Prediction of Antimicrobial Peptides Based on the Adaptive Neuro-Fuzzy Inference System Application

Fabiano C. Fernandes,[1] Daniel J. Rigden,[2] Octavio L. Franco[1]

[1]*Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia, UCB, Brasília, DF, Brazil*

[2]*Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom*

Received 15 January 2012; revised 27 March 2012; accepted 29 March 2012
Published online 9 April 2012 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/bip.22066

ABSTRACT:

Antimicrobial peptides (AMPs) are widely distributed defense molecules and represent a promising alternative for solving the problem of antibiotic resistance. Nevertheless, the experimental time required to screen putative AMPs makes computational simulations based on peptide sequence analysis and/or molecular modeling extremely attractive. Artificial intelligence methods acting as simulation and prediction tools are of great importance in helping to efficiently discover and design novel AMPs. In the present study, state-of-the-art published outcomes using different prediction methods and databases were compared to an adaptive neuro-fuzzy inference system (ANFIS) model. Data from our study showed that ANFIS obtained an accuracy of 96.7% and a Matthew's Correlation Coefficient (MCC) of 0.936, which proved it to be an efficient model for pattern recognition in antimicrobial peptide prediction. Furthermore, a lower number of input parameters were needed for the ANFIS model, improving the speed and ease of prediction. In summary, due to the fuzzy nature of AMP physicochemical properties, the ANFIS approach presented here can provide an efficient solution for screening putative AMP sequences and for exploration of properties characteristic of AMPs. © 2012 Wiley Periodicals, Inc. Biopolymers (Pept Sci) 98: 280–287, 2012.

Keywords: antimicrobial peptides; rational prediction; sequence analysis; artificial intelligence; adaptive neuro-fuzzy inference system (ANFIS)

This article was originally published online as an accepted preprint. The "Published Online" date corresponds to the preprint version. You can request a copy of the preprint by emailing the Biopolymers editorial office at biopolymers@wiley.com

INTRODUCTION

Antimicrobial peptides (AMPs) are natural antibiotics considered to be the first line of defense for the majority of living organisms, acting against pathogenic invasion, showing additional anticancer activity, immune system stimulation, inflammatory responses, and several other actions.[1–3] They have been commonly isolated from numerous species in different kingdoms.[3] AMPs can have diverse roles as promiscuous defense peptides with multiple cellular targets.[4] Some plant storage proteins can act as AMPs for plant defense and their efficiency depends on their molecular mass, amino acid sequence, charge, conformation, secondary and tertiary structures, disulfide bonds, and hydrophobicity.[5] AMPs usually have 12–100 amino acid residues, are positively charged at physiological pH and are amphipathic.[6,7] These peptides possess a wide range of secondary structures with two to four β-strands, amphipathic α-helices, loop structures, or

Additional Supporting Information may be found in the online version of this article.

Correspondence to: Octávio Luiz Franco, Centro de Análises Proteômicas e Bioquímicas, Pós-Graduação em Ciências Genômicas e Biotecnologia. SGAN 916 Av. W5 Norte, Modulo C, Sala 219, CP 70790-160, Brasilia, DF, Brazil; e-mail: ocfranco@gmail.com

extended structures.[2,6,8] Most of them may kill bacteria by lipid bilayer disruption, making them promising compounds as substitutes for conventional antibiotics. Factors to be considered during the design of novel AMPs include function prediction and optimization of parameters such as toxicity against mammalian cells, allergenicity, selectivity between microbial and human cells, and manufacturing cost.[3,9,10]

The number of possible amino acid sequences makes it hard to discover new AMPs through exclusively experimental testing. For an AMP sequence of n amino acid length, there are 20^n possible sequences to be tested, typically from 20^{12} to 20^{100}, rendering a brute force approach impracticable.[11] Many research groups are therefore seeking tools that accurately predict *in silico* peptide activity, with high-throughput, to provide specific and narrow sets of candidate structures for clinical evaluation.[12]

The advent of antimicrobial peptide databases has made it easier for scientists to understand and rationally design AMPs. In 2004 the CAMEL database[13] employed quantitative structure–activity relationship (QSAR) and artificial neural networks (ANN) to predict AMP function. In 2007, the AMPer database[14] provided hidden Markov models (HMMs) to predict individual classes of AMPs. In the same year Antibp[15] used ANN, quantitative matrices (QM), and support vector machine (SVM) to predict antimicrobial activity and used non-secretory proteins validated from Mitpred[16] algorithm as negative dataset. In 2009 the updated antimicrobial peptide database (APD2)[17,18] demonstrated that in AMPs some specific amino acid residues are more common than others. Its prediction tool is based on sequence similarity and certain known principles of AMPs. In the same year the so-called RANDOM database[19] provided *in silico* screening for AMPs on two random libraries of 1400 peptides with three-dimensional (3D) QSAR and ANN models as prediction tools. In 2010 the manually curated collection of antimicrobial peptides (CAMP)[3] database provided prediction tools based on random forests (RF), discriminant analysis (DA), and SVM and used the CD-Hit[20] clustering algorithm in order to eliminate sequences with >90% of identity in the negative dataset. In 2010 the AntiBP2[21] database, took its positive dataset from APD2[17] and its negative dataset from MitPred,[16] selecting all peptides from intracellular locations except the secretary proteins, and provided an SVM-based model as a prediction tool. The AMSDb (http://www.bbcm.univ.trieste.it/~tossi/amsdb.html) covers the sequences of gene-encoded AMPs and proteins from animal and plant sequences and provides no prediction tools. Additionally, other databases[22–24] and methodologies such as QM[21] in 2010, and weighted finite-state transducers[11] in the same year, have allowed the computational analysis of AMPs, aiming to accelerate and to rationalize the process of drug discovery and design.[25,26] However, none of the aforementioned methods and databases have employed hybrid methods that combine ANN and fuzzy inference systems such as the adaptive neuro-fuzzy inference system (ANFIS).[27]

The present study employed the ANFIS as a pattern recognition tool to classify a putative peptide as an antimicrobial peptide or non-antimicrobial peptide. Input selection heuristics for data dimensionality reduction[28] and a semisupervised k-means clustering for outlier removal[29] were chosen as the preprocessing phase. The APD2[17] database was chosen due to its links to PDB (http://www.rcsb.org)[30] database, which will allow further studies with primary and tertiary[28,29] peptide structures and their antimicrobial function.

MATERIAL AND METHODS

Datasets

For the construction of the positive antimicrobial peptide dataset, the APD2 database[17] was selected as a primary antimicrobial peptide database source. From APD2, only the AMPs with resolved structures (with corresponding PDB codes) and sequence length from 10 to 100 amino acids were selected, containing 149 tested AMPs with an average sequence length of 35 residues. The next step was clustering with CD-Hit,[20] with sequence identity cut-off of 50% in order to remove sequence redundancy in the set, resulting in 115 clusters. All the Fasta files were used to compute the peptide physicochemical properties according to the methods described by Torrent et al.,[31] such as isoelectric point (pI) from Expasy web server,[32] α-helix propensity, β-sheet propensity, turn structure propensity, and *in vitro* aggregation from Tango software,[33–35] *in vivo* aggregation propensity from AGGRESCAN,[36] while the hydrophobic mean was calculated from Gravy scale[37] and peptide length from a simple Perl script (Supporting Information Figure S1 and Table STI). Because no antimicrobial peptide negative dataset was available, a methodology was designed to select and test the sequences that did not act as an antimicrobial peptide (Supporting Information Figure S2). The PDB[30] was selected as a primary peptide database source file, and the experimental method of X-ray or NMR, polymer type of non-DNA and non-RNA, and non-mixed and sequence length from 10 to 100 amino acids was used, resulting in 1195 sequences. The next step was to remove predicted membrane or extracellular proteins using the Phobius web server,[38] which resulted in 642 predicted intracellular sequences. Redundancy removal with CD-Hit[20] as above resulted in 116 clusters (Supporting Information Table STII).

Table I A Two-Tailed Unpaired *t*-Test Results of the Negative and Positive AMP Datasets

$t_c = 1.960$ $\alpha = 0.05$	*In Vitro* Aggregation	Turn Structure Propensity	α-Helix Propensity	β-Sheet Propensity	Isoelectric Point	Length	Hydrophobic Mean	*In vivo* Aggregation
s^2	325208.223	292.781	24581.629	5184.845	4.941	421.082	458981.996	529.051
t	−13.572	−0.873	−4.248	−5.854	3.990	−7.674	−6.893	−9.442
Result	Reject null hypothesis	Accept null hypothesis	Reject null hypothesis	Reject null hypothesis	Reject null hypothesis	Reject null hypothesis	Reject null hypothesis	Reject null hypothesis
P Value	1.22384E −27	0.435	3.95641E −05	1.71669E −08	8.86602E −05	5.61515E −13	5.31341E −11	1.19648E −17

The null hypothesis stated that the average deviation between the pairs of sets is equal to zero.

With the positive and negative AMP datasets, along with the corresponding physicochemical characteristics, a two-tailed unpaired *t*-test analysis with a confidence interval of 95% was carried out. Only the turn-propensity parameter was not observed to be independent between the two databases; its *p*-value of 0.435 was greater than the significance level of 0.05, and this characteristic was therefore not considered in this study (Table I).

Classification Techniques

An ANN was constructed for comparison of performance with the ANFIS procedure. The Matlab R2010a (The MathWorks Inc., Natick, USA) software was used to build the ANN and the ANFIS as prediction tools. The ANN created is a two-layer feed-forward network with sigmoid hidden and output neurons. The ANN was trained with scaled conjugate gradient back propagation, and the hidden layer was built with 50 neurons. The training dataset is composed of the physicochemical features of 115 peptides and 58 peptides each for validation and testing sets, all randomly divided.

The ANFIS was created using two trapezoidal membership functions and trained in 10 epochs. Semisupervised *k*-means clustering for outlier detection[29] was applied to the positive and negative AMPs dataset as a pre-processing phase to improve performance. The input selection method was done using the heuristics described by Jang,[28] where only the *in vitro* aggregation together with peptide length demonstrated the smallest RMSE between the network outputs and targets after one epoch of training and therefore were used for training, testing and validating purposes (Supporting Information Table STIII, Figure 1), whereas all other physicochemical parameters were discarded from the ANFIS model. The radar graph (Figure 1) shows the ANFIS initial RMSE on the radial axis (varying from 0.0 to 0.45) and the two-by-two physicochemical features-ordered pairs (*in vitro* aggregation together with peptide length) along the circumference. The (1, 5) ordered pairs resulted in the smallest RMSE between the network outputs and targets.

RESULTS

ANN Experiment Comparison

The two datasets together contained 231 peptide sequences (115 AMPs and 116 non-AMPs). All the antimicrobial peptide sequences were assigned the value 1 and the non-AMPs sequences were assigned the value 0 for network training validation and testing purposes. The ANN model was evaluated with 115 randomly divided peptide sequences for training dataset and 58 randomly divided peptide sequences for validation and testing datasets. The resulting receiver-operating curve (ROC) (Figure 2a) and confusion plot (Figure 3a) demonstrate that even without the turn-propensity parameter, using only seven physicochemical features extracted from the peptide sequences dataset and using a novel negative dataset

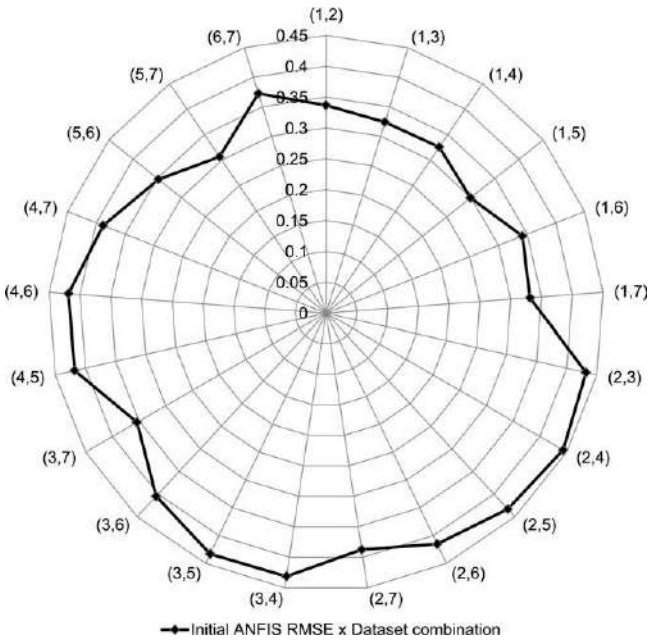


FIGURE 1 Input selection for initial ANFIS RMSE and dataset combination heuristics radar graph used for preprocessing phase, where the pair (1, 5) represents only *in vitro* aggregation together with peptide length.

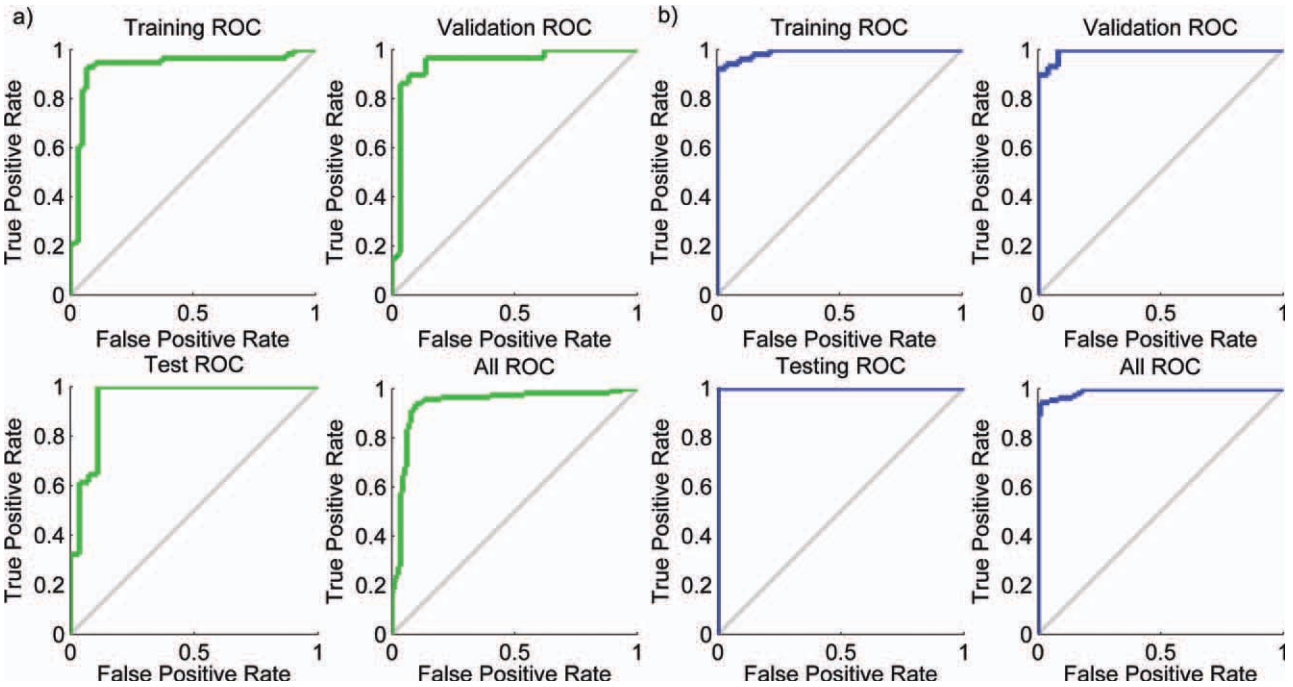


FIGURE 2 (a) ROC curves for the training, validation, testing, and global datasets showing the performance of the ANN method. (b) ROC curves for the training, validation, testing, and global datasets showing the performance of the ANFIS method.

construction methodology, the overall accuracy of the method was 90.9% and the overall Matthew's correlation coefficient (MCC) was 0.8206. The MCC for the training matrix was 0.8424, for the validation matrix 0.7685 and for the testing matrix 0.8300, which means that the results did not show overfitting. The ROC curves properly sorted the peptides into the two groups and tended to maximize the AuC (area under the ROC curve) which measures the test accuracy.

The ANN prediction tool showed very good sensitivity (a measure of how predictors or tests find true positive instances) and specificity (a measure of discovery of true negatives). The 231 peptide sequences trained the ANN, and this network can be used to test if a putative antimicrobial peptide may be synthesized and further *in vitro* evaluated against microorganisms.

ANFIS Experiment Data

The ANFIS model was then tested to see if it showed improved peptide classification ability compared to the simple ANN (above). The number of membership functions (MF) was varied, but the best accuracy was obtained with two MF (data not shown). The results obtained with two MF are shown in Figure 4.

The type of MF was varied between triangular, Gaussian, trapezoidal and sigmoidal functions, and of these the trapezoidal function produced the highest accuracy (data not shown). The best epoch number was 10. The resulting ROC (Figure 2b) and the confusion plot (Figure 3b) demonstrated that with only two physicochemical features the ANFIS network obtained better results than the ANN model with seven physicochemical features (Table II). The *in vitro* aggregation, together with peptide length, are major determinants in AMP prediction using the ANFIS model. The AMPs tended to minimize *in vitro* aggregation and presented lower peptide average length than did non-AMPs not by chance (Supporting Information Tables STI and STII). Not only the specific accuracy but also the AuC of the ROC produced better results if compared to the ANN model. All models showed good sensitivity and specificity (Figures 2a, 3a, 2b, and 3b).

The ANFIS model presented an even better performance than the ANN experiment. There is therefore a higher chance of success in verifying the function of molecules. This, in turn, will result in a better use of resources and more benefits to the pharmaceutical industry.

DISCUSSION

For the construction of the non-AMP dataset the use of the Uniprot database (http://www.uniprot.org)[39,40] was rejected, because its member peptides may not have been specifically tested as AMPs, so even selecting only the sequences not

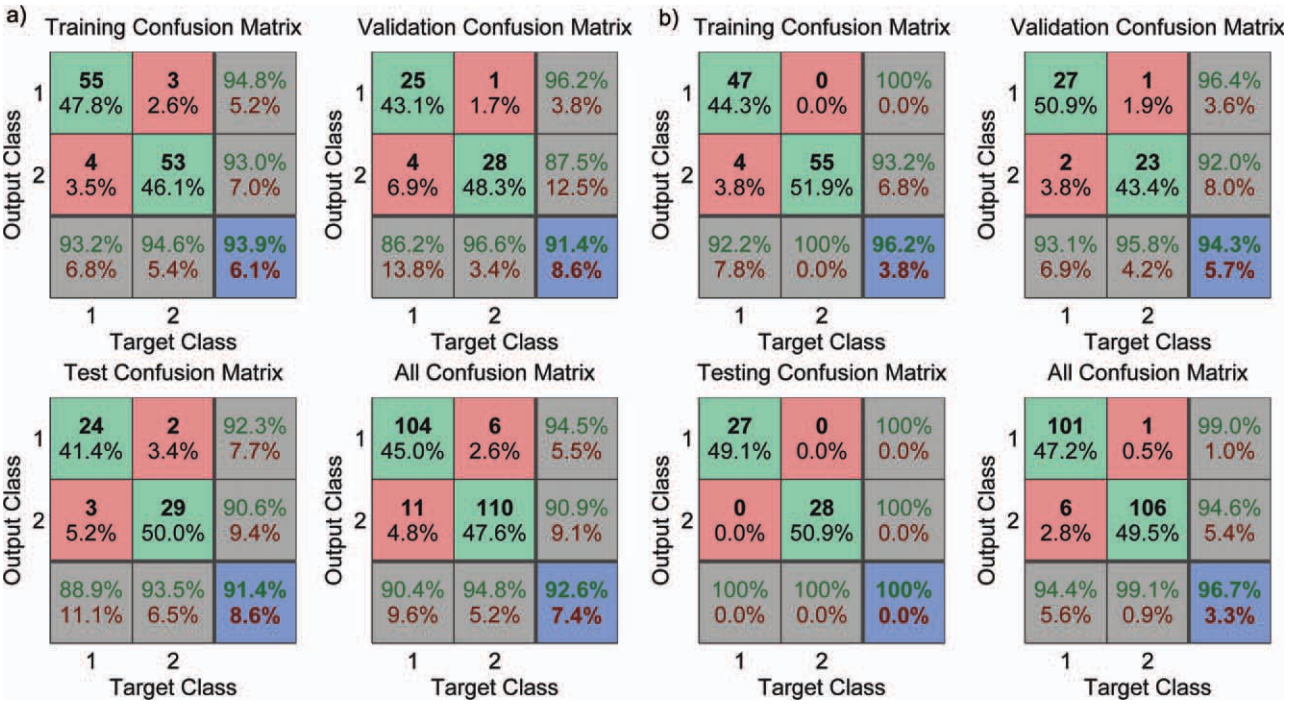


FIGURE 3 (a) Confusion plot for the training, validation, testing and global datasets showing the performance of the method comparing data proposed (ANN). The plot shows how well the ANN predictor separates antimicrobial peptides from nonantimicrobial peptides (b) Confusion plot for the training, validation, testing, and global datasets showing the performance of the ANFIS method. The plot shows how well the ANFIS predictor separates antimicrobial peptides from nonantimicrobial peptides. In this confusion matrix, the ordered pairs (output,target) with the values (1,1) are true positives, (1,2) are false positives, (2,1) are false negatives and (2,2) are true negatives. The blue square indicates the accuracy.

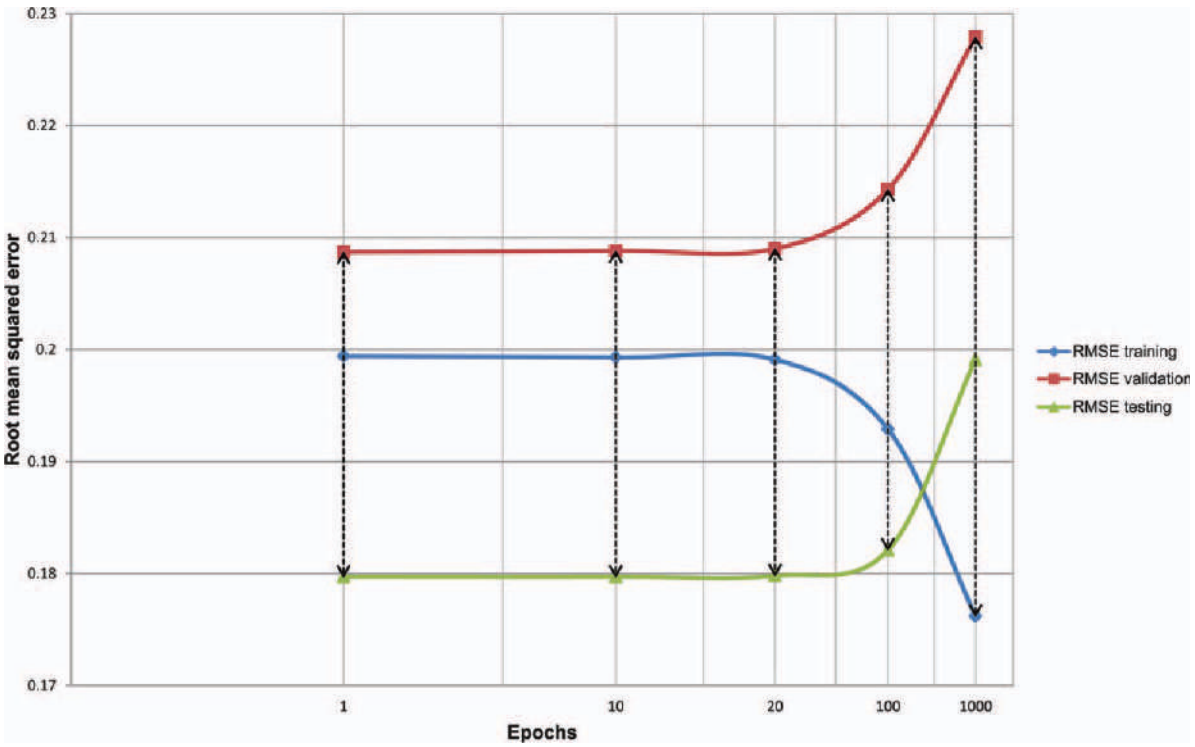


FIGURE 4 RMSE curves for ANFIS. The best ANFIS performance was obtained with two trapezoidal membership functions. The dotted lines represent the maximum and minimum values.

Table II Comparison Between Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference System (ANFIS) as AMP Prediction Tools

	ANN		ANFIS	
Datasets	RMSE	MCC	RMSE	MCC
Training	0.2591	0.8784	0.1991	0.9269
Validation	0.2663	0.8320	0.2090	0.8868
Testing	0.2853	0.8268	0.1798	1.0000

reported as AMPs or not reported as toxic there is a risk of imprecision. The methodology used in this study to build the antimicrobial peptide negative dataset used, instead, the Phobius[38] a combined signal and transmembrane predictor as the main database filter to ensure intracellular localization and thus a presumed nonantimicrobial function.[15,21,41] It was assumed that an antimicrobial peptide may act as pore forming on the cell membrane, therefore not acting as a signal peptide or transmembrane peptide.[41] The resulting nonantimicrobial peptide dataset and the antimicrobial peptide dataset become more disjoint sets and therefore present smaller *p*-values.

The present results are compared to previous work in Table III. Important differences in sequence selection and methodology are now outlined. The AntiBP2 database used random peptides from proteins belonging to intracellular location[16] assuming that the AMPs are secreted outside the cell[42] reducing the odds of antimicrobial function and presented a MCC of 0.843. The CAMP[3] database randomly selected from Uniprot[39,43] the non-secretory proteins without the annotation of "antimicrobial" and presented a MCC of 0.86. The CAMEL[13] database utilized inductive QSAR descriptors to calculate antibacterial potency and submitted them to the ANN presenting a MCC of $\sim$0.6 (80% of correct classification). The AMPER[14] database employed a Blastp algorithm from the BLAST tool[44] and a HMM to correctly assign the peptides in clusters; hence, it does not use a negative dataset with up to 99% of accuracy and a MCC of about 0.98. The Random database used the top 5% of relative inhibitory concentration at 50% (IC50) relative to control peptide as a positive dataset; the other 95% was used as a negative dataset with up to 94% of accuracy and a MCC of about 0.88 (Table III). Both Random and AMPER[14] present high-standard deviation values. The present results compare very well to published methods. The ANFIS model presented an overall accuracy of 96.7%, a RMSE of 0.1797 and a MCC of 0.9356. These results were compared to the results presented by the ANN with the same datasets (Table II) and showed a better performance (96.7% vs. 90.9%) for the ANFIS model employed in this study, confirming the latter as an outstanding pattern recognition tool. The ANN approach has good predictive power but the heterogeneous nature of the information used might have trained the ANN in local minima. This is one consideration that makes the ANFIS approach, with its fuzzy layer, such a relevant and promising approach.

The number of potential input features for an AMP prediction system is very large and the choices of how to select appropriate physicochemical features for antimicrobial prediction and build the negative dataset, even with enough class separability measures, have a tremendous impact on prediction accuracy and MCC. One can separate classes of characteristics not only related to antimicrobial function but to other physicochemical functions, resulting in higher accuracy but also in a high standard deviation or an overfitting to a specific dataset. The search for new antimicrobial descriptors that are perhaps related to secondary structure properties could help to minimize the broad variation in prediction algorithms' accuracy and to define precisely the real importance of features linked to antimicrobial functions.

From the eight selected physicochemical parameters, seven showed statistically significant differences between positive and negative datasets. They are efficient antimicrobial peptide descriptors and lead to good accuracy using the ANN method if compared to the main algorithms for AMP prediction (Table III). The ANFIS approach used semisupervised *k*-means clustering for outlier detection[29] and a heuristic solution to reduce the dimensionality of the input data, based on the first epoch training error, to select the best pairs of input data, discarding all the others.[28] The best physicochemical pairs obtained were *in vitro* aggregation and peptide length, possibly because they presented smaller *p*-values and together reached smaller local minima in network training. It obtained the highest accuracy and had the added

Table III Comparison Between ANFIS and Other Algorithms and Databases

	MCC			
Algorithm	Training Dataset	Validation Dataset	Testing Dataset	Database
HMM		$\sim$0.98 (overall)		AMPER
		$\sim$0.88 (overall)		RANDOM
ANN		$\sim$0.6 (overall)		CAMEL QSAR
DA	0.75	–	0.74	CAMP
RF	0.86	–	0.86	CAMP
SVM	0.88	–	0.82	CAMP
SVM	–	–	0.84	AntiBP2
ANFIS		0.94 (overall)		APD2
ANN		0.85 (overall)		APD2

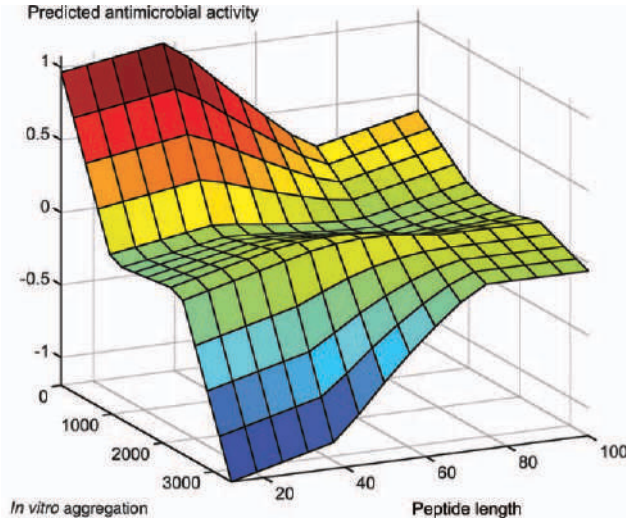


FIGURE 5 Input–output mapping surface for antimicrobial activity with peptide length and *in vitro* aggregation as input parameters.

advantage of being faster than the main algorithms available, and also used the same dataset if compared to an ANN model (Tables II and III). By using a hybrid learning procedure, the fuzzy inference system utilized fuzzy if-then rules to capture the imprecise mode of reasoning and membership functions for the "fuzzification" step to be able to model properly the system's ill-defined boundaries and generate input–output mapping for antimicrobial prediction. The fuzzy if-then rules (data not shown) and the input–output mapping surface (Figure 5) demonstrated the model's ability to describe the behavior of a complex system. In Figure 5 the vertical axis shows the predicted antimicrobial activity. The yellow to red surface areas indicate where the tool predicts that the sequence is an antimicrobial peptide. These areas are associated with smaller predicted *in vitro* aggregation values and peptide lengths more commonly less than around 40 amino acids. In the largest peptide size range, the aggregation propensity has much less impact on the predicted antimicrobial activity than at smaller peptide size range.

The *in vitro* aggregation index extracted from TANGO's AGG parameter[33–35] was first used in antimicrobial peptide prediction and classification by Torrent et al.[31] with 89.2% of prediction overall accuracy. The TANGO algorithm[33–35] measures the tendency for β-sheet aggregation according to a Boltzmann distribution from a phase-space encompassing the structural states of random coil, β-turn, α-helix, β-sheet aggregation, and α-helix aggregation. For β-sheet aggregation, the TANGO algorithm[33–35] assumed full buried on aggregation and only one face buried for α-helix aggregation. In this study, only the *in vitro* aggregation index and the peptide length were enough as input parameters to an ANFIS network to predict antimicrobial activity with an overall accuracy of 96.7%.

The *in vitro* aggregation process may act as a binary barrier for short peptides (from 10 to 50 amino acids) breaking their antimicrobial function as a result of the molecules aggregation. The left side of Figure 5 indicates that the prediction of *in vitro* aggregation has a dramatic effect on the predicted antimicrobial activity for small peptides. In contrast, for larger peptides on the right side of Figure 5, there is little influence of the predicted aggregation propensity on predicted antimicrobial activity. The tertiary interactions and intermediates in the aggregating process increase the complexity of the aggregation process and its predictability, clearly improving the prediction of antimicrobial activity in small peptides.

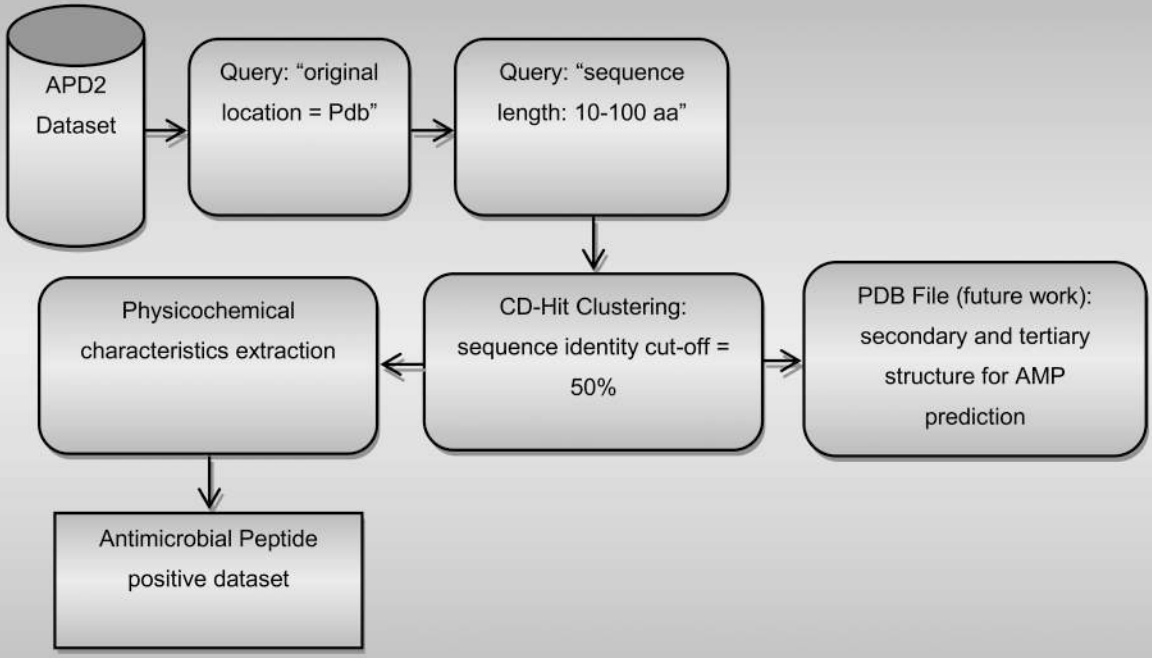
The pharmaceutical industry may currently choose from many possible AMP prediction tools and AMP databases. Even for the prediction tools that aim to be general predictors, the best choice will be based on the putative AMP characteristics and the database's closeness to the putative AMP category. Any putative AMP should clearly be submitted to as many prediction tools as possible and the results compared. The ANFIS model presented here has excellent accuracy, potentially allowing industry a better return on investments by increasing the success rate for putative AMP testing.

In summary, data reported here show that ANFIS is an efficient model for pattern recognition in antimicrobial peptide prediction. The ANFIS methodology is shown to deal effectively with the fuzzy nature of the correlation between physicochemical properties and antimicrobial activity. Notably, by selecting sequences for which 3D structures are known, future work can easily leverage structure-based descriptors and test whether their incorporation can improve predictions based on sequence alone.

REFERENCES

1. Otero-Gonzalez, A. J.; Magalhaes, B. S.; Garcia-Villarino, M.; Lopez-Abarrategui, C.; Sousa, D. A.; Dias, S. C.; Franco, O. L. FASEB J 2010, 24, 1320–1334.
2. Fjell, C. D.; Hiss, J. A.; Hancock, R. E.; Schneider, G. Nat Rev Drug Discov 2011, 11, 37–51.
3. Thomas, S.; Karnik, S.; Barai, R. S.; Jayaraman, V. K.; Idicula-Thomas, S. Nucleic Acids Res 2010, 38, D774–D780.
4. Franco, O. L. FEBS Lett 2011, 585, 995–1000.
5. Candido E de, S.; Pinto, M. F.; Pelegrini, P. B.; Lima, T. B.; Silva, O. N.; Pogue, R.; Grossi-de-Sa, M. F.; Franco, O. L. FASEB J 2011, 25, 3290–3305.
6. Yeung, A. T.; Gellatly, S. L.; Hancock, R. E. Cell Mol Life Sci 2011, 68, 2161–2176.
7. Sang, Y.; Blecha, F. Anim Health Res Rev 2008, 9, 227–235.
8. Jenssen, H.; Hamill, P.; Hancock, R. E. Clin Microbiol Rev 2006, 19, 491–511.

9. Glukhov, E.; Burrows, L. L.; Deber, C. M. Biopolymers 2008, 89, 360–371.
10. Rubinstein, M.; Niv, M. Y. Biopolymers 2009, 91, 505–513.
11. Whelan, C.; Roark, B.; Sonmez, K. Conf Proc IEEE Eng Med Biol Soc 2010, 2010, 764–767.
12. Hadley, E. B.; Hancock, R. E. Curr Top Med Chem 2010, 10, 1872–1881.
13. Cherkasov, A.; Jankovic, B. Molecules 2004, 9, 1034–1052.
14. Fjell, C. D.; Hancock, R. E.; Cherkasov, A. Bioinformatics 2007, 23, 1148–1155.
15. Lata, S.; Sharma, B. K.; Raghava, G. P. BMC Bioinformatics 2007, 8, 263.
16. Kumar, M.; Verma, R.; Raghava, G. P. J Biol Chem 2006, 281, 5357–5363.
17. Wang, G.; Li, X.; Wang, Z. Nucleic Acids Res 2009, 37, D933–D937.
18. Wang, Z.; Wang, G. Nucleic Acids Res 2004, 32, D590–D592.
19. Fjell, C. D.; Jenssen, H.; Hilpert, K.; Cheung, W. A.; Pante, N.; Hancock, R. E.; Cherkasov, A. J Med Chem 2009, 52, 2006–2015.
20. Huang, Y.; Niu, B.; Gao, Y.; Fu, L.; Li, W. Bioinformatics 2010, 26, 680–682.
21. Lata, S.; Mishra, N. K.; Raghava, G. P. BMC Bioinformatics 2010, 11(Suppl 1), S19.
22. Brahmachary, M.; Krishnan, S. P.; Koh, J. L.; Khan, A. M.; Seah, S. H.; Tan, T. W.; Brusic, V.; Bajic, V. B. Nucleic Acids Res 2004, 32, D586–D589.
23. Hammami, R.; Zouhir, A.; Le Lay, C.; Ben Hamida, J.; Fliss, I. BMC Microbiol 2010, 10, 22.
24. Juretic, D.; Vukicevic, D.; Ilic, N.; Antcheva, N.; Tossi, A. J Chem Inf Model 2009, 49, 2873–2882.
25. Hammami, R.; Fliss, I. Drug Discov Today 2010, 15, 540–546.
26. Wang, P.; Hu, L.; Liu, G.; Jiang, N.; Chen, X.; Xu, J.; Zheng, W.; Li, L.; Tan, M.; Chen, Z.; Song, H.; Cai, Y. D.; Chou, K. C. PLoS One 2011, 6, e18476.
27. Jang, J.-S. R. IEEE Trans Syst Man Cybern 1993, 23, 665–685.
28. Jang, J.-S. R. Proceedings of the Fifth IEEE International Conference on Fuzzy Systems; IEEE press: Piscataway, N.J., 1996, 1493–1499.
29. Thangavel, K.; Mohideen, A. K. In Trendz in Information Sciences & Computing (TISC); IEEE press: Piscataway, N.J., 2010, p 68–72.
30. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. Nucleic Acids Res 2000, 28, 235–242.
31. Torrent, M.; Andreu, D.; Nogues, V. M.; Boix, E. PLoS One 2011, 6, e16968.
32. Wilkins, M. R.; Gasteiger, E.; Bairoch, A.; Sanchez, J. C.; Williams, K. L.; Appel, R. D.; Hochstrasser, D. F. Methods Mol Biol 1999, 112, 531–552.
33. Fernandez-Escamilla, A. M.; Rousseau, F.; Schymkowitz, J.; Serrano, L. Nat Biotechnol 2004, 22, 1302–1306.
34. Linding, R.; Schymkowitz, J.; Rousseau, F.; Diella, F.; Serrano, L. J Mol Biol 2004, 342, 345–353.
35. Rousseau, F.; Schymkowitz, J.; Serrano, L. Curr Opin Struct Biol 2006, 16, 118–126.
36. Conchillo-Sole, O.; de Groot, N. S.; Aviles, F. X.; Vendrell, J.; Daura, X.; Ventura, S. BMC Bioinformatics 2007, 8, 65.
37. Kyte, J.; Doolittle, R. F. J Mol Biol 1982, 157, 105–132.
38. Kall, L.; Krogh, A.; Sonnhammer, E. L. Nucleic Acids Res 2007, 35, W429–432.
39. Jain, E.; Bairoch, A.; Duvaud, S.; Phan, I.; Redaschi, N.; Suzek, B. E.; Martin, M. J.; McGarvey, P.; Gasteiger, E. BMC Bioinformatics 2009, 10, 136.
40. The Uniprot Consortium. Nucleic Acids Res 2011, 39, D214–219.
41. Porto, W.; Fernandes, F.; Franco, O. In Advances in Bioinformatics and Computational Biology; Ferreira, C.; Miyano, S.; Stadler, P., Eds.; Springer: Berlin and Heidelberg, 2010; pp 59–62.
42. Bals, R. Respir Res 2000, 1, 141–150.
43. The Uniprot Consortium. Nucleic Acids Res 2012, 40, D71–D75.
44. Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. J Mol Biol 1990, 215, 403–410.


APD2
Dataset
Query: “original location = Pdb”
Query: “sequence length: 10-100 aa”
Physicochemical characteristics extraction
CD-Hit Clustering: sequence identity cut-off = 50%
PDB File (future work): secondary and tertiary structure for AMP prediction
Antimicrobial Peptide positive dataset

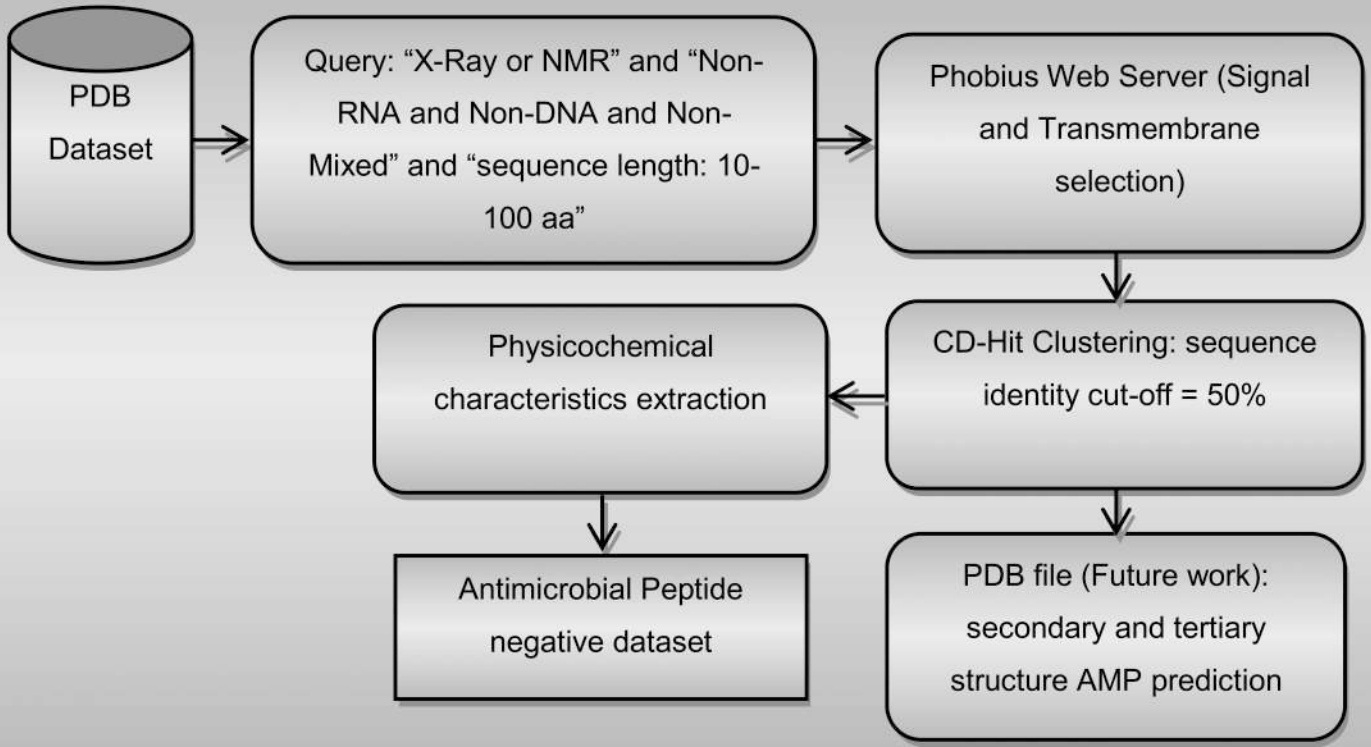

PDB Dataset
Query: “X-Ray or NMR” and “Non-RNA and Non-DNA and Non-Mixed” and “sequence length: 10-100 aa”
Phobius Web Server (Signal and Transmembrane selection)
CD-Hit Clustering: sequence identity cut-off = 50%
Physicochemical characteristics extraction
Antimicrobial Peptide negative dataset
PDB file (Future work): secondary and tertiary structure AMP prediction

Table STI Antimicrobial peptides positive dataset with selected physicochemical features used in this study.

Pdb id	*in-vitro* aggregation	turn structure propensity	α-helix propensity	β-sheet propensity	isoelectric point	Length	hydrophobic mean	*in-vivo* aggregation	DOI
1AFP:A	0.000	79.280	22.245	115.446	9.270	51.000	-914.000	-38.039	10.2210/pdb1AFP/pdb
1AYJ:A	2.061	23.385	19.532	170.073	8.510	51.000	-451.000	-18.039	10.2210/pdb1AYJ/pdb
1BNB:A	0.000	11.501	0.000	126.563	9.570	38.000	126.000	19.737	10.2210/pdb1BNB/pdb
1C01:A	135.752	51.480	23.350	173.497	9.080	76.000	-321.000	-7.632	10.2210/pdb1C01/pdb
1CIX:A	2.299	20.383	1.668	146.200	9.340	44.000	-334.000	14.773	10.2210/pdb1CIX/pdb
1CW5:A	23.126	39.489	3.649	177.552	9.700	48.000	-277.000	-14.792	10.2210/pdb1CW5/pdb
1CZ6:A	0.000	4.653	0.000	76.082	10.220	25.000	-1056.000	-58.000	10.2210/pdb1CZ6/pdb
1D6X:A	0.000	1.711	0.000	14.974	12.480	13.000	-792.000	233.077	10.2210/pdb1D6X/pdb
1D7N:A	0.577	0.238	160.339	15.422	10.300	14.000	1157.000	106.429	10.2210/pdb1D7N/pdb
1DKC:A	4.390	22.347	0.000	164.400	8.900	38.000	-232.000	-10.526	10.2210/pdb1DKC/pdb
1E4Q:A	0.000	9.186	2.692	105.550	9.510	37.000	-119.000	10.270	10.2210/pdb1E4Q/pdb
1E4S:A	0.000	8.738	1.073	157.894	8.870	36.000	-272.000	-1.667	10.2210/pdb1E4S/pdb
1EWS:A	0.000	5.920	0.000	110.750	7.540	32.000	-338.000	2.188	10.2210/pdb1EWS/pdb
1FRY:A	0.000	8.330	2.573	92.862	12.310	29.000	-210.000	6.207	10.2210/pdb1FRY/pdb
1G89:A	0.000	1.633	0.000	8.525	12.010	13.000	-1069.000	163.846	10.2210/pdb1G89/pdb
1HVZ:A	0.000	1.854	0.000	51.597	9.300	18.000	350.000	66.667	10.2210/pdb1HVZ/pdb
1I2U:A	152.813	43.678	27.674	78.863	7.770	44.000	-468.000	-23.409	10.2210/pdb1I2U/pdb
1ICA:A	0.000	54.397	2.896	46.052	8.690	40.000	48.000	-22.250	10.2210/pdb1ICA/pdb
1KFP:A	0.000	2.675	0.690	83.698	9.580	18.000	-1061.000	-90.000	10.2210/pdb1KFP/pdb
1KJ6:A	1.876	12.404	0.919	138.675	10.080	45.000	-700.000	-35.778	10.2210/pdb1KJ6/pdb
1KV4:A	236.979	9.043	4.538	84.424	11.360	42.000	-90.000	-2.857	10.2210/pdb1KV4/pdb
1LFC:A	0.000	2.103	0.651	92.935	11.840	25.000	-576.000	-30.400	10.2210/pdb1LFC/pdb
1M02:A	0.000	3.146	0.000	16.323	9.990	12.000	-1158.000	-46.667	10.2210/pdb1M02/pdb
1MAG:A	1186.400	0.759	2.366	59.833	5.490	15.000	2107.000	608.000	10.2210/pdb1MAG/pdb
1MM0:A	0.000	7.916	0.000	111.319	8.960	36.000	-153.000	8.611	10.2210/pdb1MM0/pdb
1MMC:A	0.000	19.429	5.171	71.034	8.920	30.000	-347.000	-23.333	10.2210/pdb1MMC/pdb
1NKL:A	10.892	19.951	58.436	241.768	9.170	78.000	-231.000	-1.795	10.2210/pdb1NKL/pdb
1OF9:A	569.255	31.868	96.848	130.013	5.650	77.000	406.000	11.948	10.2210/pdb1OF9/pdb
1OG7:A	0.000	54.749	11.860	76.518	8.790	43.000	-574.000	-46.977	10.2210/pdb1OG7/pdb
1PG1:A	21.507	8.570	0.000	43.629	10.660	18.000	-250.000	26.111	10.2210/pdb1PG1/pdb
1Q71:A	62.603	7.073	0.000	29.011	5.240	21.000	400.000	110.952	10.2210/pdb1Q71/pdb
1RKK:A	0.000	1.019	0.000	68.692	10.330	18.000	-833.000	57.778	10.2210/pdb1RKK/pdb
1RPB:A	338.887	12.281	0.000	48.781	3.800	21.000	1157.000	162.857	10.2210/pdb1RPB/pdb
1S6W:A	0.000	20.487	0.000	124.319	8.230	21.000	576.000	58.095	10.2210/pdb1S6W/pdb
1T51:A	0.000	3.299	0.000	12.646	8.590	13.000	777.000	173.846	10.2210/pdb1T51/pdb
1L9L:A	63.626	17.943	31.363	156.512	10.830	74.000	-747.000	-25.000	10.2210/pdb1L9L/pdb
1ZMM:A	60.470	11.545	1.018	203.836	8.980	33.000	661.000	79.394	10.2210/pdb1ZMM/pdb
1ZMP:A	0.000	12.081	0.636	117.826	8.960	32.000	-113.000	0.313	10.2210/pdb1ZMP/pdb
1ZMQ:A	3.991	3.937	2.032	245.564	8.350	32.000	0.000	35.000	10.2210/pdb1ZMQ/pdb
2MLT:A	34.103	2.017	11.337	45.617	12.020	26.000	273.000	61.538	10.2210/pdb2MLT/pdb
3HJ2:A	320.396	7.875	0.764	112.266	8.590	33.000	324.000	44.242	10.2210/pdb3HJ2/pdb
1BRZ:A	0.000	16.632	30.270	95.871	5.720	54.000	-1106.000	-47.963	10.2210/pdb1BRZ/pdb
1DQC:A	0.000	29.585	10.200	266.876	8.560	73.000	-373.000	-0.959	10.2210/pdb1DQC/pdb
1E4T:A	0.792	9.176	16.359	110.338	9.340	37.000	-476.000	-18.378	10.2210/pdb1E4T/pdb
1E68:A	1285.600	20.256	90.567	165.810	10.090	70.000	539.000	24.000	10.2210/pdb1E68/pdb
1ED0:A	0.000	30.910	1.008	106.072	9.300	46.000	-463.000	-26.739	10.2210/pdb1ED0/pdb

Table STI Antimicrobial peptides positive dataset with selected physicochemical features used in this study.

Pdb id	*in-vitro* aggregation	turn structure propensity	α-helix propensity	β-sheet propensity	isoelectric point	Length	hydrophobic mean	*in-vivo* aggregation	DOI
1G6E:A	436.731	63.920	22.806	324.979	5.440	87.000	-403.000	-8.621	10.2210/pdb1G6E/pdb
1IYC:A	4.564	20.984	9.342	54.444	9.250	36.000	-592.000	-34.722	10.2210/pdb1IYC/pdb
1K48:A	0.000	19.176	0.000	75.682	5.960	29.000	152.000	13.448	10.2210/pdb1K48/pdb
1MQZ:A	0.794	6.186	0.000	107.234	4.000	19.000	858.000	74.737	10.2210/pdb1MQZ/pdb
1MR4:A	0.000	15.440	2.063	122.359	9.080	47.000	-472.000	-24.468	10.2210/pdb1MR4/pdb
1MYN:A	0.810	35.082	34.829	68.118	7.700	44.000	-741.000	-56.591	10.2210/pdb1MYN/pdb
1P9Z:A	2.469	14.686	5.728	92.002	8.420	41.000	-66.000	-20.244	10.2210/pdb1P9Z/pdb
1PXQ:A	53.374	19.219	0.748	72.925	4.030	35.000	691.000	38.286	10.2210/pdb1PXQ/pdb
1Q3J:A	4.199	39.866	1.128	84.174	9.130	36.000	-944.000	-85.278	10.2210/pdb1Q3J/pdb
1R1F:A	3.010	25.942	7.486	197.196	4.780	37.000	-189.000	-41.892	10.2210/pdb1R1F/pdb
1TI5:A	42.147	18.105	1.459	91.733	9.060	46.000	-283.000	-5.435	10.2210/pdb1TI5/pdb
1UT3:A	0.000	11.280	11.889	88.100	11.630	38.000	-95.000	13.158	10.2210/pdb1UT3/pdb
1YP8:A	3.299	14.777	2.184	81.158	4.680	33.000	6.000	12.424	10.2210/pdb1YP8/pdb
1Z6V:A	7.133	3.830	2.377	168.218	11.240	49.000	-851.000	-41.633	10.2210/pdb1Z6V/pdb
1Z99:A	0.000	35.988	10.406	92.651	9.510	42.000	-1095.000	-59.286	10.2210/pdb1Z99/pdb
1ZA8:A	74.451	8.077	2.179	98.312	6.100	31.000	690.000	77.419	10.2210/pdb1ZA8/pdb
2B5B:A	0.000	21.759	2.850	157.053	9.390	36.000	-608.000	-15.000	10.2210/pdb2B5B/pdb
2B9K:A	721.393	83.737	1.456	107.183	9.400	47.000	-351.000	24.894	10.2210/pdb2B9K/pdb
2DCV:A	4.714	20.232	0.974	187.443	9.610	42.000	-214.000	18.333	10.2210/pdb2DCV/pdb
2E2F:A	0.000	14.845	1.378	123.257	8.350	41.000	-371.000	-26.098	10.2210/pdb2E2F/pdb
2G9P:A	0.562	9.307	25.031	29.473	11.330	26.000	-438.000	-11.923	10.2210/pdb2G9P/pdb
2JNI:A	367.843	1.192	0.000	97.346	10.850	21.000	-57.000	160.000	10.2210/pdb2JNI/pdb
2JPJ:A	126.549	19.694	22.515	86.568	10.160	39.000	-744.000	-58.974	10.2210/pdb2JPJ/pdb
2JPY:A	58.952	1.442	0.000	38.608	7.020	19.000	1105.000	193.158	10.2210/pdb2JPY/pdb
2JR3:A	437.035	18.434	11.932	150.920	5.300	42.000	545.000	34.048	10.2210/pdb2JR3/pdb
2JS9:A	0.000	71.494	59.880	183.167	6.480	99.000	-670.000	-24.040	10.2210/pdb2JS9/pdb
2K1I:A	18.566	40.812	0.000	126.606	10.440	32.000	-759.000	-65.938	10.2210/pdb2K1I/pdb
2K35:A	51.549	17.557	43.797	206.784	9.050	60.000	-948.000	-38.667	10.2210/pdb2K35/pdb
2K9B:A	3.966	6.373	103.013	25.783	10.000	33.000	191.000	-50.000	10.2210/pdb2K9B/pdb
2KCN:A	0.000	28.663	4.217	184.782	8.930	55.000	-1375.000	-74.727	10.2210/pdb2KCN/pdb
2KEG:A	490.812	32.964	10.322	29.798	10.520	32.000	-656.000	-46.250	10.2210/pdb2KEG/pdb
2KFE:A	1.461	5.964	14.368	18.306	10.170	24.000	-1300.000	-153.333	10.2210/pdb2KFE/pdb
2KJF:A	1424.940	12.892	42.589	183.550	10.000	60.000	1058.000	51.333	10.2210/pdb2KJF/pdb
2KNJ:A	51.861	23.099	123.717	159.741	5.160	90.000	-616.000	-28.889	10.2210/pdb2KNJ/pdb
2KUY:A	585.217	24.507	1.057	85.828	7.040	43.000	-319.000	10.698	10.2210/pdb2KUY/pdb
2L2R:A	0.000	13.672	5.184	70.037	9.470	37.000	-1608.000	-155.405	10.2210/pdb2L2R/pdb
2PCO:A	0.000	6.717	10.579	19.087	11.780	26.000	-1350.000	-101.923	10.2210/pdb2PCO/pdb
2RLG:A	0.000	1.858	0.000	13.036	10.820	18.000	-594.000	-18.333	10.2210/pdb2RLG/pdb
2RNG:A	1350.550	36.596	45.418	236.333	9.270	79.000	72.000	14.430	10.2210/pdb2RNG/pdb
1VM5:A	5.737	0.737	26.830	26.852	6.070	13.000	669.000	130.769	10.2210/pdb1VM5/pdb
1XKM:A	49.159	8.381	17.639	41.620	9.790	22.000	-173.000	-5.909	10.2210/pdb1XKM/pdb
1XKM:B	94.863	2.436	202.067	21.317	9.860	25.000	-464.000	-5.200	10.2210/pdb1XKM/pdb
1XV3:A	0.000	20.467	6.237	133.960	9.150	47.000	-153.000	13.404	10.2210/pdb1XV3/pdb
1YTR:A	0.000	3.542	5.437	34.524	10.400	26.000	-423.000	9.615	10.2210/pdb1YTR/pdb
1Z64:A	0.000	9.127	0.000	34.469	10.180	25.000	-68.000	2.400	10.2210/pdb1Z64/pdb
1ZFU:A	0.000	53.762	4.868	45.628	7.770	40.000	-695.000	-50.000	10.2210/pdb1ZFU/pdb

Table STI Antimicrobial peptides positive dataset with selected physicochemical features used in this study.

Pdb id	*in-vitro* aggregation	turn structure propensity	α-helix propensity	β-sheet propensity	isoelectric point	Length	hydrophobic mean	*in-vivo* aggregation	DOI
1ZRV:A	1.658	1.951	99.033	65.355	11.070	25.000	156.000	38.400	10.2210/pdb1ZRV/pdb
1ZRX:A	148.347	9.211	5.635	135.565	10.660	42.000	-2.000	-6.190	10.2210/pdb1ZRX/pdb
2AMN:A	10.083	2.713	36.166	84.026	11.600	26.000	-69.000	98.077	10.2210/pdb2AMN/pdb
2B68:A	108.802	60.520	63.212	129.670	8.730	43.000	-412.000	-40.930	10.2210/pdb2B68/pdb
2CRD:A	1.758	9.718	8.905	132.096	9.130	37.000	-832.000	-54.865	10.2210/pdb2CRD/pdb
2EEM:A	444.300	9.554	3.601	99.235	9.580	34.000	-344.000	2.353	10.2210/pdb2EEM/pdb
2ERI:A	12.735	22.546	0.764	71.106	8.330	31.000	642.000	78.065	10.2210/pdb2ERI/pdb
2G9L:A	22.602	11.982	7.203	106.809	9.510	37.000	330.000	4.595	10.2210/pdb2G9L/pdb
2GDL:A	54.671	7.390	1.139	142.829	12.850	31.000	-429.000	-11.613	10.2210/pdb2GDL/pdb
2GL1:A	0.000	22.378	11.321	152.287	8.510	47.000	-1174.000	-84.681	10.2210/pdb2GL1/pdb
2GW9:A	0.000	7.592	2.032	130.781	9.860	32.000	-469.000	-10.313	10.2210/pdb2GW9/pdb
2HFR:A	69.780	2.649	28.614	68.322	11.740	27.000	185.000	97.037	10.2210/pdb2HFR/pdb
2JOS:A	0.000	4.387	0.000	92.842	12.010	22.000	455.000	115.909	10.2210/pdb2JOS/pdb
2K10:A	1.380	16.618	20.815	59.278	9.510	32.000	444.000	30.313	10.2210/pdb2K10/pdb
2K38:A	7.118	8.034	9.828	61.909	10.300	35.000	-131.000	-3.429	10.2210/pdb2K38/pdb
2K6O:A	0.000	17.002	5.977	73.822	10.610	37.000	-724.000	-38.919	10.2210/pdb2K6O/pdb
2KEF:A	65.913	2.852	0.000	156.454	8.220	25.000	388.000	68.400	10.2210/pdb2KEF/pdb
2KSG:A	0.000	21.212	28.601	37.806	5.070	48.000	-29.000	-35.000	10.2210/pdb2KSG/pdb
2MAG:A	0.000	8.059	0.994	14.023	10.000	23.000	83.000	34.783	10.2210/pdb2MAG/pdb
8TFV:A	0.000	3.128	1.733	24.810	10.470	21.000	-900.000	-84.286	10.2210/pdb8TFV/pdb
2DCX:A	0.000	0.890	0.835	12.995	10.480	13.000	446.000	188.462	10.2210/pdb2DCX/pdb
1XC0:A	19.181	20.450	1.242	37.449	8.590	33.000	745.000	83.333	10.2210/pdb1XC0/pdb
2A2B:A	7.337	38.295	4.258	59.051	9.310	41.000	-185.000	-6.341	10.2210/pdb2A2B/pdb
Average	**104.111**	**17.747**	**16.475**	**100.168**	**9.050**	**37.226**	**-205.887**	**13.405**	
Count	**115**	**115**	**115**	**115**	**115**	**115**	**115**	**115**	
Sum	**11,972.770**	**2,040.878**	**1,894.606**	**11,519.364**	**1,040.700**	**4,281.000**	**-23,677.000**	**1,541.614**	
Variance	**72,229.650**	**298.876**	**1,045.293**	**3,845.331**	**3.896**	**299.931**	**393,789.259**	**7,684.534**	
Max	**1,186.400**	**83.737**	**202.067**	**324.979**	**12.850**	**99.000**	**2,107.000**	**608.000**	
Min	**0.000**	**0.238**	**0.000**	**8.525**	**3.800**	**12.000**	**-1,608.000**	**-155.405**	
Standard deviation	**268.756**	**17.288**	**32.331**	**62.011**	**1.974**	**17.319**	**627.526**	**87.661**	

Table STII Antimicrobial peptides negative dataset with selected physicochemical features used in this study.

Pdb id	*in-vitro* aggregation	turn structure propensity	α-helix propensity	β-sheet propensity	isoelectric point	Length	hydrophobic mean	*in-vivo* aggregation	DOI
1A11:A	1351.270	1.527	14.422	55.822	8.750	25.000	724.000	97.600	10.2210/pdb1A11/pdb
1A1W:A	478.267	23.368	594.343	180.493	6.100	91.000	-373.000	-10.000	10.2210/pdb1A1W/pdb
1AFO:A	1701.750	2.973	18.999	141.899	9.700	40.000	938.000	108.000	10.2210/pdb1AFO/pdb
1BTR:A	1022.950	0.410	0.000	81.565	5.490	20.000	2105.000	364.000	10.2210/pdb1BTR/pdb
1FCT:A	0.000	3.247	2.099	64.459	12.180	32.000	184.000	-17.813	10.2210/pdb1FCT/pdb
1FV5:A	1.085	12.019	6.206	77.214	8.660	36.000	200.000	36.111	10.2210/pdb1FV5/pdb
1H7D:A	198.802	12.621	5.405	158.443	9.500	49.000	635.000	43.673	10.2210/pdb1H7D/pdb
1N7L:A	1721.630	2.915	1592.430	77.793	10.830	53.000	840.000	66.038	10.2210/pdb1N7L/pdb
1PYV:A	1.992	27.002	26.138	48.719	12.480	54.000	-291.000	22.963	10.2210/pdb1PYV/pdb
1SKH:A	973.780	3.771	9.195	61.728	10.190	30.000	413.000	92.000	10.2210/pdb1SKH/pdb
1SPF:A	1627.040	0.812	43.243	297.334	9.700	35.000	2351.000	240.000	10.2210/pdb1SPF/pdb
1UTR:A	1030.750	42.156	103.128	346.515	4.930	96.000	289.000	8.958	10.2210/pdb1UTR/pdb
1VRY:A	2212.100	12.437	128.355	196.054	10.220	76.000	236.000	21.711	10.2210/pdb1VRY/pdb
1Y0J:A	0.000	21.280	5.836	56.368	9.250	46.000	-737.000	51.739	10.2210/pdb1Y0J/pdb
1Y4E:A	580.631	3.784	0.000	55.692	4.210	27.000	1204.000	201.852	10.2210/pdb1Y4E/pdb
1ZZA:A	1975.640	44.262	264.671	292.863	5.200	90.000	243.000	11.889	10.2210/pdb1ZZA/pdb
2HAC:B	1391.630	7.020	39.151	90.965	8.150	33.000	885.000	137.273	10.2210/pdb2HAC/pdb
2J5D:A	1042.120	12.611	40.188	88.933	11.070	45.000	624.000	74.000	10.2210/pdb2J5D/pdb
2JO1:A	1553.860	15.656	62.232	236.939	9.300	72.000	-575.000	13.889	10.2210/pdb2JO1/pdb
2JP3:A	1371.420	40.517	29.537	136.784	8.600	67.000	52.000	9.552	10.2210/pdb2JP3/pdb
2JWA:A	1736.160	4.296	31.535	129.937	11.570	44.000	784.000	72.045	10.2210/pdb2JWA/pdb
2K1K:A	1748.220	5.260	94.804	82.660	12.000	38.000	1139.000	101.053	10.2210/pdb2K1K/pdb
2K9P:A	2441.220	16.273	277.261	299.229	9.820	80.000	709.000	45.750	10.2210/pdb2K9P/pdb
2K9Y:A	1697.580	31.368	130.640	111.913	10.740	41.000	822.000	70.488	10.2210/pdb2K9Y/pdb
2KLU:A	1451.670	16.221	50.181	145.980	12.310	70.000	-140.000	3.714	10.2210/pdb2KLU/pdb
2KNC:A	1765.140	13.041	309.167	73.921	4.480	54.000	285.000	27.407	10.2210/pdb2KNC/pdb
2KNC:B	1604.620	13.658	526.723	271.868	9.160	79.000	176.000	18.734	10.2210/pdb2KNC/pdb
2KOE:A	1103.430	13.498	21.701	102.281	9.300	40.000	460.000	58.250	10.2210/pdb2KOE/pdb
2KS1:B	1413.670	8.057	601.903	78.250	11.830	44.000	591.000	58.182	10.2210/pdb2KS1/pdb
2KU8:A	81.504	25.522	61.704	110.558	6.970	61.000	910.000	16.721	10.2210/pdb2KU8/pdb
2L0E:A	761.539	9.346	7.240	73.539	9.310	31.000	161.000	30.968	10.2210/pdb2L0E/pdb
2L2T:A	1519.730	7.713	25.879	130.585	12.020	44.000	614.000	65.682	10.2210/pdb2L2T/pdb
2L35:A	2789.610	13.385	270.763	297.073	4.680	63.000	1681.000	97.460	10.2210/pdb2L35/pdb
2L6Z:C	31.539	45.452	55.032	274.293	6.090	96.000	-173.000	-2.813	10.2210/pdb2L6Z/pdb
2L9U:A	1567.060	7.038	18.313	120.320	12.000	40.000	250.000	50.750	10.2210/pdb2L9U/pdb
2L9Z:A	841.362	10.911	43.719	102.594	4.430	39.000	-131.000	-7.436	10.2210/pdb2L9Z/pdb
2LAT:A	1547.000	6.014	116.810	113.399	6.500	37.000	816.000	99.730	10.2210/pdb2LAT/pdb
2RNG:A	1350.550	36.596	45.418	236.333	9.270	79.000	72.000	14.430	10.2210/pdb2RNG/pdb
2RQO:A	873.661	23.360	13.504	192.689	5.520	48.000	796.000	27.500	10.2210/pdb2RQO/pdb
2W1O:A	520.343	70.858	79.938	105.224	5.130	70.000	17.000	-8.714	10.2210/pdb2W1O/pdb
2WBR:A	75.010	32.991	94.370	151.145	6.290	89.000	-112.000	-1.461	10.2210/pdb2WBR/pdb
3LRI:A	43.990	32.733	41.343	144.430	8.550	83.000	-17.000	2.169	10.2210/pdb3LRI/pdb
3MRA:A	1749.190	0.414	0.000	186.483	8.760	25.000	1504.000	300.800	10.2210/pdb3MRA/pdb
1A4P:A	1137.250	41.683	166.684	216.807	7.300	96.000	-384.000	-6.458	10.2210/pdb1A4P/pdb
1BCC:G	989.103	13.781	23.865	309.296	10.710	81.000	-504.000	8.642	10.2210/pdb1BCC/pdb
1BCC:H	1.444	16.666	23.758	141.445	4.500	78.000	-1054.000	8.974	10.2210/pdb1BCC/pdb
1BCC:J	1226.050	18.968	83.634	180.836	9.520	62.000	-105.000	11.290	10.2210/pdb1BCC/pdb
1BE3:I	24.215	15.648	27.949	162.768	11.540	78.000	374.000	8.590	10.2210/pdb1BE3/pdb
1BE3:K	221.736	51.406	17.868	66.401	10.150	56.000	-88.000	11.964	10.2210/pdb1BE3/pdb
1BL8:A	3352.220	46.613	261.220	395.377	8.200	97.000	911.000	38.041	10.2210/pdb1BL8/pdb

1BT6:C	0.000	0.580	7.607	40.417	5.380	13.000	223.000	68.462	10.2210/pdb1BT6/pdb
1CD1:B	627.889	93.079	62.948	332.421	7.130	99.000	-778.000	11.111	10.2210/pdb1CD1/pdb
1CO7:I	1002.320	45.573	120.788	177.630	9.160	99.000	-126.000	-1.313	10.2210/pdb1CO7/pdb
1EF2:C	1572.100	39.608	278.881	202.993	5.360	100.000	-2.000	-1.000	10.2210/pdb1EF2/pdb
1EZV:G	1043.450	38.588	21.650	195.257	9.700	93.000	-556.000	4.624	10.2210/pdb1EZV/pdb
1EZV:H	0.942	23.346	33.691	135.158	5.070	74.000	-999.000	5.811	10.2210/pdb1EZV/pdb
1EZX:B	590.944	5.349	2.505	135.280	9.820	36.000	-194.000	29.444	10.2210/pdb1EZX/pdb
1GAQ:B	31.345	48.571	4.493	409.786	3.880	98.000	-289.000	14.286	10.2210/pdb1GAQ/pdb
1H0X:A	944.581	36.853	217.285	309.509	10.330	100.000	-59.000	4.100	10.2210/pdb1H0X/pdb
1H2S:B	3065.030	10.682	121.543	140.076	3.570	60.000	1703.000	74.833	10.2210/pdb1H2S/pdb
1I72:B	166.156	39.211	56.636	139.226	4.610	67.000	-490.000	18.060	10.2210/pdb1I72/pdb
1IJD:A	1560.900	8.535	36.364	140.845	10.000	53.000	872.000	58.868	10.2210/pdb1IJD/pdb
1IJD:B	1323.110	5.045	25.047	123.896	5.640	42.000	562.000	74.286	10.2210/pdb1IJD/pdb
1IZL:K	1422.450	2.204	6.579	67.545	4.560	37.000	1286.000	141.081	10.2210/pdb1IZL/pdb
1JB0:C	52.179	30.785	16.863	205.878	5.680	80.000	14.000	81.125	10.2210/pdb1JB0/pdb
1JB0:E	4.017	41.275	0.000	247.191	9.520	75.000	-520.000	86.533	10.2210/pdb1JB0/pdb
1JB0:I	1211.870	9.669	5.248	88.755	3.790	38.000	1408.000	170.789	10.2210/pdb1JB0/pdb
1JB0:J	1444.140	10.821	7.434	131.768	6.690	41.000	815.000	158.293	10.2210/pdb1JB0/pdb
1JB0:K	1426.280	25.008	51.228	221.758	6.510	83.000	933.000	78.193	10.2210/pdb1JB0/pdb
1JB0:M	860.174	1.129	32.609	84.650	5.820	31.000	1148.000	209.355	10.2210/pdb1JB0/pdb
1JB0:X	1767.240	2.118	36.682	103.773	9.820	35.000	971.000	185.429	10.2210/pdb1JB0/pdb
1KVD:A	1088.170	37.348	37.733	151.807	8.960	63.000	375.000	7.778	10.2210/pdb1KVD/pdb
1KVD:B	51.797	76.155	48.071	123.450	4.570	77.000	-584.000	6.364	10.2210/pdb1KVD/pdb
1KYO:I	1727.740	18.025	16.486	219.976	9.900	57.000	93.000	31.579	10.2210/pdb1KYO/pdb
1M56:D	1727.280	17.010	17.544	190.946	6.370	51.000	675.000	32.941	10.2210/pdb1M56/pdb
1O82:A	1285.600	20.256	90.567	165.810	10.090	70.000	539.000	24.000	10.2210/pdb1O82/pdb
1OCC:F	703.083	54.652	102.229	145.949	6.070	98.000	-499.000	16.122	10.2210/pdb1OCC/pdb
1OCC:H	6.327	38.712	160.716	176.596	8.780	85.000	-895.000	18.588	10.2210/pdb1OCC/pdb
1OCC:I	1199.000	15.625	147.843	156.604	10.280	73.000	-274.000	21.644	10.2210/pdb1OCC/pdb
1OCC:J	415.815	22.290	32.647	129.183	8.710	59.000	-215.000	26.780	10.2210/pdb1OCC/pdb
1OCC:K	1139.690	23.404	7.707	144.106	8.150	56.000	-362.000	28.214	10.2210/pdb1OCC/pdb
1OCC:L	777.070	28.412	75.224	107.520	9.820	47.000	-28.000	33.617	10.2210/pdb1OCC/pdb
1OCC:M	893.995	10.298	13.394	105.811	9.600	46.000	52.000	34.348	10.2210/pdb1OCC/pdb
1Q90:G	1320.240	4.520	11.058	108.384	4.370	37.000	1276.000	204.595	10.2210/pdb1Q90/pdb
1Q90:L	1613.370	3.410	9.919	136.335	9.520	32.000	1894.000	236.563	10.2210/pdb1Q90/pdb
1Q90:M	1516.650	5.014	6.957	73.939	4.330	39.000	1095.000	194.103	10.2210/pdb1Q90/pdb
1Q90:N	1592.660	5.222	2.402	155.059	5.990	31.000	1039.000	244.194	10.2210/pdb1Q90/pdb
1Q90:R	500.906	28.426	29.589	43.818	6.110	49.000	378.000	154.490	10.2210/pdb1Q90/pdb
1QCR:D	138.139	10.011	129.902	129.424	9.840	75.000	-324.000	-1.600	10.2210/pdb1QCR/pdb
1QLE:D	1561.360	12.106	64.879	120.810	6.740	43.000	405.000	34.419	10.2210/pdb1QLE/pdb
1QO1:K	3147.320	24.912	88.342	134.216	4.440	79.000	1267.000	52.532	10.2210/pdb1QO1/pdb
1RH5:B	1238.580	11.586	383.673	147.262	9.560	74.000	96.000	23.514	10.2210/pdb1RH5/pdb
1RH5:C	1065.880	9.239	15.428	274.700	8.160	53.000	321.000	32.830	10.2210/pdb1RH5/pdb
1SFK:A	937.336	15.515	151.260	156.484	12.180	76.000	113.000	14.868	10.2210/pdb1SFK/pdb
1V54:G	139.531	40.780	4.004	218.573	10.510	85.000	-524.000	-9.882	10.2210/pdb1V54/pdb
1VF5:E	1757.790	1.885	22.637	111.177	9.700	32.000	2391.000	164.375	10.2210/pdb1VF5/pdb
1VF5:F	1518.330	4.726	117.790	45.367	4.490	35.000	903.000	150.286	10.2210/pdb1VF5/pdb
1VF5:H	1645.770	9.774	7.740	184.191	4.370	29.000	1352.000	181.379	10.2210/pdb1VF5/pdb
1VRZ:A	1319.110	2.887	0.000	25.830	5.520	21.000	1848.000	287.619	10.2210/pdb1VRZ/pdb
1VSY:3	922.815	32.962	41.132	173.199	6.010	76.000	-529.000	-3.421	10.2210/pdb1VSY/pdb
1W5C:E	1442.110	18.402	13.668	233.184	5.150	84.000	-11.000	15.238	10.2210/pdb1W5C/pdb
1W8X:M	0.000	13.871	5.306	155.465	6.970	83.000	-269.000	-4.819	10.2210/pdb1W8X/pdb

1XME:C	1735.320	6.505	39.772	114.433	8.250	34.000	1312.000	167.647	10.2210/pdb1XME/pdb
1XOW:B	0.000	2.832	0.000	14.333	9.750	11.000	-73.000	20.909	10.2210/pdb1XOW/pdb
1Y34:I	241.612	7.982	60.101	255.627	9.300	64.000	108.000	12.500	10.2210/pdb1Y34/pdb
1YCE:A	1745.240	27.022	320.390	152.987	4.940	89.000	913.000	29.775	10.2210/pdb1YCE/pdb
2AXT:F	1240.390	5.208	22.131	147.554	8.520	45.000	518.000	142.889	10.2210/pdb2AXT/pdb
2AXT:H	1608.240	18.766	494.942	152.267	8.190	66.000	705.000	97.424	10.2210/pdb2AXT/pdb
2AXT:I	1670.340	8.076	2.473	188.270	8.190	38.000	534.000	169.211	10.2210/pdb2AXT/pdb
2AXT:J	1529.980	21.025	2.547	149.974	5.750	40.000	1333.000	160.750	10.2210/pdb2AXT/pdb
2AXT:L	1545.380	31.507	1367.150	13.973	5.900	37.000	503.000	173.784	10.2210/pdb2AXT/pdb
2AXT:M	1527.270	3.230	19.385	121.367	4.530	36.000	892.000	178.611	10.2210/pdb2AXT/pdb
2AXT:T	1774.200	0.876	17.435	183.563	9.190	32.000	1062.000	200.938	10.2210/pdb2AXT/pdb
2AXT:Z	3263.390	17.928	57.572	209.960	5.710	62.000	1579.000	103.710	10.2210/pdb2AXT/pdb
2BKY:X	529.846	21.671	74.846	348.267	9.390	89.000	-247.000	-3.371	10.2210/pdb2BKY/pdb
2C0W:A	1398.590	16.786	92.949	157.038	6.280	50.000	428.000	23.600	10.2210/pdb2C0W/pdb
average	1122.611	19.518	104.156	155.541	7.882	57.948	408.603	68.965	
count	116	116	116	116	116	116	116	116	
sum	130222.913	2264.031	12082.121	18042.813	914.350	6722.000	47398.000	7999.889	
variance	575,986.984	292.796	47,896.480	6,479.969	5.979	541.180	523,607.841	6,226.000	
max	3352.22	93.078	1592.43	409.786	12.48	100	2391	364	
min	0	0.409	0	13.972	3.57	11	-1054	-17.812	
standard deviation	758.938	17.111	218.852	80.498	2.445	23.263	723.607	78.905	

Table STIII Input selection heuristics for ANFIS model.

Physicochemical Property	*in-vitro* aggregation	α-helix propensity	β-sheet propensity	isoelectric point	Length	hydrophobic mean	*in-vivo* aggregation
Dataset #	1	2	3	4	5	6	7