SUPPORTING INFORMATION:

Large-Scale Screening of Antifungal Peptides Based on Quantitative Structure-Activity Relationship

Jin Zhang[3], Longbing Yang[1], Zhuqing Tian[1], Wenjing Zhao[1], Chaoqin Sun[1], Lijuan Zhu[1], Mingjiao Huang[1], Guo Guo[1,2,4] and Guiyou Liang[4*]*

1. School of Basic Medical Sciences, Guizhou Medical University, Guiyang 550025, China

2. The Key and Characteristic Laboratory of Modern Pathogen Biology, Guizhou Medical University, Guiyang 550025, China

3. School of Public Health, Guizhou Medical University, Guiyang, 550025, China

4. Translational Medicine Research Center, Guizhou Medical University, Guiyang 550025, China

**Corresponding author. Guo Guo and Guiyou Liang*

Table of contents

Experimental Details

Text S1. Model establishment for antifungal classification……………….………..…S-3

Text S2. Model establishment for antifungal activity prediction………………..…S-6

Text S3. Screening protocol of antifungal peptides……………………….……..…S-8

Text S4. Antifungal peptide screening and experimental validation ………..….…S-10

Tables

Table S1. Peptide descriptors used in this study……….…………..……………S-12

Table S2. Usage of peptide descriptors in antifungal identification………….…S-14

Table S3. Results of antifungal peptide classification models.……...……………S-15

Table S4. Results of the antifungal activity prediction model. ……….………..S-16

Reference ……………………………………………………………....S-17

Text S1. Model establishment for antifungal classification.

Data collection for antifungal peptide classification was conducted. In detail, dataset 1, containing 5775 antifungal peptides and 5775 negative peptides without antifungal activity, was collected. The antifungal peptides were obtained from four antimicrobial databases, *i.e.*, DBAASP[1], APD3[2], DRAMP[3] and CAMP[4] by restricting the activity type to antifungal. These antifungal peptides in the dataset are those reported in the literature and collected in specified antimicrobial databases, but without considering specific activity values. Then, duplicates and peptides with sequence lengths more than 150 or less than 11 amino acid residues were removed. Negative peptides were collected from the manually reviewed peptide dataset (Swiss-Prot section) in UniProt knowledgebase[5] by inputting the query string of “NOT goa:("response to fungus [9620]") existence: "Evidence at protein level [1]" length: [11 TO 150] AND reviewed: yes”. It will ensure that the obtained sequences are of proper length and not antimicrobial peptides. Therefore, the model built based on the dataset can only identify antifungal peptides without considering the degree of activity.

Peptide descriptors were calculated for the whole collected sequences. A total of 9516 descriptors in 11 categories (Table S1) were obtained for each sequence by resorting to python packages of modlamp 4.3.0 [6] and propy 1.0.0a2 [7] with the suggested parameters.

Data preprocessing, including sample partition, descriptor normalization and feature selection, were carried out before calibration. Peptides in dataset 1 were divided into calibration and validation set in a ratio of 4:1 by using Kennard-Stone (KS) algorithm[8] for building the prediction models and validating the model efficiency, respectively. The peptide descriptors were also normalized to avoid variance impact on calibration.

Classification method, support vector machine (SVM) with radial basis function (rbf) kernels, was adopted to build antifungal peptide classification models by using the python package of scikit-learn 0.24.2[9]. The descriptor usage in calibration can be referred to Table S2. Hyper-parameters C and γ were used to control the regularization strength and kernel function scale, respectively. Grid search with ten-fold cross validation was used to optimize the parameters.

Metrics including accuracy, sensitivity (*aka* recall), specificity (*aka* selectively), F1 score, and Matthews correlation coefficient (MCC) were used to evaluate the performance of the built classification models[10], as defined in Eq. (1).

$$
\begin{aligned}
\text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\
\text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}} \\
\text{Specificity} &= \frac{\text{TN}}{\text{FP} + \text{TN}} \\
\text{F1 score} &= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \\
\text{MCC} &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + FP)(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}
\end{aligned}
\qquad (1)
$$

where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative entries predicted by the models, respectively. In addition, the receiver operating characteristic (ROC) curve and area under the ROC curve (AUC) were also adopted to indicate the efficiency of the classifiers by plotting the curve of true positive rate (TPR) against false positive rate (FPR) at various thresholds. The

AUC value is in the range of 0.5-1.0, where 0.5 and 1.0 represent a random and perfect classifier, respectively.

Text S2. Model establishment for antifungal activity prediction.

Minimum inhibitory concentration (MIC) is commonly used to evaluate the activity of isolated peptides by determining the lowest drug concentration that prevents visible microorganisms after overnight incubation. Using four common fungi as the targets, i.e., *Candida albicans*, *Candida krusei*, *Cryptococcus neoformans* and *Candida parapsilosis*, sequences with experimental antifungal activity were picked from dataset 1 to form dataset 2. The information of fungal species was not considered due to the insufficient amount of data. Restricting the activity type to MIC, 1583, 95, 275, and 148 records were remained with targets of *C. albicans*, *C. krusei*, *C. neoformans* and *C. parapsilosis*, respectively. The records with the same sequence but different activity values were kept as independent entries.

For comparability, activity units were unified to μM. Then, MIC was converted to pMIC by a logarithmic transform as defined in Eq. (2). Three benefits might be achieved by the unit transform, such as (1) narrowing the MIC range of multiple orders of magnitude to the pMIC range of one order of magnitude, thus fitting a better model, (2) making a more accurate prediction for the low MIC peptides of interest because the difference between small MIC values will be relatively enlarged, and (3) the prediction error of MIC should be exponential of 2, which is consistent with the fact of step-by-step broth dilution in experimental MIC measurement.

$$\mathrm{pMIC} = -\log_2(\mathrm{MIC}) \qquad (2)$$

Peptides in dataset 2 were divided into calibration and validation set by KS algorithm in a ratio of 4:1. Normalization was also carried out on peptides descriptors.

Feature selection was performed by using variable influence on projection (VIP) method [11].

Support vector regression (SVR) with rbf kernel function was adopted to build the regression models for predicting pMIC values against the four specified fungi. The optimal super-parameters, C and γ, were also determined by the grid search method with ten-fold cross validation. For unknown peptides, pMIC values against the four specified fungi and probability can be predicted by the built models, and then transformed to MIC value for comparison.

Root mean square error (RMSE) and determination coefficient (R^2) were used to assess the prediction performance of the developed models.

Text S3. Screening protocol of antifungal peptides.

With the established models, a stepwise protocol for large-scale antifungal peptide screening was integrated:

(1) For a candidate sequence, if the prediction of antifungal classification is true, proceed to the next step, else drop it.

(2) If the predicted MIC value against *C. albicans* is smaller than 32 μM and probability larger than 50%, proceed to the next step; else, drop it.

(3) If the predicted MIC value against *C. krusei* in step 3 is smaller than 32 μM and probability larger than 50%, proceed to the next step; else, drop it.

(4) If the predicted MIC value against *C. neoformans* is smaller than 32 μM and probability larger than 50%, proceed to the next step; else, drop it.

(5) If the predicted MIC value against *C. parapsilosis* is smaller than 32 μM and probability larger than 50%, keep it and continue the next loop of screening; else, drop it. The above steps are repeated until all candidates are screened.

Many potential sequences may be obtained after a step-by-step screening. A final ranking of the screened peptides is still required to select the outperforming *N* sequences for further investigation. Antifungal index (AFI) was firstly defined for comprehensively assessing the antifungal ability against the considered fungi, as Eqs. (3):

$$\begin{aligned} \mathrm{AFI} &= 2^{-1 \times \frac{\mathrm{pMIC}_{Ca} + \mathrm{pMIC}_{Ck} + \mathrm{pMIC}_{Cn} + \mathrm{pMIC}_{Cp}}{4}} \\ p_{\mathrm{AFI}} &= p \times p_{Ca} \times p_{Ck} \times p_{Cn} \times p_{Cp} \end{aligned} \tag{3}$$

where $\mathrm{pMIC}_{\mathrm{Ca}}$, $\mathrm{pMIC}_{\mathrm{Ck}}$, $\mathrm{pMIC}_{\mathrm{Cn}}$, $\mathrm{pMIC}_{\mathrm{Cp}}$ and p_{Ca}, p_{Ck}, p_{Cn}, p_{Cp} are predicted pMIC values and the corresponding probability against *C. albicans*, *C. krusei*, *C. neoformans* and *C. parapsilosis*, respectively, and p_{AFI} is the probability of the calculated AFI. In fact, the AFI is to some extent the average of all predicted antifungal activities, which is consistent with the fact that most broad-spectrum antifungal peptides show similar bioactivity. A smaller AFI and p_{AFI} suggest a more promising broad-spectrum antifungal peptide. In this study, the AFI threshold is set to 3 μM to identify prominent antifungal peptides.

The screened peptides are sorted according to AFI from smallest to largest, and the three top-ranking peptides are chosen for further experimental validation.

Text S4. The details about chemical synthesis of the screened antifungal peptides.

With the proposed protocol, a demonstrative screening application was conducted on peptides from the UniProt knowledgebase with sequence lengths in the range of 11-75 amino acid residues. The screening was performed on a personal computer (Ubuntu 20.04, 3.00 GHz Intel i9-10980XE, 32G×4 memory).

Chemical synthesis was employed for these screened peptides. In detail, the screened peptides were synthesized using standard 9-fluorenylmethoxycarbonyl (Fmoc) solid phase peptide synthesis. All syntheses were performed at room temperature under nitrogen bubbling. The Fmoc resin was firstly deprotected twice one minute and four minutes using a deprotection cocktail containing 20% piperidine in N,N-dimethylformamide (DMF). For each amino acid, a coupling was performed using three times of the corresponding Fmoc protected amino acid, three times benzotriazol-1-yloxytripyrrolidinophosphonium hexafluorophosphate (PyBOP), and six times N,N-Diisopropylethylamine (DIEA) in DMF. Deprotection steps (double deprotection, five minutes, and ten minutes) were achieved using the same cocktail described above. After the last deprotection, peptides were cleaved from the resin using 10 mL of a mixture of trifluoroacetic acid/triisopropylsilane/mQ water (TFA/TIS/H_2O) with the corresponding ratios 94/5/1 during three hours. Peptides were then precipitated using approximately 25 mL of cold ethyl ether and centrifuged 10 minutes at 4400 rpm. Supernatant was removed and peptides were washed twice with15 ml of cold ethyl ether before lyophilization.

Experimental validation was performed by measuring their minimum inhibitory concentration (MIC) against *C. albicans* SC5314, *C. krusei* IFM56881, *C. neoformans* H99, and *C. parapsilosis* ATCC22019. In detail, a single actively growing microbial colony was inoculated into 5 ml sterile SDB medium and incubated overnight at 37°C. The turbidity of the fungal solution was adjusted to 1 - 5×10^6 colony forming units (CFU)/ml using a blood cell counting plate. The fungal suspension was then diluted with sabouraud dextrose broth (SDB) to 0.5–2.5 × 103 CFU/ml. An aliquot of 100 µl of the final suspension was added into each well of a sterile 96-well plate containing 100 µl of medium containing antimicrobial agents at double-diluted concentrations. Phosphate-buffered saline (PBS) was used as a negative control and fluconazole as a positive control. The plate was assessed for MIC values after 24 h or 48 h of incubation at 37°C. The MIC value was determined to be the minimum concentration at which microscopic growth could not be observed by the naked eye, as recommended by the Clinical Laboratory and Standards Institute CLSI (2008) methods. The experiment was repeated three times, three biological replicates at a time. The units of experimental MIC were converted from µg/ml to µM to facilitate comparison with prediction results.

Table S1. Peptide descriptors used in this study**.**

Categories	Descriptor category	No. descriptors	Descriptor name[a]
1	common descriptor	9	Length, ChargeDensity, Isoelectric point, InstabilityInd, Aromaticity, AliphaticInd, BomanInd, HydRatio
2	amino acid composition	20	A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, V
3	dipeptide composition	400	AA, AR, …, YV, VV
4	tripeptide composition	8000	AAA, AAR, …, WVV, YVV
5	composition transition distribution	147	_PolarizabilityC1~3, _SolventAccessibilityC1~3, _SecondaryStrC1~3, _ChargeC1~3, _PolarityC1~3, _NormalizedVDWVC1~3, _HydrophobicityC1~3, _PolarizabilityT12~23, …
6	Geary autocorrelation	240	GearyAuto_Hydrophobicity1~30, GearyAuto_AvFlexibility1~30, GearyAuto_Polarizability1~30, GearyAuto_FreeEnergy1~30, GearyAuto_ResidueASA1~30, GearyAuto_ResidueVol1~30, GearyAuto_Steric1~30, GearyAuto_Mutability1~30
7	Moran autocorrelation	240	MoranAuto_Hydrophobicity1~30, MoranAuto_AvFlexibility1~30, MoranAuto_Polarizability1~30, MoranAuto_FreeEnergy1~30, MoranAuto_ResidueASA1~30, MoranAuto_ResidueVol1~30, MoranAuto_Steric1~30, MoranAuto_Mutability1~30
8	Normalized Moreau-Broto autocorrelation	240	MoreauBrotoAuto_Hydrophobicity1~30, MoreauBrotoAuto_AvFlexibility1~30, MoreauBrotoAuto_Polarizability1~30, MoreauBrotoAuto_FreeEnergy1~30, MoreauBrotoAuto_ResidueASA1~30, MoreauBrotoAuto_ResidueVol1~30, MoreauBrotoAuto_Steric1~30, MoreauBrotoAuto_Mutability1~30
9	Type I Pseudo amino acid composition	30	PAAC1~30
10	Quasi sequence order	100	QSOSW1~50, QSOgrant1~50

11	Sequence order coupling numbers	90	tausw1~45, taugrant1~45

[a] More information about these descriptors refers to the website of modlamp (https://pypi.org/project/modlamp/) and propy (https://pypi.org/project/propy3/).

Table S2. Usage of peptide descriptors in antifungal identification.

Descriptor category	No. descriptors	Used descriptors	Usage percentage %
Common descriptor	9	2	22.2
Amino acid composition	20	3	15.0
Dipeptide composition	400	80	20.0
Tripeptide composition	8000	1611	20.1
Composition transition distribution	147	39	26.5
Geary autocorrelation	240	31	12.9
Moran autocorrelation	240	46	19.2
Moreau-Broto autocorrelation	240	20	8.3
Type I Pseudo amino acid composition	30	7	23.3
Quasi sequence order	100	20	20.0
Sequence order coupling numbers	90	2	2.2

Table S3. Results of the antifungal peptide classification model.

Sequence lengths	Calibration					Validation				
(Sample size)	Acc.	Sen.	Spec.	F1	MCC	Acc.	Sen.	Spec.	F1	MCC
All (*n*=9240, 2310)[a]	0.95	0.95	0.95	0.95	0.90	0.89	0.90	0.89	0.89	0.79
≤50 (*n*=7874, 2162)	0.94	0.94	0.94	0.94	0.89	0.89	0.90	0.89	0.89	0.78
>50 & ≤ 100 (*n*=1189, 141)	0.98	0.98	0.99	0.98	0.97	0.91	0.90	0.92	0.90	0.82
> 100 (*n*=177, 7)	0.99	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00	1.00

[a] Sample size of calibration and validation set, respectively.

Table S4. Results of antifungal activity prediction models.

Targets	Calibration		Validation	
	RMSE	R^2	RMSE	R^2
C. albicans (*n*=1266, 317)[a]	0.69	0.90	1.23	0.66
C. krusei (*n*=76, 19)	0.48	0.94	1.10	0.69
C. neoformans (*n*=220, 55)	0.82	0.90	0.89	0.89
C. parapsilosis (*n*=118, 30)	0.73	0.90	1.17	0.69

[a] Sample size of calibration and validation set, respectively.

Reference

(1) Pirtskhalava, M.; Amstrong, A. A.; Grigolava, M.; Chubinidze, M.; Alimbarashvili, E.; Vishnepolsky, B.; Gabrielian, A.; Rosenthal, A.; Hurt, D. E.; Tartakovsky, M., DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **2020,** *49* (D1), D288-D297.

(2) Wang, G. S.; Li, X.; Wang, Z., APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **2015,** *44* (D1), D1087-D1093.

(3) Kang, X. Y.; Dong, F. Y.; Shi, C.; Liu, S. C.; Sun, J.; Chen, J. X.; Li, H. q.; Xu, H. M.; Lao, X. Z.; Zheng, H., DRAMP 2.0, an updated data repository of antimicrobial peptides. *Sci. Data* **2019,** *6* (1), 148.

(4) Waghu, F. H.; Barai, R. S.; Gurung, P.; Idicula-Thomas, S., CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res.* **2015,** *44* (D1), D1094-D1097.

(5) Consortium, T. U., UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2020,** *49* (D1), D480-D489.

(6) Müller, A. T.; Gabernet, G.; Hiss, J. A.; Schneider, G., modlAMP: Python for antimicrobial peptides. *Bioinformatics* **2017,** *33* (17), 2753-2755.

(7) Cao, D. S.; Xu, Q. S.; Liang, Y. Z., propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* **2013,** *29* (7), 960-962.

(8) Kennard, R. W.; Stone, L. A., Computer aided design of experiments.

Technometrics **1969,** *11* (1), 137-148.

(9) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V., Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011,** *12*, 2825-2830.

(10) Lever, J.; Krzywinski, M.; Altman, N., Classification evaluation. *Nat. Methods* **2016,** *13* (8), 603-604.

(11) Wold, S.; Sjöström, M.; Eriksson, L., PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001,** *58* (2), 109-130.