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

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

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

[2] Pós-Graduação em Ciências Genômicas e Biotecnologia Universidade Católica de Brasília, 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 776 entries: 388 from known AMPs and another 388 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. Our results demonstrate the system's applicability, 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 can adopt amphipathic structures, with a positive net charge and they are composed of dozens of amino acid residues.[5] 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 prior to synthesis requires 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.[14] Most AMP prediction systems have been based on physicochemical/structural properties, such as charge, hydrophobicity and secondary structure formation[5,15,16] and, regardless of machine learning algorithm (e.g. neural network, support vector machine), the systems have the same or similar prediction assessments.[17]

However, the linguistic model for the rational design of antimicrobial peptides[12] pointed a problem that hinders the correct classification on AMP prediction systems: shuffled versions of AMP sequences have the same physicochemical properties (e.g. charge and hydrophobicity), but the antimicrobial activity is abolished.[12] In this context, our group previously demonstrated that the AMP prediction tools often fail when challenged with shuffled sequences, even in the case of iAMP-2L, which uses the correlation between amino acid residues based on Chou's pseudo amino acid composition.[17] Therefore, exploring genomes and or sequence databases with current antimicrobial activity prediction tools is not feasible, because sequences with a similar composition could generate a huge number of false positive hits.

In this context, in 2010, our group proposed the inclusion of vector properties, such as hydrophobic moment, as developed by Eisenberg et al.,[18] as a way to avoid this bias of identical composition of shuffled sequences.[19] However, despite the hydrophobic moment is strongly connected with the sequence, this property is not correlated with antimicrobial activity.[20,21] Therefore, we hypothesized that the hydrophobic moment could be used as a classifier between ordered and shuffled sequences. Therefore, here we report the development of a sensitive system for distinguishing ordered and shuffled peptides, with high sensitivity for AMP prediction.

2 Results

2.1 Hydrophobic moment discriminates AMPs and their shuffled versions

In order to discriminate shuffled sequences, the sequences from Loose et al. would be used as training data, because this data set is the proof of principle that shuffled sequences lost their activities. However, this data set has few sequences for training the system, and we preferred to use these sequences as a blind set for testing the system. Therefore, instead of using the Loose data set, we used sequences from the APD as the positive data set (PS); and as the negative data set (NS), we used a theoretical set, generated by shuffling sequences from the positive data set (Figure 1A). The datasets are available on supplementary Tables 1-3.

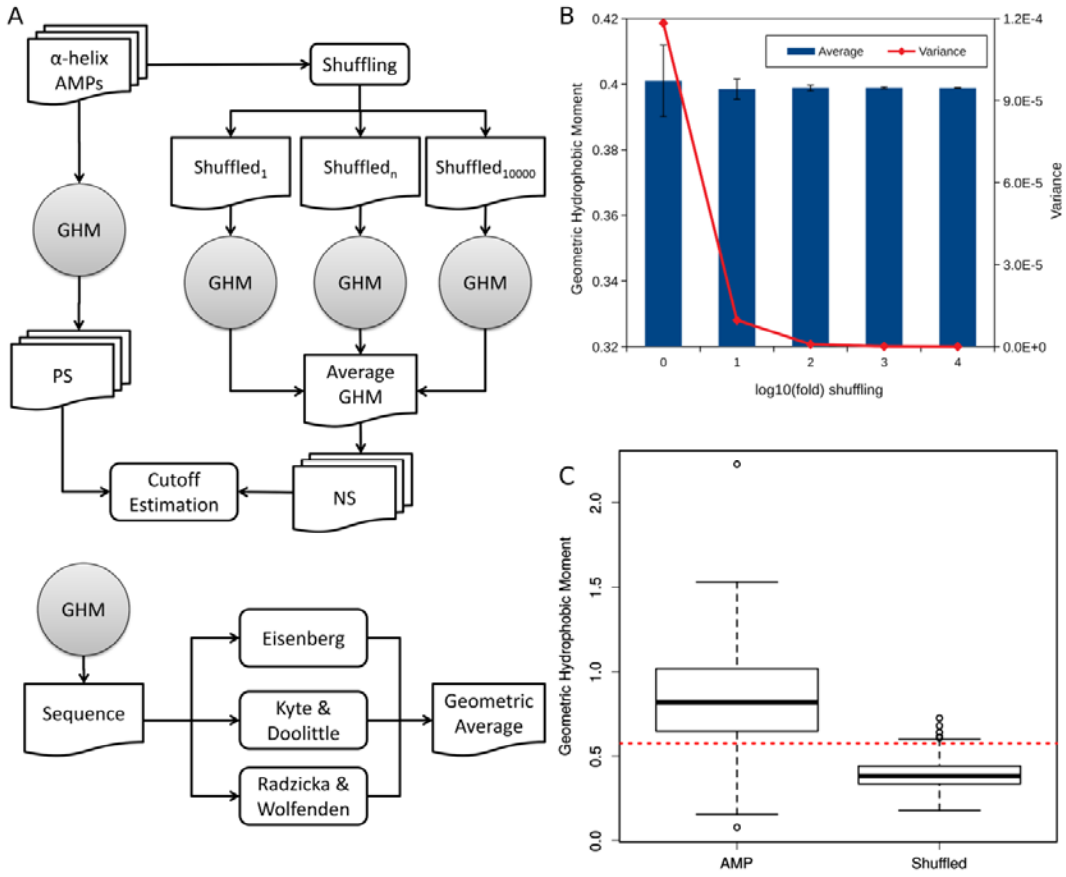


Figure 1. Discriminating AMPs and their shuffled versions using the geometric hydrophobic moment. (A) The dataset construction flowchart. The α-helical peptides were subjected to geometric hydrophobic moment calculation, generating the positive dataset (PS). The same set of peptides was subjected to a shuffling process in order to generate the negative dataset (NS). A total of 10,000 rounds of shuffling were applied in order to abolish the variance generated by the randomization process. Then, PS and NS were used for threshold definition. Because the geometric hydrophobic moment (GHM) is a single value, there is no need to use any machine learning engine for classification. (B) Reducing data variance from shuffled sequences. Data represents the average of 100 independent shuffling experiments. In each experiment, each sequence was shuffled n-folds, where n is expressed as the $\log_{10}$(fold) shuffling. Once our negative data set was composed of shuffled sequences, increase the number of shuffling rounds reduces the variance in the set, allowing reproducibility, even with the bias of randomness data. Shuffling the peptides by 10,000 folds almost abolished the variance and consequently the standard deviation. (C) Boxplot of 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 24 out of 388 shuffled sequences (6.19 %) had their geometric hydrophobic moments increased after shuffling. The dotted red line indicates the best cut-off value (0.590).

For PS generation, initially, all sequences with at least 70 % of helical content on PSIPRED predictions were selected, then the redundant sequences at 70 % were removed, and finally, the set was curated by hand, remaining 388 sequences (Supplementary Table S1).

For NS generation, 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. In addition, being a random process, shuffling increases the dataset variance, which could hinder reproducibility. Therefore, to reduce the variance and in order to obtain homogeneous hydrophobic moment values, the shuffling process was repeated by 1, 10, 100, 1,000 and 10,000 times for each sequence, and then the set was averaged (Figure 1). Despite there is no statistical difference between the sets regardless of the repetition times, increasing the repetitions abolished the variance caused by shuffling, and the set with 10,000 repetitions was selected.

Because we are assessing a single descriptor, there is no need for a machine learning classifier. Overall, geometric hydrophobic moment values in NS were reduced in comparison with PS (Figure 2B). A similar trend was observed using individual scales (data not shown). According to the receiver-operating characteristic (ROC) curve (Figure 2), the system presented a very good performance, showing an area under the curve of 0.891, where the threshold of 0.59 showed an accuracy of 88.79 %.

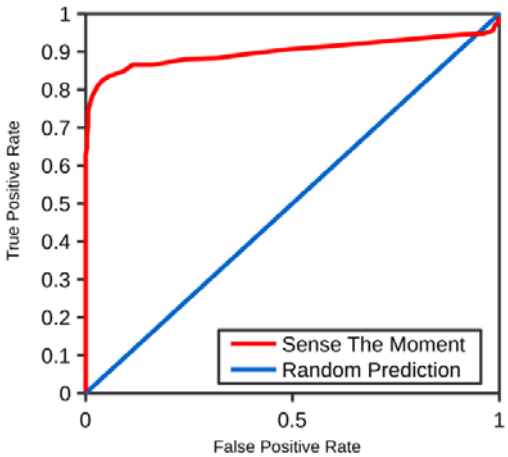


Figure 2. Receiver-operating characteristic (ROC) curve concerning the training set. The blue line indicates a random prediction and the red line indicates the Sense the Moment prediction. Sense the Moment presented an AUC of 0.891. The results are consistent with the 5-fold cross-validation.

2.2 Sense the Moment is Comparable to Deep learning approaches

Tables 1 and 2 summarise the assessments of Sense the Moment against blind data sets from Loose and Nagarajan, respectively. The first evaluation, taking into account only the training data, slightly overestimated the model, because accuracies on Loose and Nagarajan data sets were 80.77 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 74.51 and 89.47 % on Loose and Nagarajan data sets, respectively. Interestingly, methods based on the hydrophobic moment, Sense the Moment and DBAASP, outperformed the systems based on deep learning algorithms AMP Scanner Vr. 2 and Deep-AmPEP. DBAASP showed the best performance against Loose and Nagarajan data sets, while Sense The Moment was ranked 2nd on Loose (Table 1); and 3rd on Nagarajan data set (Table 2).

Table 1. Prediction assessments of Sense the Moment and its benchmarking against Deep learning methods on Loose data set.

Algorithm	Sensitivity	Specificity	Accuracy	PPV	MCC
Sense The Moment	95.00	65.79	80.77	74.51	0.64
DBAASP	100.00	73.68	87.18	80.00	0.78
Deep-AmPEP	77.50	23.68	51.28	51.67	0.01
AMP Scanner Vr. 2	100.00	36.84	69.23	62.60	0.48

Table 2. Prediction assessments of Sense the Moment and its benchmarking against Deep learning methods on Nagarajan data set.

Algorithm	Sensitivity	Specificity*	Accuracy	PPV	DOR**
Sense The Moment	94.44	0.00	85.00	89.47	2.33
DBAASP	88.89	50.00	85.00	88.89	6.60
Deep-AmPEP	88.89	50.00	85.00	94.12	6.60
AMP Scanner Vr. 2	83.33	0.00	75.00	88.24	0.89

* There are only two negative entries in the set, thus, specificity would be 0, 50 or 100%.
** Because of the high prevalence of positive entries on this data set (90 %), DOR was used as a prevalence independent measurement.

3 Discussion

In the last decade, numerous computational methods to predict the antimicrobial activity of AMPs have been reported.[5,15,16,19,21–23] 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.[17] Because a shuffled sequence has the same scalar properties as the native sequence, the sequences would have the same or similar predictions.[17] 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,[5,15,16] resulting in the compositional bias.[17] This is a problem when genomes and databases are used as a source for such molecules.[24]

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.[19] In this context, there are some properties that change when sequence order is altered, including hydrophobic moment[18] and Chou's pseudo amino acid composition.[25] The iAMP-2L server[26] 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.[17] Currently, deep learning is the state of the art in terms of algorithms for prediction, and such approach is capable of extract data direct from sequences, with no need to convert the sequence into descriptors.[27,28] However, the required number of sequences for training is a limitation for deep learning application to AMPs, and therefore, this could be the main reason because Deep-AmPEP and AMP Scaner Vr. 2 were outperformed by hydrophobic moment-based systems (Tables 1 and

2).

Here, we posited hydrophobic moment as a way to differentiate designed and shuffled sequences, because this property is almost a fingerprint of amino acid order. Thus, assuming α-helical AMPs have higher hydrophobic moment values, shuffled sequences would have their hydrophobic moments reduced. Indeed, it is only in 6.19 % 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 1C).

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.[20] Although hydrophobicity scales are strongly correlated,[29–33] 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. The differences between the scales may be due to reasons such as pH or experimental system. 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. In practical terms, the individual scales have similar predictive power (data not shown). However, the use of the geometric average is important for those sequences which are near the threshold value, because the little differences among the scales could influence the prediction of such sequences (data not shown). Finally, due to the strong correlations between the scales, the choice of scales has only a little contribution on the results.[31] Therefore, we applied the three scales which are commonly used for studying AMPs: Eisenberg,[18] Kyte-Doolittle[34] and Radzicka-Wolfenden.[35,36]

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 some sequences were unable to form α-helixes.[17,37] Therefore, Sense the Moment is restricted to short peptides with low-complexity secondary structures, in particular disulphide-free peptides.

In fact, the system had very good accuracy in both blind sets, considering designed/shuffled (Table 1) and AMP/non-AMP (Table 2). However, the strength of the system relies on its sensitivity, which is what gives the system its name. 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* or even used in combination with other prediction tools.

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. In fact, this could be the reason why DBAASP predictor outperformed Sense the Moment in both datasets (Tables 1 and 2). DBAASP takes into account charge density and depth-dependent potential in addition to the hydrophobic

moment itself. Interestingly, the DBAASP predictor and Sense the Moment were constructed using a different database, which confirms the binary prediction capacity of hydrophobic moment and also validates our approach for the construction of a negative data set (Figure 1A and 1B). However, we have to remember that the hydrophobic moment works best in classifying shuffled and designed sequences, not AMPs themselves, because not all designed sequences from Loose were indeed active against bacteria.[12] In addition there is no direct correlation between hydrophobic moment and antimicrobial activity.[20,21]

Therefore, Sense the Moment could be used together with other prediction systems, such as AmPEP and AMP Scanner, allowing the best characteristics of each system to be exploited, in particular with the 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 solved this bias. 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[38] was used as the primary source of AMP sequences. Initially, the secondary structure of all sequences was predicted using PSIPRED,[39] 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.[40] This set was manually curated, removing incomplete sequences, totalling 388 sequences. For each sequence, three average hydrophobic moments were calculated, using three hydrophobic scales: Eisenberg,[18] Kyte-Doolittle[34] and Radzicka-Wolfenden[35,36] (Figure 1A). The average hydrophobic moment was calculated using Eisenberg's equation:[18]

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

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 are referred 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 (Figure 1A).

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 1B). Ten thousand shuffled versions of each positive sequence were generated. The NS was composed of 388 entries, representing the average of 10,000 geometric hydrophobic moment values of shuffled versions of AMP sequences. The shuffling was performed using a PERL script, available at Git Hub (http://www.github.com/williamfp7/stm). 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).

5.2 Prediction Cut-off Estimation

By using positive and negative sets together, the cut-off value was evaluated in intervals of 0.02, 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.

5.2.1 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 38 shuffled peptides. The second blind data set was derived from Nagarajan et al.[37] (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.

5.3 Assessments against blind sets

The prevalence of both blind sets was measured according to Eq. 2:

$$Prevalence = \left[\frac{Positive\ Entries}{Total\ Entries}\right] \times 100 \quad (2)$$

The prevalence indicates how often AMPs occur in the data set and could be used as an indication of how statistical tests should be used. The model was challenged against the blind data sets and the following parameters were measured:

$$Sensitivity = \left[\frac{TP}{(TP+FN)}\right] \times 100 \quad (3)$$

$$Specificity = \left[\frac{TN}{(TN+FP)}\right] \times 100 \quad (4)$$

$$Accuracy = \left[\frac{(TP+TN)}{(TP+TN+FP+FN)}\right] \times 100 \quad (5)$$

$$PPV = \left[\frac{TP}{(TP+FP)}\right] \times 100 \quad (6)$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt[2]{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \quad (7)$$

$$DOR = \frac{(TP+0.5) \times (TN+0.5)}{(FP+0.5) \times (FN+0.5)} \quad (8)$$

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). Finally, the diagnostic odds ratio (DOR) is a prevalence independent indicator of performance, ranging from 0 to infinity, where values lower than 1 indicate an inverse prediction; values equals to 1, a random prediction; and values higher than 1, a correct prediction, with higher values indicating better discriminatory test performance.[41] Since some divisions by zero may occur, the approximated DOR is calculated by adding 0.5 to all counts.[41]

5.4 Benchmarking

The Sense the Moment performance was compared with the state-of-the-art methods for antimicrobial activity prediction, AMP Scanner Vr. 2[27] and Deep-AmPEP[28], which are based on deep learning, using Eqs. 3-7. In addition, the DBAASP prediction tool,[42] which is based on the hydrophobic moment, charge density and depth-dependent potential, was also assessed against the blind data sets.

5.5 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>.

6 Acknowledgments

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

15. 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).

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

17. 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).

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

19. 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.) vol. 6268 59–62 (Springer Berlin Heidelberg, 2010).

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

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

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

23. 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).

24. 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).

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

26. 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).

27. Veltri, D., Kamath, U. & Shehu, A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* **34**, 2740–2747 (2018).

28. Yan, J. *et al.* Deep-AmPEP30: Improve Short Antimicrobial Peptides Prediction with Deep Learning. *Mol. Ther. - Nucleic Acids* **20**, 882–894 (2020).

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. Simm, S., Einloft, J., Mirus, O. & Schleiff, E. 50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification. *Biol. Res.* **49**, 31 (2016).

31. MacCallum, J. L. & Tieleman, D. P. Hydrophobicity scales: a thermodynamic looking glass into lipid–protein interactions. *Trends Biochem. Sci.* **36**, 653–662 (2011).

32. Peters, C. & Elofsson, A. Why is the biological hydrophobicity scale more accurate than earlier experimental hydrophobicity scales? *Proteins Struct. Funct. Bioinforma.* **82**, 2190–2198 (2014).

33. Zamora, W. J., Campanera, J. M. & Luque, F. J. Development of a Structure-Based, pH-Dependent Lipophilicity Scale of Amino Acids from Continuum Solvation Calculations. *J. Phys. Chem. Lett.* **10**, 883–889 (2019).

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

35. 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).

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

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

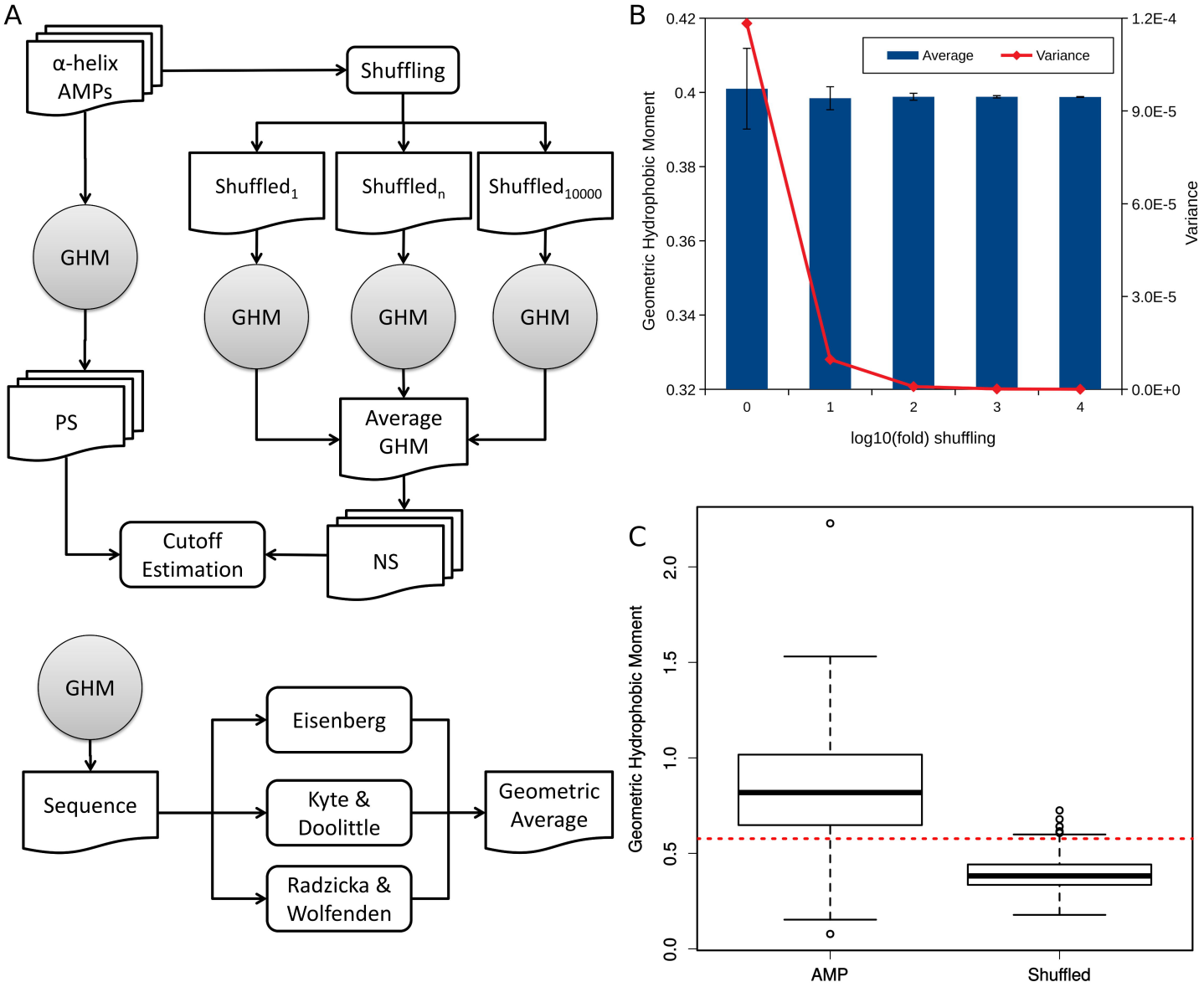
38. Wang, G., Li, X. & Wang, Z. APD2: the updated antimicrobial peptide database and its application in peptide design. *Nucleic Acids Res.* **37**, D933-7 (2009).

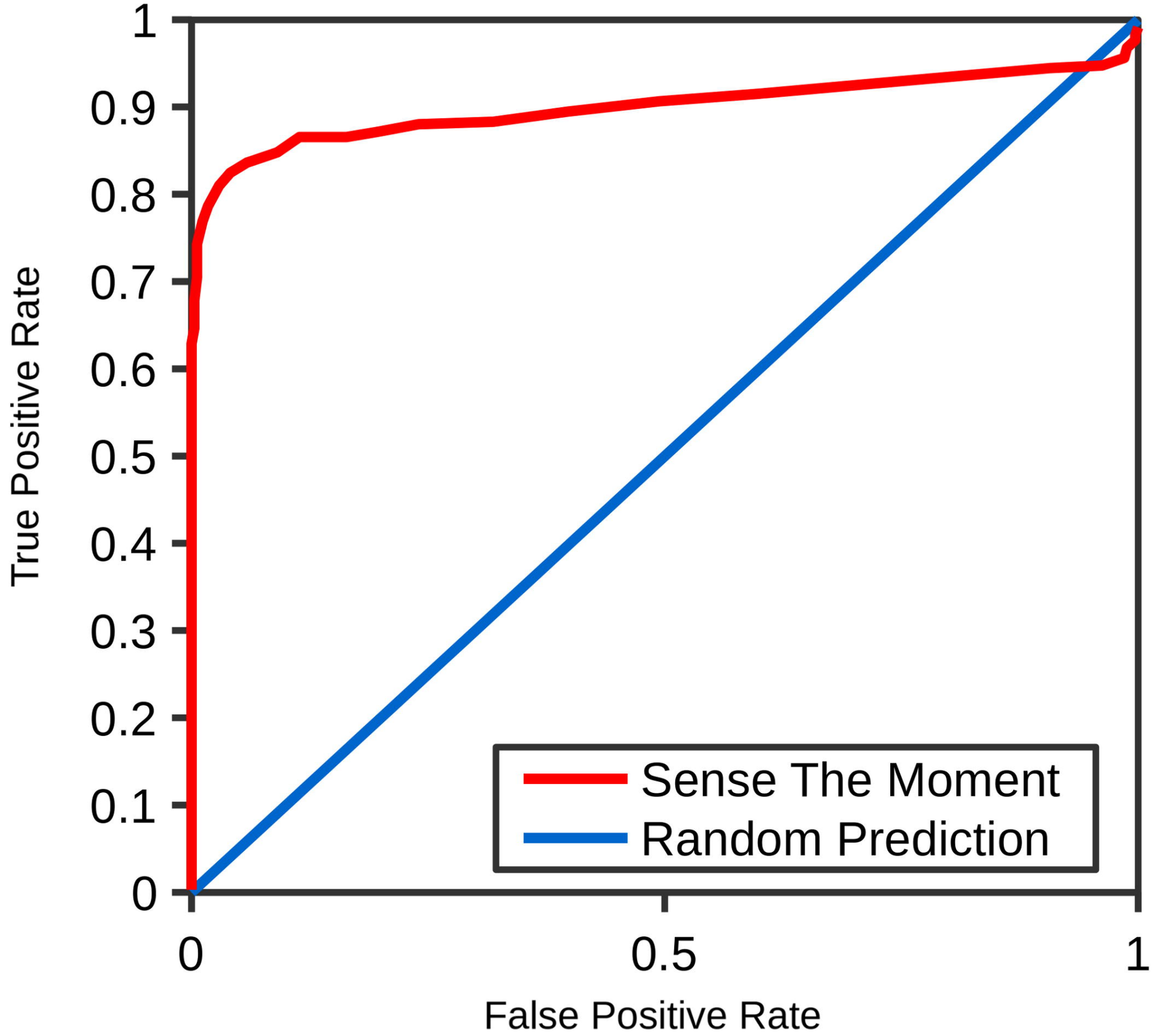
39. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices 1 1Edited by G. Von Heijne. *J. Mol. Biol.* **292**, 195–202 (1999).

40. 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).

41. Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J. & Bossuyt, P. M. M. The diagnostic odds ratio: a single indicator of test performance. *J. Clin. Epidemiol.* **56**, 1129–1135 (2003).

42. Vishnepolsky, B. & Pirtskhalava, M. Prediction of Linear Cationic Antimicrobial Peptides Based on Characteristics Responsible for Their Interaction with the Membranes. *J. Chem. Inf. Model.* **54**, 1512–1523 (2014).


A
α-helix AMPs
Shuffling
Shuffled1
Shuffledn
Shuffled10000
GHM
GHM
GHM
GHM
PS
Average GHM
Cutoff Estimation
NS
GHM
Sequence
Eisenberg
Kyte & Doolittle
Radzicka & Wolfenden
Geometric Average
B
Average
Variance
Geometric Hydrophobic Moment
Variance
log10(fold) shuffling
0.42
0.4
0.38
0.36
0.34
0.32
1.2E-4
9.0E-5
6.0E-5
3.0E-5
0.0E+0
0
1
2
3
4
C
Geometric Hydrophobic Moment
2.0
1.5
1.0
0.5
0.0
AMP
Shuffled


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
True Positive Rate
0
0.5
1
False Positive Rate
Sense The Moment
Random Prediction