The 6th International Conference on Biochemistry and Molecular Biology

Screening of non-cytotoxic membrane penetrating peptides using machine learning approach

Patcha Poungsombat[1], Wanapinun Nawae[2], Marasri Ruengjitchatchawalya[1,3,*]

[1]Bioinformatics and Systems Biology program, School of Bioresources and Technology/ School of Information Technology, King Mongkut's University of Technology Thonburi (KMUTT), Bang Khun Thian, Bangkok, Thailand
[2]HRH Princess Chulabhorn College of Medical Science, Bangkok, Thailand
[3]Biotechnology Program, School of Bioresources and Technology, KMUTT, Bang Khun Thian, Bangkok, Thailand
*E-mail: marasri.rue@kmutt.ac.th

Abstract

For peptide drug development, two machine learning methods were applied in this study to clarify critical features that are responsible for cell-penetrating activity and non-cytotoxicity of peptide. Lacking in dataset for peptides that exhibit both non-toxic and penetrating activities, therefore, two separate models were developed for the activity prediction. To construct cell penetrating peptides (CPPs) prediction model, 1,054 positive and 1,009 negative datasets were retrieved from CPPsite and CPPsite2.0. While 1,528 toxic peptides and 1,528 non-toxic peptides retrieving from UniprotKB were used to generate toxicity prediction model. The feature extraction was performed using Interpol package, summarized the feature values by Moreau- Broto autocorrelation (AC), and selected Correlation-base feature selection (CFS) subsetEval. As a result of feature selection, 15 and 17 features were acquired for modelling CPPs and toxicity activity, respectively. The models for CPPs and toxic peptide prediction constructed by Artificial Neural Network (ANN) yielded 94.7% and 95.5% ROC, respectively, demonstrating our applicable models to predict cell penetrating and cytotoxicity activities of unknown peptides.

Introduction

Peptide and protein drugs holding the high specificity and low toxicity properties lead them become a keystone of pharmaceutical manufacture [1]. The major barrier of drug development is their low efficiency delivery and low stability within a cell [2]. Cyclotides, a family of disulfide rich, head-to-tail cyclized peptides, become the key to improve the drug development in recent years regarding their cyclic cystine knot (CCK) motif that make them a group of ultra-stable peptides. Nawae et.al. (2014) reported that the binding of KalataB1 (KB1), a member of cyclotides family, to the membrane interfacial zone can cause membrane disruption [3].

Cell-penetrating peptides (CPPs), also known as protein transduction domains (PTDs) or membrane translocating sequences (MTSs) or Trojan peptides), are short peptides, characterized by highly cationic, rich in arginine and lysine amino acids in their sequences. They have an exceptional ability to cross cell membrane that made them to be importer for drug cargos [4]. There are two major uptake mechanisms of the CPPs including endocytosis and direct penetration which is depend on positive charge and hydrophobicity of peptides. Several studies demonstrated that positive residues in the peptide are significantly important for their uptake into the cell. For example, a study showed that poly-lysines can interact with negative

charged atoms in the membrane surface [5]. However, replacing of lysine with arginine cuased an increase of the uptake rate [6], which the poly-arginine of 7–15 residues is an optimal for the uptake requirement. Moreover, increasing of tryptophan residues in oligo-arginine sequences was reported to enhance the uptake efficiency [7]. Generally, hydrophobicity caused peptides sticking on the plasma membrane [8] and showed strong toxic properties via membrane disruption. Some report showed that amphipathic peptide disrupts the plasma membrane through the mechanism resemble to the pore forming of antimicrobial peptide [9]. Wherewith an internalization feature of CPPs without cytotoxic properties might be critical step for cyclotide modification in pharmaceutical application.

Due to the massive accumulation of biological data generated by high-throughput technology, computational approaches become useful tool for biological data analysis. Machine learning-based approaches have been used in proteins functional prediction. Support Vector Machine (SVM) applied for CPPs prediction with 95.94% accuracy based on dataset of 111 sequences was reported [10]; whereas other study found designing highly effective cell penetrating peptides with accuracy of 97.40% using the hybrid model that combines motif information and binary profile of the peptides [8]. However, cytotoxicity of CPPs have not been attentive concerned in any reported. In this study, we therefore aim to develop a prediction model for non-toxic CPPs with high accuracy and precision.

Methodology:

Data Set Compilation

A dataset for constructing CPPs prediction model were CPPs and non-CPPs sequences obtained from CPPsite[11] and CPPsite 2.0[12] with sequence length of not more than 50 amino acids. After redundant and noise were filtered, a total of 1,054 CPPs or positive set were randomly divided into two separate datasets containing 954 and 100 sequences to be used as training set and validation set, respectively. While 1,009 non-CPPs or negative set were randomly divided in to 909 and 100 sequences for being training set and validation set, respectively. For cytotoxicity prediction model, a dataset of toxic peptides was constructed from the amino acids sequence having less than 50 amino acids in length obtained from UniprotKB[13]. Likewise, the above, total 1,528 toxic peptides each of positive and negative sets were cleaned and randomly categorized into 1,375 and 153 sequences for being training set and validation set, respectively.

Feature extraction and selection

The total of 533 physicochemical properties were calculated for every peptide sequence using the Interpol package [14] then the features values were summarized by Moreau-Broto autocorrelation [15]. The feature selection was carried out using Waikato Environment for Knowledge Analysis (WEKA) software package [16] by using Correlation-base feature selection (CFS) subsetEval method [17].

Model training

Training and test set were applied to two machine learning algorithms: Artificial Neural Network (ANN) with hyper parameter: Hidden Layer a, Learning Rate 0.3, Momentum 0.2, Training Time 500, and Validation Threshold 20. Support Vector Machine (SVM) with hyper parameter: Epsilon 1.0E-12, Num Flod -1, Random Seed 1, and Tolerance Parameter 0.001 [Figure 1A]. The training sets was assessed using 10-fold cross validation and evaluated on validation/independent dataset [Figure 1A].

Model performance measurement

The factors for evaluating the performance of a machine learning program are calculated from confusion matrix. Four factors were applied in this work: precision, recall, F-Measure, and Receiver Operating Characteristic (ROC).

Finally, CPPs prediction and toxicity prediction models were connected and applied for predict non-toxic CPPs as shown in Figure 1B.

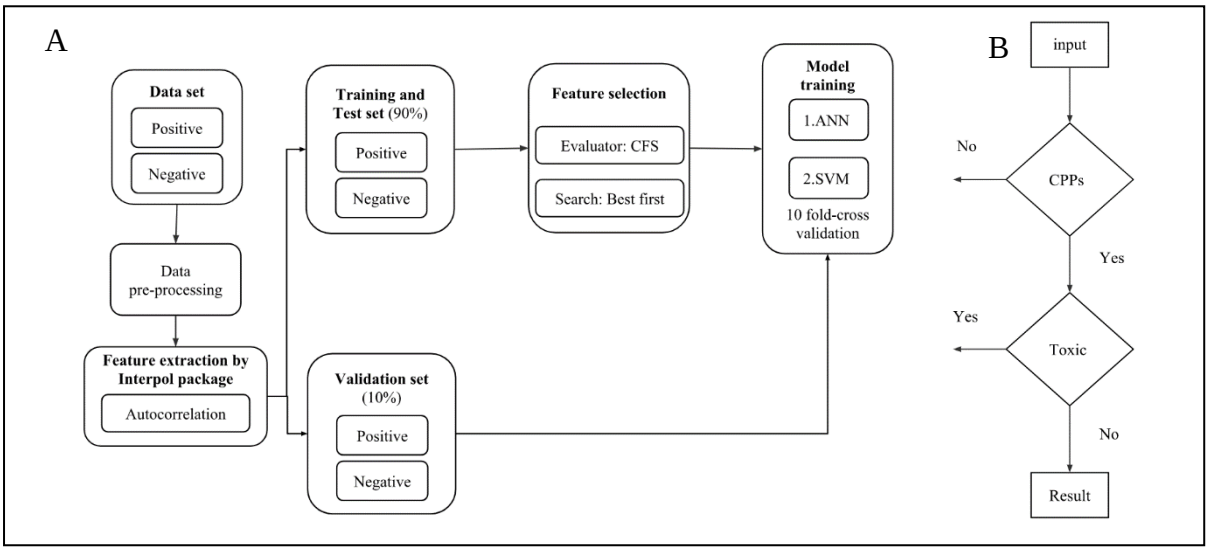

Figure 1. A. Overview of prediction model construction and B. Final model of non-Toxic CPPs prediction.

Results and Discussion

To construct an efficient model for predicting non-toxic and cell-penetrating ability of any peptide sequence, two models were developed separately to predict CPPs and toxicity. More than 500 physiochemical features were extracted for each amino acid sequence in the dataset. Out of these features, 15 features (Table 1) were found strongly correlate with membrane penetration activity of the peptides in dataset while their toxicity were observed to correlate with 17 features (Table 2).

Table 1: The feature set of CPPs prediction model extracted by using Interpol package and summary by using Moreau-Broto autocorrelation method.

Feature	Description
X131	Transfer_free_energy
X267	Weights_for_alpha-helix_at_the_window_position_of_3
X281	Weights_for_beta-sheet_at_the_window_position_of_4
X295	Weights_for_coil_at_the_window_position_of_5
X316	Transfer_free_energy_from_vap_to_chx
X317	Transfer_free_energy_from_chx_to_oct
X318	Transfer_free_energy_from_vap_to_oct
X320	Energy_transfer_from_out_to_in(95
X340	Information_measure_for_N-terminal_helix
X343	Information_measure_for_extended
X389	Hydration_potential
X392	Principal_property_value_z3
X397	Dependence_of_partition_coefficient_on_ionic_strength
X507	Hydrophobicity_index
X513	Weights_from_the_IFH_scale

Table 2: The feature set of Toxic peptide prediction model extracted by using Interpol package and summary by using Moreau-Broto autocorrelation method.

Feature	Description
X13	Retention_coefficient_in_HFBA
X22	Polarizability_parameter
X68	Consensus_normalized_hydrophobicity_scale
X74	Optical_rotation
X220	Optimized_transfer_energy_parameter
X221	Optimized_average_non-bonded_energy_per_atom
X259	Weights_for_alpha-helix_at_the_window_position_of_-5
X261	Weights_for_alpha-helix_at_the_window_position_of_-3
X263	Weights_for_alpha-helix_at_the_window_position_of_-1
X276	Weights_for_beta-sheet_at_the_window_position_of_-1
X284	Weights_for_coil_at_the_window_position_of_-6
X320	Energy_transfer_from_out_to_in(95
X341	Information_measure_for_middle_helix
X347	Information_measure_for_N-terminal_turn
X444	Side-chain_contribution_to_protein_stability
X521	ALTFT_index
X522	ALTLS_index

CPPs prediction model

Table 3 shows that the ANN model gained a better performance than SVM, especially, the Receiver Operating Characteristic (ROC) of 0.947 in comparison to that of 0.825. ANN is better than SVM in non-linear classification scenario because ANN provide multi-layer connection to deal with nonlinear problems, while SVM is based on the statistical learning theory for separating the data points into two different classes. Due to False negative (F_N) of ANN CPPs prediction model were very low and ROC was calculated from true positive rate and false positive rate, led to the significantly high ROC value than precision, recall and F.

Table 3. Performance of CPPs prediction model evaluated on validation dataset

Prediction algorithm	T_P	F_N	T_N	F_P	Precision	Recall	F	ROC
ANN	89	11	82	18	0.857	0.855	0.855	0.947
SVM	74	26	91	9	0.835	0.825	0.824	0.825

Toxic peptide prediction model

Similar to the CPPs prediction model, precision, recall, F-measure and ROC provided by ANN model were higher than those by SVM model (Table 4).

Table 4. Performance of toxic peptide prediction model evaluated on validation dataset

Prediction algorithm	T_P	F_N	T_N	F_P	Precision	Recall	F	ROC
ANN	138	15	140	13	0.909	0.908	0.908	0.955
SVM	139	14	136	17	0.899	0.899	0.899	0.899

Conclusion

The success of machine learning in model prediction relies on 3 things: the data set, feature set and machine learning algorithms. Suitable algorithm is necessary to achieve the good results. However, the algorithm classify the training set base on the features; the feature was extracted and selected from data set. This study provides classification model that apply set of critical physiochemical features (data not shown) for CPPs and toxic peptide prediction. Our results demonstrated that ANN achieved good accuracy for both CPPs and toxic peptide predictions. The models can be used to predict CPPs activity and toxicity of any unknown peptides with sequence length of less than 50 amino acids.

References

1. Katsila T, Siskos AP, Tamvakopoulos C. Peptide and protein drugs: the study of their metabolism and catabolism by mass spectrometry. Mass spectrometry reviews. 2012;31(1):110-33.
2. Bechara C, Sagan S. Cell-penetrating peptides: 20 years later, where do we stand?. FEBS letters. 2013;587(12):1693-702.
3. Nawae W, Hannongbua S, Ruengjitchatchawalya M. Defining the membrane disruption mechanism of kalata B1 via coarse-grained molecular dynamics simulations. Scientific reports. 2014;4:3933.
4. Lundberg P, Langel Ü. A brief introduction to cell-penetrating peptides. Journal of Molecular Recognition. 2003;16(5):227-33.
5. Deshayes S, Morris MC, Divita G, Heitz F. Interactions of amphipathic CPPs with model membranes. Biochimica et Biophysica Acta (BBA)-Biomembranes. 2006;1758(3):328-35.
6. Rydberg HA, Matson M, Åmand HL, Esbjörner EK, Nordén B. Effects of tryptophan content and backbone spacing on the uptake efficiency of cell-penetrating peptides. Biochemistry. 2012;51(27):5531-9.
7. Bechara C, Sagan S. Cell-penetrating peptides: 20 years later, where do we stand?. FEBS letters. 2013;587(12):1693-702.
8. Gautam A, Chaudhary K, Kumar R, Sharma A, Kapoor P, Tyagi A, Raghava GP. In silico approaches for designing highly effective cell penetrating peptides. Journal of translational medicine. 2013;11(1):74.

9. Hansen M, Kilk K, Langel Ü. Predicting cell-penetrating peptides. Advanced drug delivery reviews. 2008;60(4-5):572-9.
10. Ryser HJ, Hancock R. Histones and basic polyamino acids stimulate the uptake of albumin by tumor cells in culture. Science. 1965;150(3695):501-3.
11. Gautam A, Singh H, Tyagi A, Chaudhary K, Kumar R, Kapoor P, Raghava GP. CPPsite: a curated database of cell penetrating peptides. 2012.
12. Agrawal P, Bhalla S, Usmani SS, Singh S, Chaudhary K, Raghava GP, Gautam A. CPPsite 2.0: a repository of experimentally validated cell-penetrating peptides. Nucleic acids research. 2015;44(D1):D1098-103.
13. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. Uniprotkb/swiss-prot. InPlant bioinformatics. 2007;89-112.
14. Heider D, Hoffmann D. Interpol: An R package for preprocessing of protein sequences. BioData mining. 2011;4(1):16.
15. Li ZR, Lin HH, Han LY, Jiang L, Chen X, Chen YZ. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Research. 2006;34:W32-7.
16. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. 2009;11(1):10-8.
17. Hall MA. Correlation-based feature selection for machine learning.

Acknowledgements

Authors are thankful to Bioinformatics & Systems Biology Program; School of Bioresources and Technology and King Mongkut's University of Technology Thonburi; and National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Thailand, for providing scholarship and facilities for the research. We would also like to show our gratitude to Asst. Prof. Dr.Teeraphan Laomettachit and Dr.Sawannee Sutheeworapong for the greatly comment.