistically higher than NS according to the Wilcoxon-Mann-Whitney non-parametric test, showing the same p-value for the three used scales (p-value <2.2e-16). In fact, only 20 out 342 shuffled sequences (5.85 %) had their geometric hydrophobic moments increased after shuffling. Dotted red line indicates the cut-off with the best accuracy (0.577) (B) Model ROC curve in relation to the training set. Blue line indicates a random prediction and the red line indicates the Sense the Moment prediction. Sense the Moment presented an AUC of 0.899.

2.2 Sense the Moment has Excellent Sensitivity

Table 1 summarises the assessments of Sense the Moment against blind data

sets. The first evaluation with a training data set slightly overestimated the model, because accuracies on Loose and Nagarajan data sets were 82 and 85 %, respectively. However, sensitivity drew attention, due to excellent rates in both blind sets, reaching 95 and 94.4 % on Loose and Nagarajan data sets, respectively. However, if the MCC value was taken into account, the system was better in discriminating designed and shuffled peptides than antimicrobial and non-antimicrobial peptides. Furthermore, the model has a PPV of 76.0 and 89.5 % on Loose and Nagarajan data sets, respectively.

Table 1. Prediction assessments of Sense the Moment on the blind data sets.

Data Set	Sensitivity	Specificity	Accuracy	PPV	MCC
Loose	95.0	69.2	82.3	76.0	0.7
Nagarajan	94.4	0.0	85.0	89.5	-0.1

2.3 True Negative Sequences are Sense the Moment targets in Peptide Screening

To evaluate the predictive power of Sense the Moment for antimicrobial peptides in a real-world scenario, a set of peptides was screened based on sequences derived from NCBI's non-redundant database (NR). Initially, 3879 peptide sequences were retrieved. Since the hydrophobic moment requires a secondary structure, the α-helix propensity of all peptides was calculated in order to select the most α-helical peptides. Finally, a subset containing 11 sequences with α-helix propensities ranging from 0.286 to 0.312 $Kcal.mol^{-1}$ was chosen. This set has two peptides with solved structure (PDB 1DU1 and 2FLY) showing α-helix conformation (Table 2). These sequences were classified by Sense the Moment and then synthesized and tested against *P. aeruginosa* (Table 2). Interestingly, only the peptide CBL20939 was more potent than magainin 2, with MIC of 25 $\mu g.mL^{-1}$. The other AMP, 2FLY_A, was not predicted as an AMP by Sense the Moment, which considered it a false negative result. However the peptide 2FLY_A showed a relatively weak MIC (100 $\mu g.\ mL^{-1}$).

Table 2. Peptides identified by direct screening.

GenBank ID	Sequence	Helix Propensity	Score	Prediction	MIC*
Magainin 2 (2MAG)	GIGKFLHSAKKFGKAFVGEIMNS		0.922	AMP	100
CBL20939	MIKSVAKKAAQGLRRRKERK	0.295	0.820	AMP	25
CBL36218	MTAAQAVREETAALMRSFYA	0.290	0.706	AMP	>200
BAB89365	MATRNLEKLASIDAQLRALV	0.304	0.628	AMP	>200
1DU1_A	TSAQKAKAEERKRRKMSRGL	0.309	0.609	AMP	>200
ZP_01169002	MKNSKAAANILQLLFFFAKL	0.304	0.573	Non-AMP	>200
ZP_02493962	MYDVARETALERARNLSARL	0.308	0.429	Non-AMP	>200
ZP_01787376	ARAKAKKLAKTQATLENNQE	0.294	0.467	Non-AMP	>200
NP_740433	AAYAGRARAQKQALQVLDIQ	0.295	0.262	Non-AMP	200
ZP_02357056	MIAAFSVAAEADSDALALDE	0.305	0.222	Non-AMP	>200
2FLY_A	ARLDVAAEFRKKWNKWALSR	0.310	0.158	Non-AMP	100
CBK77156	MEKQAALAAAKNACADGKYT	0.312	0.206	Non-AMP	>200

*The MIC value was considered the concentration in ug.mL^{-1} that reduced 90% of luminescence.

3 Discussion

In the last decade, numerous computational methods to predict the antimicrobial activity of AMPs have been reported.[5,18,19,22–25] The majority of such methods use different machine learning algorithms, however this results in only a slight increase or decrease in prediction performance. Besides, these prediction tools are not able to handle shuffled sequences.[20] Because a shuffled sequence has the same scalar properties as native sequence, the sequences would have the same or similar predictions.[20] Thus, any arbitrary sequence with a similar composition of an AMP would be predicted as an AMP. Hence, considering the choice of machine learning algorithm with little or no effect on prediction performance, training is the Achilles' heel of all these systems. Most systems use scalar properties for training algorithms,[5,18,19] resulting in the compositional bias.[20] This is a problem when genomes and databases are used as a source for such molecules.[26]

Therefore, the use of descriptors that take the amino acid order into account is desirable to discriminate between AMPs and non-AMPs with similar sequences.[22] In this context, there are some properties that change when sequence order is altered,

including hydrophobic moment[21] and Chou's pseudo amino acid composition.[27] The iAMP-2L server[28] was constructed using Chou's pseudo amino acid composition; however, in our previous report, this server was not able to distinguish between designed and shuffled sequences.[20]

Here, we posited hydrophobic moment as a way to differentiate AMPs and shuffled sequences, because this property incorporates three important AMP characteristics: hydrophobicity, net charge and amino acid order. However, there are two biases associated with the hydrophobic moment, since its value is connected, firstly, to the hydrophobicity scale used [29] and, secondly, to the peptide's secondary structure.[30] Although hydrophobicity scales are strongly correlated,[29] the hydrophobicity of a given amino acid could differ within scales. For instance, in Eisenberg's scale, phenylalanine is more hydrophobic than leucine, in contrast to Kyte-Doolittle's. To overcome this bias, Sense the Moment was developed using the geometric average of three hydrophobic moments, each one calculated with a different scale; and the geometric average was essential, because the three values have the same weight independently of the scale used.

The second bias, related to secondary structure, was handled by selecting AMP sequences with α-helix as secondary structure, because δ in eq. 1 was set to 100°. Even so, Sense the Moment could, in theory, be applied to β-stranded peptides. Mathematically, β-strand structure could present elevated values of hydrophobic moment even with δ adjusted to an α-helix, because in an amphipathic α-helix, residues X_i and X_{i+4} would be on the same side (e.g. X_1 and X_5), while in an amphipathic β-strand, the residues on the same side are X_i and X_{i+2}. Following this reasoning, in a β-strand, residues X_i, X_{i+2} and X_{i+4} would be on the same side and, therefore, generating elevated values for the β-strand even with δ adjusted to an α-helix. In fact, in Loose and

Nagaranjan data sets there were sequences that do not form α-helixes.[20,31] Therefore, Sense the Moment is restricted to short peptides with low-complexity secondary structures, in particular disulphide-free peptides.

Besides these specific biases, there are two general biases for AMP prediction, which are the variable length of sequences and the absence of a data set of non-antimicrobial peptides.[32] Because the geometric hydrophobic moment is independent of sequence length, the length bias was avoided. In the case of a negative data set, the Loose data set could be used; however, this data set has few sequences for training the system, and we preferred to use it as a blind set for testing the system. Thus, instead of using the Loose data set, we used a theoretical set, generated by shuffling sequences from the positive data set. However, a single round of shuffling did not ensure the generation of a shuffled peptide, especially if the original sequence was composed of a set of restricted and repetitive amino acids. The shuffling process, therefore, had to be repeated 10,000 times for each sequence in order to obtain homogeneous hydrophobic moment values (Figure S1). Thus, assuming α-helical AMPs have higher hydrophobic moment values, shuffled sequences would have their hydrophobic moments reduced. Indeed, it is only in 5.85 % of AMP sequences that the geometric hydrophobic moment increases when the sequence is shuffled and, therefore, there are statistical differences in hydrophobic moments, when comparing AMPs and shuffled peptides (Figure 1).

In fact, the system had very good accuracy in both blind sets, considering designed/shuffled and AMP/non-AMP (Table 1). However, the strength of the system relies on its sensitivity, which is what give the system its name. Thus, in order to demonstrate its application in a peptide screening process, a direct search on NR was conducted. Because Sense the Moment has excellent sensitivity and only reasonable specificity, a few false negative predictions were expected. With this in mind, Sense the

Moment should be used for removing peptides predicted as non-AMP, reducing the number of sequences to evaluate *in vitro*. With regard to direct screening (Table 1), there was a single false negative, the hormone proadrenomedullin (Genbank id 2FLY_A), which showed low potency (MIC of 100 $\mu g.ml^{-1}$). Interestingly, this hormone falls into the group that, once shuffled, had increased hydrophobic moment (0.158 in native form, 0.525 in average shuffled form). However, this peptide would be easily discarded in a high throughput screening due to its MIC value. On the other hand, one of four peptides predicted as AMPs showed reasonable activity (MIC of 25 $\mu g.ml^{-1}$), which was more potent than magainin 2 (MIC of 100 $\mu g.ml^{-1}$).

There were three false positives in direct screening, which could be due to the false positives generated by hydrophobic moment when used to predict sequences as AMPs.[30] In addition, this highlights some adjustments that could be necessary in the cut-off of Sense the Moment cut-off, selecting a value higher than 0.577. However, the main limitation of Sense the Moment is that it merely takes the geometric hydrophobic moment into account, which has only two dimensions, while the three-dimensional structure of AMPs could be more complex and might interfere in antimicrobial activity. Despite that, Sense the Moment could be valuable in eliminating sequences with no antimicrobial activity from peptide screening projects, as demonstrated in our direct screening. Besides, our system could be used together with other prediction systems, allowing the best characteristics of each system to be exploited, in particular with application of Sense the Moment RESTFull API, which opens the possibility of generating consensus methods for AMP prediction.

4 Conclusions

The application of computers for identifying and/or designing AMPs has helped the field to move forward. These computational models can explore vast quantities of

data and generate the next generation of AMPs. However, this is a recent field in machine learning applications, with nebulous and unexplored regions. At the beginning, the main biases were sequence length variation and the absence of a non-AMP database. By now, the main bias is discriminating AMPs from sequences with the same composition in a different order. Due to the small overlap between AMPs and their shuffled versions, Sense the Moment partially solves this bias. Nevertheless, the application of Sense the Moment in AMP screening has proven to be useful, with a direct application in identifying natural or encrypted sequences. Henceforth, Sense the Moment could be a valuable tool for identifying drug candidates to help to solve the antibiotic resistance crisis.

5 Material and Methods

5.1 Data Sets

5.1.1 Positive Data Set

The Antimicrobial Peptides Database[33] was used as the primary source of AMP sequences. Initially, the secondary structure of all sequences was predicted using PSIPRED,[34] and for the construction of the positive data set (PS), only sequences with at least 70% of helical content were selected. Then, sequences with redundancy greater than 70 % were removed through CD-HIT.[35] This set was manually curated, removing incomplete sequences, totalling 342 sequences. For each sequence, three average hydrophobic moments were calculated, using three hydrophobic scales: Eisenberg,[21] Kyte-Doolittle[36] and Radzicka-Wolfenden.[37,38] The average hydrophobic moment was calculated using Eisenberg's equation:[21]

$$Hm = \frac{\sqrt[2]{\left[\sum_{i=1}^{I} H_i \times cos(\delta i)\right]^2 + \left[\sum_{i=1}^{I} H_i \times sin(\delta i)\right]^2}}{I} \quad (1)$$

Where δ represents the angle between the amino acid side chains (100° for α-helix, in average); *i* represents the residue number in the position *i* from the sequence; H_i represents the ith amino acid's hydrophobicity on a hydrophobicity scale; and *I* represents the total number of residues. Then, the three hydrophobic moments were used to calculate the geometric average of the hydrophobic moment, which will be referred to as "geometric hydrophobic moment".

5.1.2 **Negative Data Set**

Given that shuffled versions of AMPs often lose their activity, the negative data set (NS) was derived from the positive data set by shuffling sequences. Because the shuffling is a random process, variance must be reduced to ensure reproducibility. Hence, an initial analysis was performed to verify the number of repetitions necessary to minimize this variance (Figure S1). Ten thousand shuffled versions of each positive sequence were generated. The NS was composed of 342 entries, representing the average of 10,000 geometric hydrophobic moment values of shuffled versions of AMP sequences. A one-sided Wilcoxon-Mann-Whitney non-parametric test was applied to verify the differences between geometric hydrophobic moment in PS and NS sets, with a critical value of 0.05, using R package for statistical computing (http://www.r-project.org).

5.1.3 Blind Data Sets

Two blind data sets were used to assess the predictive capacity of Sense the Moment, regarding the differentiation between designed and shuffled sequences and the

differentiation between antimicrobial and non-antimicrobial peptides. The first blind data set was derived from Loose et al.[12] (Loose data set). The set was manually curated, removing sequences that had not been experimentally validated, resulting in 40 designed peptides and 39 shuffled peptides. The second blind data set was derived from Nagarajan et al.[31] (Nagarajan data set), corresponding to 20 peptides tested against 30 organisms of Gram-negative, Gram-positive, mycobacterial, and fungal origin, where 18 were active against at least one organism and two were completely inactive.

5.2 Prediction Cut-off Estimation

By using positive and negative sets together, the cut-off value was evaluated in intervals of 0.021, where for each interval true positive and true negative rates were measured. The cut-off value with the most correct predictions was selected as default. The area under the curve (AUC) was measured using the rollmean function of zoo library from R package for statistical computing.

5.3 Assessments against blind sets

The model was challenged against the blind data sets and the following parameters were measured:

$$Sensitivity = \left[\frac{TP}{(TP + FN)} \right] \times 100 \tag{2}$$

$$Specificity = \left[\frac{TN}{(TN + FP)} \right] \times 100 \tag{3}$$

$$Accuracy = \left[\frac{(TP + TN)}{(TP + TN + FP + FN)} \right] \times 100 \tag{4}$$

$$PPV = \left[\frac{TP}{(TP + FP)} \right] \times 100 \tag{5}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt[2]{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (6)$$

Where TP is the number of true positives; FN, the false negatives; TN, the true negatives; FP, the false positives; PPV, the probability of correct positive prediction; NPV, the probability of correct negative prediction; and MCC, Matthews Correlation Coefficient, which measures the quality of the prediction (the higher the value, the better the prediction is).

5.4 Web interface and RESTFull API

Sense the Moment was developed as a web application, using the PHP programming language. The system can be accessed by a human-friendly interface with an HTML form or by a RESTFull API, using GET or POST HTTP request methods, for single sequence or fasta file submission, respectively. The system is available at <http://portoreports.com/stm>.

5.5 Peptide Screening

The NCBI's non-redundant protein database was used as a source of sequences. Sequences containing 20 amino acid residues were selected and their α-helix propensities were calculated using the α-helix propensity scale from Pace and Scholtz.[39] A subset containing 11 peptides with a strong propensity to form α-helices was selected and a small array of 12 peptides (11 peptides plus magainin 2, the positive control) was designed and produced by Kinexus Bioinformatics Corporation (Vancouver, BC) using the SPOT synthesis approach onto cellulose membranes as previously described by Winkler et al.[40] The antimicrobial activity of the synthesized peptides was evaluated against an engineered luminescent *Pseudomonas aeruginosa* strain in 96-well microplates, as described previously, with few modifications.[41] For a detailed

description, please see supplementary methods.

6 Acknowledgments

The authors would like to thank Dr. Robert E. W. Hancock and Dr. Evan F. Haney for their support in the experimental section, including peptide array and bioluminescent antimicrobial assays. Porto Reports is hosted at Free Web Hosting Area <http://freewebhostingarea.com>. This work was supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) and Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul (FUNDECT).

7 References

1. Fjell, C. D., Hiss, J. A., Hancock, R. E. W. & Schneider, G. Designing antimicrobial peptides: form follows function. *Nat. Rev. Drug Discov.* **11**, 37–51 (2011).

2. Porto, W. F., Silva, O. N. & Franco, O. L. Prediction and Rational Design of Antimicrobial Peptides. in *Protein Structure* (ed. Faraggi, E.) 377–396 (InTech, 2012). doi:10.5772/2335

3. Franco, O. L. Elucidating Novel Bacterial Targets and Designing Unusual Antimicrobial Peptides: Two Faces of the Same Proteomic Coin. *J. Proteomics Bioinform.* **s8**, (2014).

4. Silva, O. N. *et al*. Exploring the pharmacological potential of promiscuous host-defense peptides: from natural screenings to biotechnological applications. *Front. Microbiol.* **2**, 232 (2011).

5. Porto, W. F., Pires, Á. S. & Franco, O. L. CS-AMPPred: an updated SVM model for antimicrobial activity prediction in cysteine-stabilized peptides. *PLoS One* **7**, e51444 (2012).

6. Pinto, M. F. S. *et al.* Cyclotides: From Gene Structure to Promiscuous Multifunctionality. *J. Evid. Based. Complementary Altern. Med.* **17**, 40–53 (2011).

7. Cândido, E. S. *et al.* Structural and functional insights into plant bactericidal peptides. in *Science against Microbial Pathogens: Communicating Current Research and Technological Advances* (ed. Méndez-Vilas, A.) 951–960 (Formatex, 2011).

8. Ribeiro, S. M. *et al.* Plant Antifungal Peptides. in *Handbook of Biologically Active Peptides* (ed. Kastin, A. J.) 169–179 (Elsevier Inc., 2013).

9. Brand, G. D. *et al.* Probing protein sequences as sources for encrypted antimicrobial peptides. *PLoS One* **7**, e45848 (2012).

10. Okubo, B. M. *et al.* Evaluation of an antimicrobial L-amino acid oxidase and peptide derivatives from Bothropoides mattogrosensis pitviper venom. *PLoS One* **7**, e33639 (2012).

11. Mandal, S. M. *et al.* Identification of multifunctional peptides from human milk. *Peptides* **56**, 84–93 (2014).

12. Loose, C., Jensen, K., Rigoutsos, I. & Stephanopoulos, G. A linguistic model for the rational design of antimicrobial peptides. *Nature* **443**, 867–9 (2006).

13. Cherkasov, A. *et al.* Use of artificial intelligence in the design of small peptide antibiotics effective against a broad spectrum of highly antibiotic-resistant superbugs. *ACS Chem. Biol.* **4**, 65–74 (2009).

14. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–40 (2014).

15. Porto, W. F., Souza, V. A., Nolasco, D. O. & Franco, O. L. In silico identification of novel hevein-like peptide precursors. *Peptides* **38**, 127–136 (2012).

16. Porto, W. F., Fensterseifer, G. M. & Franco, O. L. In silico identification, structural characterization, and phylogenetic analysis of MdesDEF-2: a novel defensin from the Hessian fly, Mayetiola destructor. *J. Mol. Model.* **20**, 2339 (2014).

17. Chou, K.-C. Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol*. **273**, 236–47 (2011).

18. Torrent, M., Andreu, D., Nogués, V. M. & Boix, E. Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PLoS One* **6**, e16968 (2011).

19. Waghu, F. H. *et al.* CAMP: Collection of sequences and structures of antimicrobial peptides. *Nucleic Acids Res.* **42**, D1154-8 (2014).

20. Porto, W. F., Pires, Á. S. & Franco, O. L. Antimicrobial activity predictors benchmarking analysis using shuffled and designed synthetic peptides. *J. Theor. Biol*. **426**, 96–103 (2017).

21. Eisenberg, D., Weiss, R. M., Terwilliger, T. C. & Wilcox, W. Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc*. **17**, 109 (1982).

22. Porto, W. F., Fernandes, F. C. & Franco, O. L. An SVM Model Based on Physicochemical Properties to Predict Antimicrobial Activity from Protein Sequences with Cysteine Knot Motifs. in *Advances in Bioinformatics and Computational Biology* (eds. Ferreira, C. E., Miyano, S. & Stadler, P. F.) **6268**,

59–62 (Springer Berlin Heidelberg, 2010).

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

24. Fjell, C. D. *et al.* Identification of novel antibacterial peptides by chemoinformatics and machine learning. *J. Med. Chem.* **52**, 2006–15 (2009).

25. Fernandes, F. C., Rigden, D. J. & Franco, O. L. Prediction of antimicrobial peptides based on the adaptive neuro-fuzzy inference system application. *Biopolymers* **98**, 280–7 (2012).

26. Porto, W. F., Pires, A. S. & Franco, O. L. Computational tools for exploring sequence databases as a resource for antimicrobial peptides. *Biotechnol. Adv.* **35**, (2017).

27. Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* **43**, 246–55 (2001).

28. Xiao, X., Wang, P., Lin, W.-Z., Jia, J.-H. & Chou, K.-C. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* **436**, 168–77 (2013).

29. Porto, W. F., Fensterseifer, I. C. M., Ribeiro, S. M. & Franco, O. L. Joker: An algorithm to insert patterns into sequences for designing antimicrobial peptides. *Biochim. Biophys. Acta - Gen. Subj.* **1862**, (2018).

30. Porto, W. F. *et al.* In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nat. Commun.* **9**, 1490 (2018).

31. Nagarajan, D., Nagarajan, T., Nanajkar, N. & Chandra, N. A Uniform In Vitro Efficacy Dataset to Guide Antimicrobial Peptide Design. *Data* **4**, 27 (2019).

32. Lata, S., Sharma, B. K. & Raghava, G. P. S. Analysis and prediction of antibacterial peptides. *BMC Bioinformatics* **8**, 263 (2007).

33. Wang, G., Li, X. & Wang, Z. APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res.* **37**, D933-7 (2009).

34. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices 1 1Edited by G. Von Heijne. *J. Mol. Biol.* **292**, 195–202 (1999).

35. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–9 (2006).

36. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–32 (1982).

37. Radzicka, A. & Wolfenden, R. Comparing the Polarities of the Amino Acids: Side-Chain Distribution Coefficients between the Vapor Phase, Cyclohexane, 1 - 0ctano1, and Neutral Aqueous Solutiont. 1664–1670 (1988).

38. Boman, H. G. Antibacterial peptides: basic facts and emerging concepts. 197–215 (2003).

39. Pace, C. N. & Scholtz, J. M. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75**, 422–7 (1998).

40. Winkler, D. F. H., Hilpert, K., Brandt, O. & Hancock, R. E. W. Synthesis of peptide arrays using SPOT-technology and the CelluSpots-method. *Methods Mol. Biol.* **570**, 157–74 (2009).

41. Hilpert, K. & Hancock, R. E. W. Use of luminescent bacteria for rapid screening and characterization of short cationic antimicrobial peptides synthesized on cellulose using peptide array technology. *Nat. Protoc.* **2**, 1652–60 (2007).