Ensemble AI Approach for Predicting Hemolysis Using Sequence and Concentration of Functional Peptides

You-Lin Zeng[1], Wen-Chih Cheng[2], Chung-Yen Lin[2]

1. CSIE, National Taiwan University, 2. Institute of Information Science, Academia Sinica, TAIWAN

Key Challenges

When designing antimicrobial peptides (AMPs), AMPs that indiscriminately kill all cells cannot be utilized as drugs. However, it is costly to validate if the sequence is hemolytic. Hence, this study aims to perform early screening on computed peptide sequences to rule out sequences that might induce hemolysis which lead to fail on drug discovery.

Dataset

The data came from DBAASP [1]. The labels depend on which threshold we are looking at. For example, if a sequence will kill 6% of erythrocytes, it will be labeled as hemolytic at threshold 5%, non-hemolytic otherwise. Each row in the dataset (2984 in total) includes amino acids sequence, concentration, hemolysis percentage.

Hemolysis	Pos+Neg	Train/Test	
		Pos	Neg
5%	2186	874/219	874/219
10%	2782	1112/279	1113/278
20%	2434	973/244	974/243
30%	2246	898/225	898/225
40%	2146	858/215	858/215

Brief Algorithm

Given the difficulty in accurately predicting the actual percentage of hemolysis using regression models, we have introduced a novel approach for estimation, as illustrated in Figure 1. We developed five distinct classification models employing different thresholds—specifically, 5%, 10%, 20%, 30%, and 40%. Each model can determine whether a given sequence, at a given concentration, is likely to induce hemolysis beyond a specified threshold. Hemolysis estimation is subsequently derived from the results generated by this algorithm.

To encode the sequences, we employed the PC6 method [3], where each amino acid is represented by six features. These features are then concatenated with the normalized concentration after PC6 transformation. As a result, each sequence (limited to a maximum of 50 amino acids) is encoded into a total of 301 features. We employed an ensemble of six different machine learning models: Support Vector Machine (SVM), Random Forest (RF), Multi-Layer Perceptron (MLP), k-Nearest Neighbors (KNN), XGBoost, and AdaBoost. The ensemble technique used is soft voting. We systematically explored the power set of these six models to determine the most effective combination.

The results indicate that, across all thresholds, the models consistently achieve an accuracy rate of approximately 0.8.

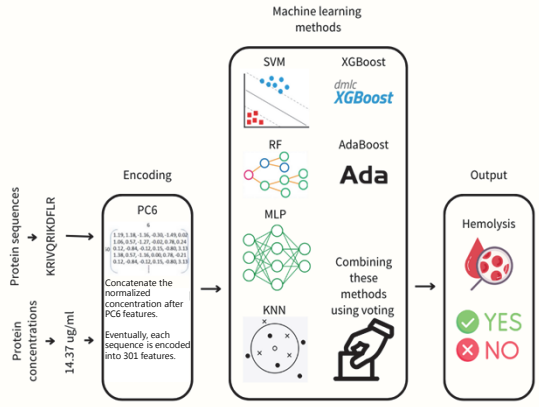

Figure 1. The ensemble mode used in this study.

Comparison with current study

The results were obtained through a 10-fold cross-validation process and then averaged. It's important to note that the reference data for HAPPENN was labelled based on the criteria outlined in their article [2], while our models were labelled using the actual hemolysis percentages from the original dataset. For performance comparisons with other thresholds, you can refer to our GitHub repository.

Threshold 20%	Acc.	Prec.	F1	MCC	recall	spec.
HAPPENN	0.67	0.55	0.67	0.44	0.54	**0.92**
This study	**0.77**	**0.77**	**0.76**	**0.53**	**0.77**	0.76

GitHub

Peptide-Hemolysis

[1] Pirtskhalava M;Amstrong AA;Grigolava M;Chubinidze M;Alimbarashvili E;Vishnepolsky B;Gabrielian A;Rosenthal A;Hurt DE;Tartakovsky M; (2021) DBAASP V3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of New Therapeutics, Nucleic acids research. Available at: https://pubmed.ncbi.nlm.nih.gov/33151284/

[2] Timmons, P.B. and Hewage, C.M. (2020) *Happenn is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks*, *Nature News*. Available at: https://www.nature.com/articles/s41598-020-67701-3.

[3] Lin TT;Yang LY;Lu IH;Cheng WC;Hsu ZR;Chen SH;Lin CY; (2021) *AI4AMP: An antimicrobial peptide predictor using physicochemical property-based encoding method and Deep Learning*, *mSystems*. Available at: https://pubmed.ncbi.nlm.nih.gov/34783578/.