

Analysis and Prediction of Highly Effective Antiviral Peptides Based on Random Forests

Kuan Y. Chang*, Je-Ruei Yang

Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan

Abstract

The goal of this study was to examine and predict antiviral peptides. Although antiviral peptides hold great potential in antiviral drug discovery, little is done in antiviral peptide prediction. In this study, we demonstrate that a physicochemical model using random forests outperform in distinguishing antiviral peptides. On the experimental benchmark, our physicochemical model aided with aggregation and secondary structural features reaches 90% accuracy and 0.79 Matthew's correlation coefficient, which exceeds the previous models. The results suggest that aggregation could be an important feature for identifying antiviral peptides. In addition, our analysis reveals the characteristics of the antiviral peptides such as the importance of lysine and the abundance of α -helical secondary structures.

Citation: Chang KY, Yang J-R (2013) Analysis and Prediction of Highly Effective Antiviral Peptides Based on Random Forests. PLoS ONE 8(8): e70166. doi:10.1371/journal.pone.0070166

Editor: Mark Isalan, Center for Genomic Regulation, Spain

Received: March 16, 2013; **Accepted:** June 16, 2013; **Published:** August 5, 2013

Copyright: © 2013 Chang, Yang. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The work was supported by National Taiwan Ocean University. Je-Ruei Yang was partially supported by Center for Excellence for the Oceans at National Taiwan Ocean University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: kchang@ntou.edu.tw

Background

Antiviral peptides (AVPs) are an unconventional perspective for treating viral infections. Antiviral researches have undergone for more than half a century [1–3]. Although the traditional trial-and-error biochemical approach has led to the discovery of several antiviral nucleoside and non-nucleoside analogues such as brivudine against varicella-zoster virus [4], acyclovir against herpes simplex virus (HSV) [5], and azidothymidine (AZT) [6], stavudine [7–9] and efavirenz [10] against human immunodeficiency virus (HIV), the process is costly and time-consuming. Besides, severe toxicity is often a problem [11]. Instead, lower toxicity of antiviral peptides or proteins such as enfuvirtide against HIV virus [12] and DRACO [13], a potential panacea for all viruses, become an appealing alternative [14].

AVPs are known to fight against various viruses. All of the AVPs are derived from either synthetic combinatorial libraries or segments of natural proteins and their homologues. A list of highly effective antiviral peptides against HIV [15], HSV [16], hepatitis C virus [17], influenza virus [18–20], rabies virus [21], and west Nile virus [22] has been compiled into an online database AVPpred [23]. Recently, there is an dedicated AVP database HIPdb for HIV, comprehensively collecting the experimentally validated HIV inhibiting peptides [24].

Several mechanisms are available for AVPs to fight against viruses. Antiviral therapeutics agents are known to block the attachment of viruses, prevent from the fusion of viruses to host cells, interrupt the signaling process of viruses, or inhibit the replication of viruses in host cells which may involve DNA polymerase, reverse transcriptase, integrase, and protease [14]. Currently studies have shown that AVPs inhibited the fusion of viruses to the cells [25,26]; others have shown that AVPs interfered the replication of viruses [27–29].

Little is done in predicting and examining antiviral peptides. Broadly speaking, antiviral peptides should be a part of antimicrobial peptides, which fight against bacteria, fungi, parasites, and viruses. Several studies have been done in antimicrobial peptides [30–35], but a recent study by Thakur *et al.* demonstrated that antimicrobial peptide predictors are not suitable to assess AVPs [23]. In addition, this study was the first to explore four different approaches to predict effective AVPs: motif, sequence alignment, amino acid composition, and physicochemical features. Their results demonstrated that a support vector machine (SVM) approach using physicochemical features was a powerful method to identify AVPs. However, it is not clear whether key residues exist in AVPs and whether other methods can outperform SVM in predicting AVPs.

In this study, we demonstrate that our random forests (RF) model based on physicochemical properties works better for identifying AVPs. Physicochemical properties of peptides are a useful means to identify AVPs. A previous study demonstrated that predicting antimicrobial peptides (AMP) could depend on sequence-derived physicochemical properties and this study also suggested that aggregation could be important for classifying AMPs [33]; A recent study indeed showed that identifying AVPs using physicochemical properties of peptides worked [23]. Here we further investigated this finding.

Materials and Methods

Training, validation, and test data sets

The data sets were obtained from the study by Thakur *et al.* [23]. 1,056 peptides were validated experimentally, containing 604 highly effective AVPs and 452 non-effective peptides; another 604 peptides without experimental validation were non-effective from

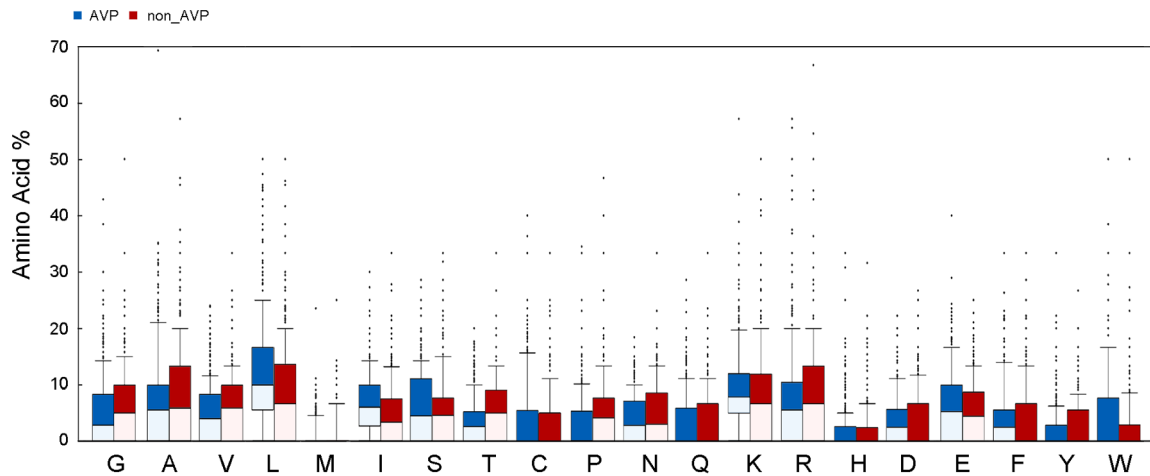


Figure 1. Statistical distribution of the amino acid composition of AVPs and non-AVPs. The blue and red bars represent the amino acid composition of 604 antiviral peptides and 452 non-antiviral peptides. doi:10.1371/journal.pone.0070166.g001

the study by Lata *et al.* [32]. Each of the peptides in the data sets was different from one another.

Two training sample sets and two independent test sets were established based on the data described above. Here we followed the same nomenclature used in the study by Thakur *et al.* [23]. 10-fold cross-validation was performed in our analysis, where the training and validation sets came from either of the two sample sets $T^{544P+407N}$ and $T^{544P+544N^*}$. $T^{544P+407N}$ consisted of 544 highly effective AVPs and 407 non-effective experimental peptides; $T^{544P+544N^*}$ contained the same 544 positive AVPs but different 544 non-experimental negative peptides. The independent test sets $V^{60P+45N}$ and $V^{60P+60N^*}$ were used for the benchmark. $V^{60P+45N}$ consisted of 60 highly effective AVPs and 45 non-effective peptides; $V^{60P+60N^*}$ contained 60 positive peptides and 60 non-experimental negative peptides.

Viral proteomes

The viral proteins were obtained from the viral genome database at the NCBI Entrez (January 2013), which consisted of 3251 viral genomes and viroids. All the 41316 viral proteins

expressed by these genomes and viroids were retrieved. Non-standard amino acids such as 'B', 'J', 'U' and 'X' in the viral proteins, which were less than 0.01% of the overall residues, were eliminated during the analysis. The viral proteomes were treated as non-effective antiviral peptides under a simple assumption that the viral proteins would not discourage themselves to develop.

RF classifier

RF is a classification method using an ensemble of unpruned decision trees with randomly selected features. The RF algorithm integrates random subspace method [36] into the concept of bootstrap aggregating or bagging [37] to generate an ensemble of decision trees. Independently each decision tree is best split by a small fraction of randomly selected features, trained on cases chosen by random sampling with replacement to all the available data. The aggregated decision trees then determine which class the predicted case belongs. The RF classifier known to avoid overfitting is a highly accurate method in many classification problems [38,39]. In this study, the randomForest R package version 4.6–7 was utilized [40], which was based on Breiman and

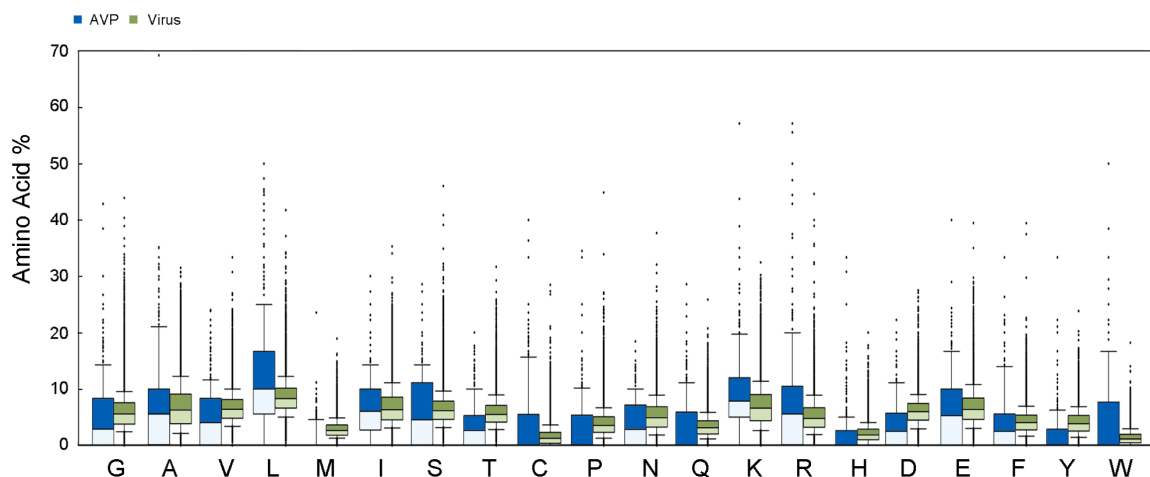


Figure 2. Statistical distribution of the amino acid composition of AVPs and viral proteomes. The blue and green bars represent the amino acid composition of 604 antiviral peptides and 41316 viral proteins. doi:10.1371/journal.pone.0070166.g002

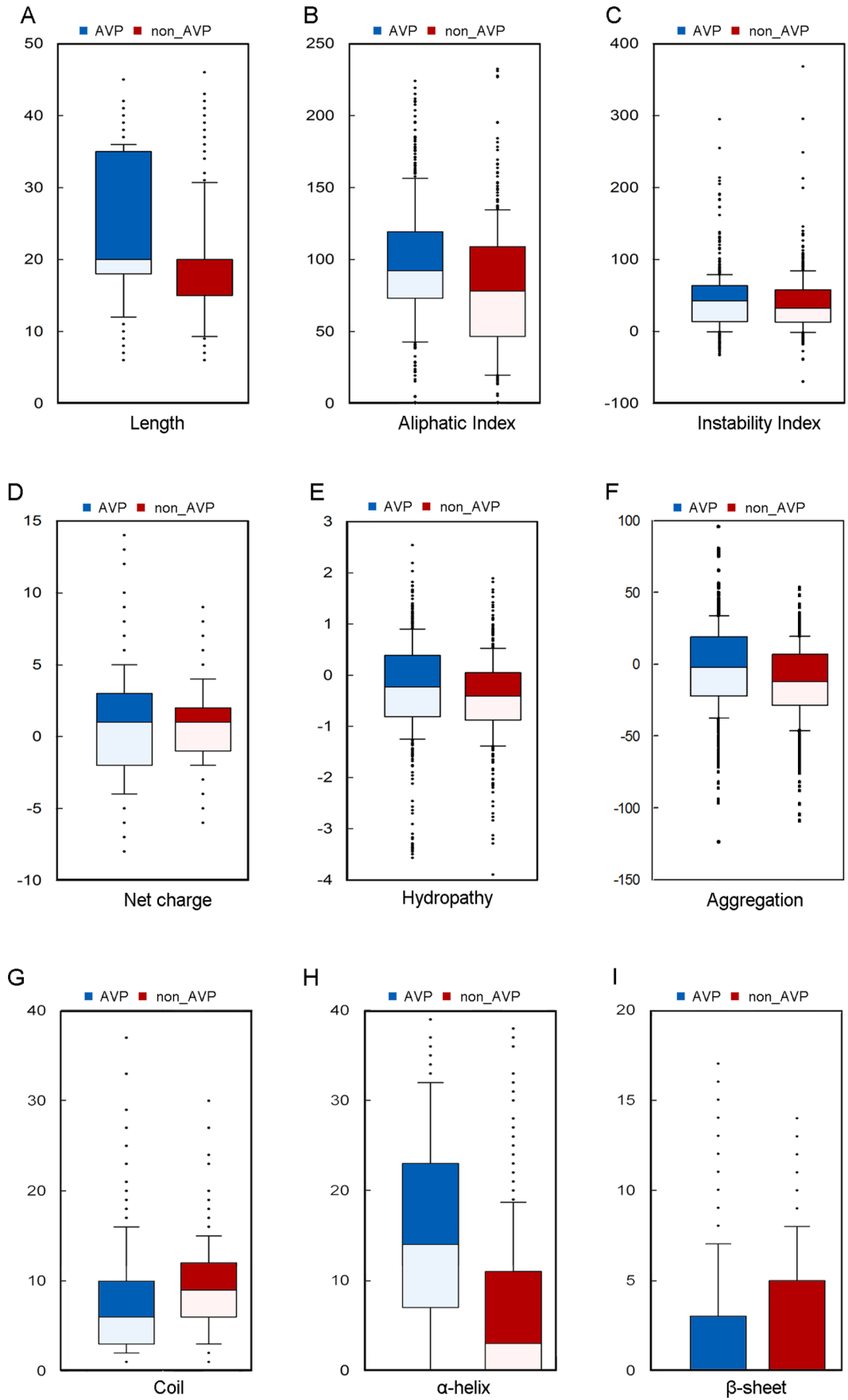


Figure 3. Statistical distribution of the physicochemical properties of AVPs and non-AVPs. 604 antiviral peptides and 452 non-antiviral peptides in the AVPpred study are compared, including (A) length, (B) aliphatic index, (C) instability index, (D) net charge, (E) hydrophathy, (F) aggregation, (G) random coil, (H) α -helix, and (I) β -sheet. doi:10.1371/journal.pone.0070166.g003

Cutler's algorithm [39]. Two parameters the number of *ntree* decision trees and the number of *mtry* selected features were set as follows: *ntree* = 100 and *mtry* = $\sqrt{\text{number_of_features}}$ as recommended [39]. One additional advantage of the RF model is that the model is possible to interpret the importance of the features using measures such as decrease mean accuracy or Gini importance.

Artificial Neural Network (ANN) classifier

In this study, ANN was trained by the backpropagation algorithm. Its learning rate and momentum rate were equal to 0.3 and 0.2 respectively. The number of hidden units was set to half of the number of features and the number of classes.

Linear Discriminant Analysis (LDA) classifier

The MASS R package version 7.3–26 was utilized to build the LDA models in this study. The LDA models seek the best linear combination of the features to separate AVPs from others.

Gini importance

Gini importance or the mean decrease of Gini index (MDGI) is a robust quantity to measure variable importance in the RF model [41]. Gini index is an impurity quantity defined as follows:

$$G = 1 - \sum_i P_i^2$$

where *i* contains all the classes and P_i is the fraction of class *i*. In our case there are two classes: AVPs and non-AVPs. The Gini index ranges from 0 to 1 and the index is closer to 0 if the sample is purer. The decrease of Gini index in a tree model refers to the difference between the Gini index of a parent node and the weighted Gini index of its descending nodes. Given a particular variable, the MDGI is the sum of the decrease of Gini index over the number of the decision trees in the RF. The larger the MDGI is, the more important the variable is.

Amino Acid Composition (AAC)

Amino acid composition is the ratio of each amino acid in a peptide. The ratio of an amino acid with type *T* in a peptide *X* is calculated as follows:

$$AAC(X_T) = \frac{N_T}{L}$$

where N_T is the number of the amino acid with type *T* and *L* is the length of peptide *X*.

Physicochemical properties

In the first attempt, 544 physicochemical indices in amino acid index database (AAindex version 9.1) [42] were examined. Each index contains 20 numerical values to represent the specific physicochemical properties of 20 amino acids. A simple linear addition model, the scalar product of each of the 544 indices and each of the amino acid composition of peptides, was used for the evaluation. Since the number of the parameters generated by these 544 indices was much larger than the number of the training cases, identifying crucial indices using feature selection was recom-

mended. We tested few feature selection methods such as minimum Redundancy Maximum Relevance to identify the crucial indices [43]. However, neither the top-ranked indices from our study nor those by Thakur *et al.* [23] provided a satisfactory predictive performance in the RF models. In fact, our results suggested that the RF models with more indices performed worst (data not shown), which followed similar trends as indicated in the previous finding [34].

In the next attempt, the following physicochemical properties were selected: length, net charges, instability index, aliphatic index, and hydrophaticity. Our basic physicochemical models were built by utilizing the five properties combined with amino acid composition. The advanced features such as secondary structures of peptides described below were considered.

The instability index, an estimate of peptide stability, is calculated as follows:

$$II = \frac{10}{L} \sum_{i=1}^{L-1} DIWV(X_i, X_{i+1})$$

where *L* is the length of peptide and *DIWV* from the study by Guruprasad *et al.* is an instability weight value of a dipeptide starting at position *i* [44]. Peptides with II values greater than 40 are considered to be unstable.

The aliphatic index, the relative volume of aliphatic residues in a peptide, is calculated as follows:

$$AI = X_{Ala} + a * X_{Val} + b * (X_{Leu} + X_{Ile})$$

where *a* and *b* are the constants, which represent the relative volume of valine and leucine or isoleucine to alanine. X_{Ala} , X_{Val} , X_{Leu} , and X_{Ile} are the fractions of alanine, valine, leucine and isoleucine multiplied by 100, respectively [45].

Grand average of hydrophaticity index (GRAVY) is used to represent the hydrophobicity value of a peptide, which calculates the sum of the hydrophathy values of all the amino acids divided by the sequence length. GRAVY was calculated using the hydrophathy values from Kyte and Doolittle [46]. Positive GRAVY values indicate hydrophobic; negative values mean hydrophilic. All these physicochemical values could be obtained directly from the ExpASY website (<http://www.expasy.org>) [47].

Aggregation

AGGRESCAN was utilized to estimate the aggregation tendencies of a peptide [48]. AGGRESCAN, which applied aggregation propensities of amino acids derived from the experimental data of β -amyloid peptides, is a good indicator of *in vivo* aggregation.

Secondary Structure

Protein secondary structure prediction (PSSpred version 2.0), which was a neural network classifier integrated into the famous I-TASSER server [49], was utilized to predict the secondary structure of a peptide [50]. Each amino acid on the peptide was classified into α -helix, β -sheet, or random coil. The number of each type of secondary structure was then recorded.

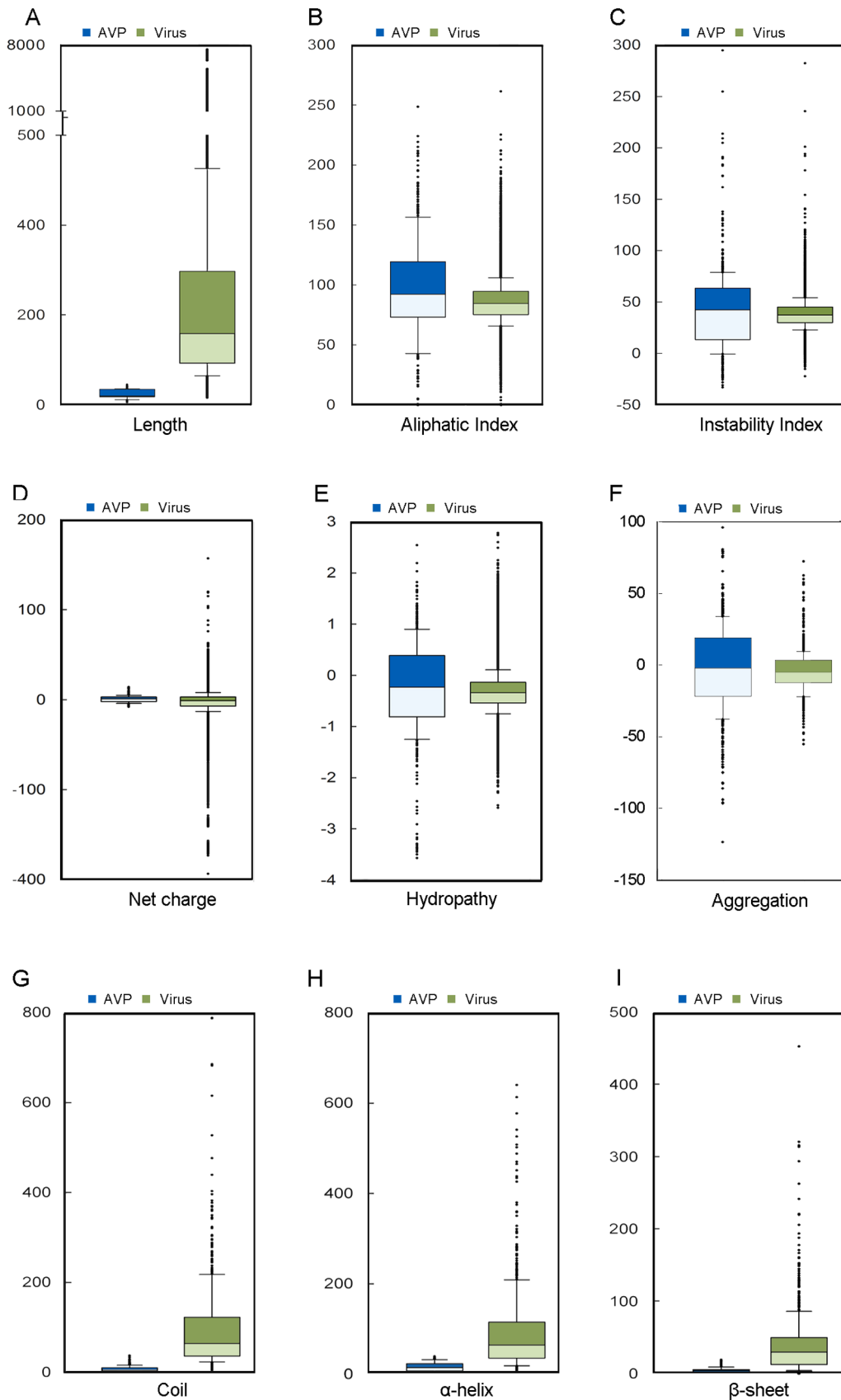


Figure 4. Statistical distribution of the physicochemical properties of AVPs and viral proteomes. 604 antiviral peptides and 41316 viral proteins are compared, including (A) length, (B) aliphatic index, (C) instability index, (D) net charge, (E) hydropathy, (F) aggregation, (G) random coil, (H) α -helix, and (I) β -sheet. Due to intense computing, the viral data at (F-I) were based on randomly selected 604 proteins instead of 41316 viral proteins. doi:10.1371/journal.pone.0070166.g004

Validation

The performance characteristics of the models such as sensitivity, specificity, accuracy, and MCC are defined as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{FP + TN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negative, respectively. Sensitivity, specificity, and accuracy are the percentage of the correct predictions on positive data, negative data, and all of the data, respectively. MCC is used to evaluate the performance of the binary classifier. It works even when the sizes of classification classes differ. The value of MCC is between -1 and 1 . The larger the MCC value, the better the classifier.

Results

Amino acid composition of the AVPs

The comparison of amino acid composition of AVPs was shown in Fig. 1 and Fig. 2. The viral proteomes were chosen to be an optional baseline. Compared to the non-effective experimental

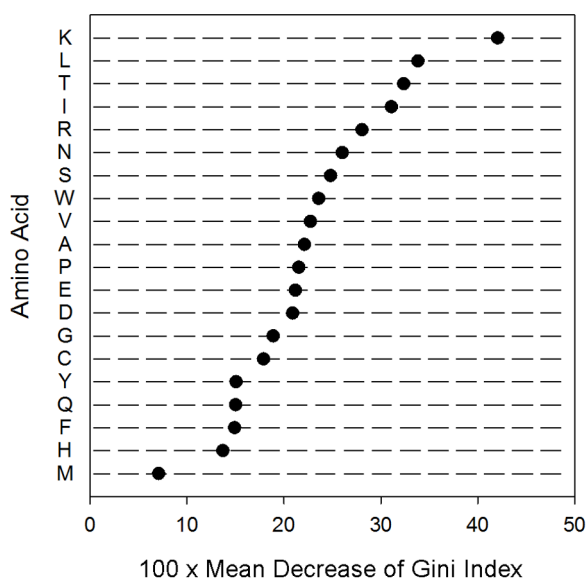


Figure 5. Feature importance of amino acid composition. The importance of each amino acid is measured using the mean decrease of Gini index (MDGI) of the 20 amino acids ranked by the RF model built for the AVPs and non-AVPs. The larger the MDGI value, the more important the residue.
doi:10.1371/journal.pone.0070166.g005

peptides and the viral proteomes, the AVPs consistently showed higher percentages of leucine and lysine, but lower percentages of threonine, proline, and valine. Leucine, medium-sized hydrophobic residue, was the most abundant residues in the AVPs. In addition, a clear preference was shown for a basic residue lysine, which was also known as the key characteristics for anti-microbial peptides.

Physicochemical properties, aggregation, and secondary structures of the AVPs

The statistical analysis of the physicochemical properties of the AVPs was done (Fig. 3 and Fig. 4). The following properties were examined, including length, aliphatic index, instability, net charge, hydrophathy, aggregation, and secondary structure. The lengths of the 604 AVPs ranged from 6 to 45 amino-acid residue, with an average length of 24.2 residues; the 452 non-effective AVPs had similar range with an average length of 18.2 residues; the viral proteins were longer with an average length over 200 residues. The AVPs had the strongest aliphatic tendency among them all (Fig. 3B and Fig. 4B). Indeed, the AVPs had the largest portions of aliphatic residues such as leucine, isoleucine, and valine (Fig. 1 and Fig. 2). Besides, only the AVPs had an average instability value over 40, which considered to be unstable; both the non-effective AVPs and viral proteins had an instability value less than 40 (Fig. 3C and Fig. 4C). In terms of residue charge, the AVPs had higher tendency to be charged positively than the non-effective peptides; the viral proteins preferred negative net charges (Fig. 3D and Fig. 4D). The AVPs had a slightly negative GRAVY score, but higher than both the non-effective peptides and the viral proteins (Fig. 3E and Fig. 4E). In the aggregation analysis, the AVP also had a slightly negative AGGRESCAN score on average, but higher than the non-effective peptides and the viral proteins

Table 1. Performance of the random forests models using 10-fold cross-validation.

Data	Model	Sensitivity	Specificity	Accuracy	MCC
T ^{544P+407N}	RFcompo	85.3	82.6	84.1	0.68
	RFcompo+agg	86.4	82.8	84.9	0.69
	RFcompo+structure	86.8	81.8	84.6	0.69
	RFcompo+structure+agg	86.6	83.0	85.1	0.70
	RFphysico	85.9	81.8	84.2	0.69
	RFphysico+agg	85.5	81.6	83.8	0.67
	RFphysico+structure	86.6	82.1	84.6	0.69
	RFphysico+structure+agg	86.6	81.6	84.4	0.68
	T ^{544P+544N*}	RFcompo [#]	89.5	92.6	91.1
RFcompo+agg [#]		89.5	92.6	91.1	0.82
RFcompo+structure [#]		89.0	94.1	91.5	0.83
RFcompo+structure+agg [#]		89.0	94.1	91.5	0.83
RFphysico [#]		88.1	93.8	90.0	0.82
RFphysico+agg [#]		88.1	93.8	90.0	0.82
RFphysico+structure [#]		89.0	93.9	91.5	0.83
RFphysico+structure+agg [#]		89.0	93.9	91.5	0.83

T^{544P+544N*} contained non-experimental peptides. The models trained by T^{544P+544N*} were marked by the number sign #.

doi:10.1371/journal.pone.0070166.t001

Table 2. Performance of the standard models by different learning methods on validation sets $V^{60P+45N}$ and $V^{60P+60N^*}$.

Data	Model	Sensitivity	Specificity	Accuracy	MCC
$V^{60P+45N}$	AVPcompo	83.3	88.9	85.7	0.72
	AVPphysico	88.3	82.2	85.7	0.71
	ANNcompo	76.7	75.6	76.2	0.52
	ANNphysico	85.0	80.0	82.9	0.65
	LDACompo	85.0	44.4	67.6	0.33
	LDAPhysico	88.3	53.3	73.3	0.45
	RFcompo	86.7	80.0	83.8	0.67
	RFphysico	93.3	77.8	86.7	0.73
$V^{60P+60N^*}$	AVPcompo [#]	83.3	98.3	90.8	0.83
	AVPphysico [#]	93.3	91.7	92.5	0.85
	ANNcompo [#]	81.7	93.3	87.5	0.76
	ANNphysico [#]	91.7	90.0	90.8	0.82
	LDACompo [#]	78.3	66.7	72.5	0.45
	LDAPhysico [#]	81.7	75.0	78.3	0.57
	RFcompo [#]	93.3	93.3	93.3	0.87
	RFphysico [#]	90.0	95.0	92.5	0.85

$V^{60P+60N^*}$ contained non-experimental peptides. The models trained by T544P+544N^{*} were marked by the number sign #; the models trained by T544P+407N had no marks. All the AVP models here were built using SVM [23]. doi:10.1371/journal.pone.0070166.t002

(Fig. 3F and Fig 4F).

Each residue of the peptides was classified into α -helix, β -sheet, or coil according to the secondary structure prediction. More than half of all the residues of the AVPs were α -helix (Fig. 3H), but the AVPs had lower portions of coil than the non-effective peptides and the viral proteins (Fig. 3G and Fig. 4G). β -sheet propensity was less clear, for few residues were classified into β -sheet (Fig. 3I and Fig 4I).

Feature importance of amino acid composition

The analysis of Gini importance on the RF model built with amino acid composition was performed. Lysine, a positively charged molecule, was the most important residue among the 20 amino acids for classifying AVPs. The result was consistent with those evaluated by different methods, which ranked the importance of amino acid similarly albeit slightly different (Table S1). Medium-size hydrophobic residues such as leucine and isoleucine also played a role in distinguishing AVPs from non-AVPs. Besides, the importance of threonine was also identified (Fig. 5).

In addition, the same Gini analysis of a RF model on an AMP database was performed. Although both lysine and arginine were abundant in the AMPs, the Gini analysis showed that neither lysine nor arginine was the most effective residue to tell apart AMPs from non-AMPs (Fig. S1). Our results indicated relatively less abundant residues in the AMPs such as methionine, aspartic acid, glutamic acid, and threonine are important to distinguish AMPs (Fig. S1). This differs from what we have seen in the AVPs, where lysine was the most effective residue.

Performance analysis on the training samples

All eight RF models were examined using 10-fold cross-validation trained by either T^{544P+407N} or T^{544P+544N^{*}} as summarized in Table 1. Those models trained by T^{544P+544N^{*}} were marked with the number sign #; those trained by T^{544P+407N}

Table 3. Performance of the random forests models on validation sets $V^{60P+45N}$ and $V^{60P+60N^*}$.

Data	Model	Sensitivity	Specificity	Accuracy	MCC
$V^{60P+45N}$	RFcompo	86.7	80.0	83.8	0.67
	RFcompo+agg	88.3	82.2	85.7	0.71
	RFcompo+structure	90.0	86.7	88.6	0.77
	RFcompo+structure+agg	91.7	86.7	89.5	0.79
	RFcompo [#]	93.3	48.9	74.3	0.48
	RFcompo+agg [#]	91.7	51.1	74.3	0.48
	RFcompo+structure [#]	93.3	42.2	71.4	0.43
	RFcompo+structure+agg [#]	93.3	48.9	74.3	0.48
	RFphysico	93.3	77.8	86.7	0.73
	RFphysico+agg	90.0	82.2	86.7	0.73
	RFphysico+structure	91.7	82.2	87.6	0.75
	RFphysico+structure+agg	91.7	82.2	87.6	0.75
	RFphysico [#]	90.0	53.3	74.3	0.48
	RFphysico+agg [#]	88.3	53.3	73.3	0.45
	RFphysico+structure [#]	90.0	40.0	68.6	0.35
	$V^{60P+60N^*}$	RFcompo	86.7	56.7	71.7
RFcompo+agg		88.3	56.7	72.5	0.47
RFcompo+structure		90.0	60.0	75.0	0.52
RFcompo+structure+agg		91.7	56.7	74.2	0.52
RFcompo [#]		93.3	93.3	93.3	0.87
RFcompo+agg [#]		91.7	95.0	93.3	0.87
RFcompo+structure [#]		93.3	88.3	90.8	0.82
RFcompo+structure+agg [#]		93.3	86.7	90.0	0.80
RFphysico		91.7	41.7	66.7	0.39
RFphysico+agg		90.0	45.0	67.5	0.39
RFphysico+structure		91.7	48.3	70.0	0.44
RFphysico+structure+agg		91.7	48.3	70.0	0.44
RFphysico [#]		90.0	95.0	92.5	0.85
RFphysico+agg [#]		88.3	91.7	90.0	0.80
RFphysico+structure [#]		90.0	88.3	89.2	0.78
RFphysico+structure+agg [#]		95.0	86.7	90.8	0.82

$V^{60P+60N^*}$ contained non-experimental peptides. The models trained by T^{544P+544N^{*}} were marked by the number sign #; the models trained by T^{544P+407N} had no marks.

doi:10.1371/journal.pone.0070166.t003

were without the sign. In order to easily compare AVPpred, our models named similarly after the models of AVPpred. The abbreviations compo, physico, structure, and agg represented amino acid composition, the amino acid composition plus five basic physicochemical properties, secondary structure, and aggregation respectively. For example, RFphysico was a RF model based on five basic physicochemical properties and amino acid

composition. RFcompo+structure+agg was based on amino acid composition, secondary structure, and aggregation.

In the 10-fold cross-validation, the performances of the eight RF models were close. In the $T^{544P+407N}$ training data, the best model RFcompo+structure+agg had an accuracy of 92% and 0.83 MCC. Both RFcompo+structure and RFphysico+structure reached a slightly higher accuracy of 85% and 0.69 MCC than RFcompo and RFphysico. In the $T^{544P+544N^*}$ training data, similar trends were found. Four models RFcompo+structure#, RFcompo+structure+agg#, RFphysico+structure#, and RFphysico+structure+agg# achieved a high accuracy of 92% and 0.83 MCC. As suggested before, different levels of the performances on the two training data sets might be due to $T^{544P+544N^*}$ containing non-experimental peptides [23]. Those non-experimental peptides were retrieved from non-secreted proteins, for the original study assumed antimicrobial peptides to be naturally secreted proteins [32]. The non-secreted proteins might be very different from the AVPs, thus distinguishing AVPs from those non-experimental peptides would be easy.

Performance evaluation on the independent test sets

Table 2 shows the performances of four learning methods with the standard models. The models using ANN, LDA, and RF were built on the same datasets. Two best SVM models of AVPpred, the only other tool available for predicting AVPs, were also included. All these models were evaluated on the two test sets $V^{60P+45N}$ and $V^{60P+60N^*}$. $V^{60P+45N}$ had only experimentally verified peptides, but $V^{60P+60N^*}$ contained non-experimental peptides. As seen in Table 2, the LDA models performed fairly, the ANN models gave a satisfactory prediction, and the SVM or RF models were very good in distinguishing AVPs. In general, the comparison results demonstrated that the RF models were superior to the other models in this problem.

On the $V^{60P+45N}$ test set, the RF models performed well (Table 2). The best RF models consistently outperformed the best models of AVPpred. For example, RFphysico surpassed AVPphysico, which was the performance indicator of AVPpred with an accuracy of 86% and 0.71 MCC, while both models utilized only 25 physicochemical properties.

On the $V^{60P+60N^*}$ test set containing non-experimental peptides, similar results were seen. For this test set, the two best models of AVPpred trained by $T^{544P+544N^*}$, AVPcompo# and AVPphysico#, were included. Our best model RFcompo# achieved the maximal accuracy of 93% and the highest 0.87 MCC. RFcompo# outperformed AVPcompo# while RFphysico and AVPphysico were comparable.

Whether additional features such as aggregation and secondary structure could improve the prediction was examined. Table 3 shows a performance comparison among the eight RF models. Generally speaking, adding aggregation to the RF models tended to improve the performance slightly. For example, on both test sets, RFcompo+agg surpassed RFcompo while RFcompo+agg# and RFcompo# were comparable. Adding secondary structure could also improve the RF models. For example, RFcompo+structure outperformed RFcompo. In addition, on the $V^{60P+45N}$ test set, our best model RFcompo+structure+agg achieved the maximal accuracy of 90% and the highest 0.79 MCC. On the $V^{60P+60N^*}$ test set, our best model RFcompo# and RFcompo+agg# achieved the maximal accuracy of 93% and the highest 0.87 MCC. The overall results demonstrated that these features could be useful for identifying the AVPs.

Discussion

This was the first study to apply RF into AVP prediction. Our RF models were based on size, amino acid composition, net charge, aliphaticity, instability, hydrophobicity, and secondary structure. The results demonstrated that predicting AVPs by the RF models through the basic physicochemical properties worked. On the independent test data provided by AVPpred, our evaluation indicated that RF worked better than SVM in distinguishing AVPs using these physicochemical properties. Our result supports that RF, a robust classifier, excels in many problems [38].

In order to reach optimal performances, several training designs for the RF models were explored. One option was to remove highly similar sequences from the training cases. The performance dropped as the RF models were trained by the non-redundant cases, which were generated by removing highly similar cases over 80% identities, leaving out more than one-third of the entire training cases. Additional analyses on the model attributes were also done such as increasing the number of attributes by AAindex and replacing the secondary-structure counts with the secondary-structure percentages. However, none enhanced performance. This suggests that improving AVP prediction is not a trivial task.

To compare AVPpred fully, the effects of different training sets on the RF models were examined. One set had the experimental data; the other contained the negatives without experimental verification. This led to different levels of specificity. Since true negatives in biological context are often limited and difficult to obtain, hypothetical negatives or negatives without experimental verification are the substitutes. However, ideally all true negatives for building the models should be verified experimentally to avoid performance errors. More attention, therefore, should be paid to the models with the experimental data. Our RF models outperformed the previous ones in either training set, but more obvious in the experimental training data.

Feature importance of amino acid composition of the AVPs was analyzed using the experimental data. Lysine, also known as a key residue in antimicrobial peptides, was the most important residue in distinguish AVPs. The residues with great importance need not to be the abundant residues. Our analyses showed that the AVPs were abundant in leucine, lysine, alanine, and glutamic acid (Fig. 1 and Fig. 2), which supports the previous finding by HIPdb [24]. Several other interesting properties were found in our analysis, but we only emphasized lysine, for this finding was supported by various feature ranking methods. However, why the AVPs were biased toward lysine is not clear. One possible explanation is that the positively charged AVPs might interact with enveloped viruses like HIV [30], inhibiting the entry of the viruses into the cells.

The aggregation propensities of the AVPs were revealed. Our results suggested that the AVPs had higher tendencies to be aggregated *in vivo* than the non-effective peptides and the viral proteins. Our RF models also indicated that aggregation could be an important feature for classifying the AVPs. It is not clear how aggregation affects the AVPs to restrain the viruses. However, aggregation has been suggested to be an important feature of AMPs, for accumulating and clumping peptides together could affect the peptide availability [33].

In addition, the abundance of α -helix secondary structures in the AVPs was discovered in this study. What role the secondary-structure preference plays in fighting viral infection is not understood. It has been suggested that the α -helical structures are a common element for protein-protein interactions [51] and the α -helical structures of antimicrobial peptides are linked to the permeability to the membranes to abrupt the cell membranes of

bacteria [35]. Whether the α -helical structures of the AVPs, a subset of antimicrobial peptides, interact with the enzymes in the virus replication or have similar impact on host cell membranes still need to be investigated. However, given secondary structure information, the performance of the RF models could further improve. On the experimental benchmark, our model trained by amino acid composition, aggregation, and secondary structure achieved the optimal performances. The importance of secondary structure was also supported by the previous study—several top physicochemical indices were related to secondary structure [23].

Supporting Information

Figure S1 Feature importance of amino acid composition of AMPs. The importance of each amino acid is measured using the mean decrease of Gini index (MDGI) of the 20 amino acids

ranked by the RF model built for the AMPs and non-AMPs. The larger the MDGI value, the more important the residue. (TIF)

Table S1 Feature importance of amino acid composition measured by different methods. (DOCX)

Text S1 Supplementary Information for Figure S1. (DOCX)

Author Contributions

Conceived and designed the experiments: KYC JRY. Performed the experiments: KYC JRY. Analyzed the data: KYC JRY. Contributed reagents/materials/analysis tools: KYC. Wrote the paper: KYC JRY.

References

- Wainman MA (2009) Perspectives on antiviral drug development. *Antiviral Res* 81: 1–5.
- De Clercq E (2011) A 40-year journey in search of selective antiviral chemotherapy. *Annu Rev Pharmacol Toxicol* 51: 1–24.
- Field HJ, De Clercq E (2004) Antiviral drugs – a short history of their discovery and development. *Microbiol Today*: 58–61.
- de Clercq E, Degreef H, Wildiers J, de Jonge G, Drochmans A, et al. (1980) Oral (E)-5-(2-bromovinyl)-2'-deoxyuridine in severe herpes zoster. *Br Med J* 281: 1178.
- Schaeffer HJ, Beauchamp L, de Miranda P, Elion GB, Bauer DJ, et al. (1978) 9-(2-hydroxyethoxymethyl) guanine activity against viruses of the herpes group. *Nature* 272: 583–585.
- Mitsuya H, Weinhold KJ, Furman PA, St Clair MH, Lehrman SN, et al. (1985) 3'-Azido-3'-deoxythymidine (BW A509U): an antiviral agent that inhibits the infectivity and cytopathic effect of human T-lymphotropic virus type III/lymphadenopathy-associated virus in vitro. *Proc Natl Acad Sci U S A* 82: 7096–7100.
- Baba M, Pauwels R, Herdewijn P, De Clercq E, Desmyter J, et al. (1987) Both 2',3'-dideoxythymidine and its 2',3'-unsaturated derivative (2',3'-dideoxythymidinene) are potent and selective inhibitors of human immunodeficiency virus replication in vitro. *Biochem Biophys Res Commun* 142: 128–134.
- Hamamoto Y, Nakashima H, Matsui T, Matsuda A, Ueda T, et al. (1987) Inhibitory effect of 2',3'-dideoxy-2',3'-dideoxynucleosides on infectivity, cytopathic effects, and replication of human immunodeficiency virus. *Antimicrob Agents Chemother* 31: 907–910.
- Lin TS, Schinazi RF, Prusoff WH (1987) Potent and selective in vitro activity of 3'-deoxythymidin-2'-ene (3'-deoxy-2',3'-dideoxythymidine) against human immunodeficiency virus. *Biochem Pharmacol* 36: 2713–2718.
- Young SD, Britcher SF, Tran LO, Payne LS, Lumma WC, et al. (1995) L-743,726 (DMP-266): a novel, highly potent nonnucleoside inhibitor of the human immunodeficiency virus type 1 reverse transcriptase. *Antimicrob Agents Chemother* 39: 2602–2605.
- Lewis W, Dalakas MC (1995) Mitochondrial toxicity of antiviral drugs. *Nat Med* 1: 417–422.
- Wild C, Greenwell T, Matthews T (1993) A synthetic peptide from HIV-1 gp41 is a potent inhibitor of virus-mediated cell-cell fusion. *AIDS Res Hum Retroviruses* 9: 1051–1053.
- Rider TH, Zook CE, Boettcher TL, Wick ST, Pancoast JS, et al. (2011) Broad-spectrum antiviral therapeutics. *PLoS One* 6: e22572.
- Castel G, Chtoumi M, Heyd B, Tordo N (2011) Phage display of combinatorial peptide libraries: application to antiviral research. *Molecules* 16: 3499–3518.
- Dwyer JJ, Wilson KL, Davison DK, Freil SA, Seedorff JE, et al. (2007) Design of helical, oligomeric HIV-1 fusion inhibitor peptides with potent activity against enfuvirtide-resistant virus. *Proc Natl Acad Sci U S A* 104: 12772–12777.
- Akkarawongsa R, Pocar NE, Case G, Kolb AW, Brandt CR (2009) Multiple peptides homologous to herpes simplex virus type 1 glycoprotein B inhibit viral infection. *Antimicrob Agents Chemother* 53: 987–996.
- Portal-Nunez S, Gonzalez-Navarro CJ, Garcia-Delgado M, Vizmanos JL, Lasarte JJ, et al. (2003) Peptide inhibitors of hepatitis C virus NS3 protease. *Antivir Chem Chemother* 14: 225–233.
- Drummond JE, Shaw EE, Antonello JM, Green T, Page GJ, et al. (2008) Design and optimization of a multiplex anti-influenza peptide immunoassay. *J Immunol Methods* 334: 11–20.
- Jones JC, Settles EW, Brandt CR, Schultz-Cherry S (2011) Identification of the minimal active sequence of an anti-influenza virus peptide. *Antimicrob Agents Chemother* 55: 1810–1813.
- Wunderlich K, Juozapaitis M, Ranadheera C, Kessler U, Martin A, et al. (2011) Identification of high-affinity PB1-derived peptides with enhanced affinity to the PA protein of influenza A virus polymerase. *Antimicrob Agents Chemother* 55: 696–702.
- Real E, Rain JC, Battaglia V, Jallet C, Perrin P, et al. (2004) Antiviral drug discovery strategy using combinatorial libraries of structurally constrained peptides. *J Virol* 78: 7410–7417.
- Bai F, Town T, Pradhan D, Cox J, Ashish, et al. (2007) Antiviral peptides targeting the west nile virus envelope protein. *J Virol* 81: 2047–2055.
- Thakur N, Qureshi A, Kumar M (2012) AVPpred: collection and prediction of highly effective antiviral peptides. *Nucleic Acids Res* 40: W199–204.
- Qureshi A, Thakur N, Kumar M (2013) HIPdb: A Database of Experimentally Validated HIV Inhibiting Peptides. *PLoS One* 8: e54908.
- Rausch DM, Hwang KM, Padgett M, Voltz AH, Rivas A, et al. (1990) Peptides derived from the CDR3-homologous domain of the CD4 molecule are specific inhibitors of HIV-1 and SIV infection, virus-induced cell fusion, and postinfection viral transmission in vitro. Implications for the design of small peptide anti-HIV therapeutic agents. *Ann N Y Acad Sci* 616: 125–148.
- Pang W, Tam SC, Zheng YT (2009) Current peptide HIV type-1 fusion inhibitors. *Antivir Chem Chemother* 20: 1–18.
- Louis JM, Dyda F, Nashed NT, Kimmel AR, Davies DR (1998) Hydrophilic peptides derived from the transframe region of Gag-Pol inhibit the HIV-1 protease. *Biochemistry* 37: 2105–2110.
- Oz Gleenberg I, Avidan O, Goldgur Y, Herschhorn A, Hizi A (2005) Peptides derived from the reverse transcriptase of human immunodeficiency virus type 1 as novel inhibitors of the viral integrase. *J Biol Chem* 280: 21987–21996.
- Gleenberg IO, Herschhorn A, Hizi A (2007) Inhibition of the activities of reverse transcriptase and integrase of human immunodeficiency virus type-1 by peptides derived from the homologous viral protein R (Vpr). *J Mol Biol* 369: 1230–1243.
- Hancock RE, Diamond G (2000) The role of cationic antimicrobial peptides in innate host defenses. *Trends Microbiol* 8: 402–410.
- Thomas S, Kamik S, Barai RS, Jayaraman VK, Idicula-Thomas S (2010) CAMP: a useful resource for research on antimicrobial peptides. *Nucleic Acids Res* 38: D774–780.
- Lata S, Mishra NK, Raghava GP (2010) AntiBP2: improved version of antibacterial peptide prediction. *BMC Bioinformatics* 11 Suppl 1: S19.
- Torrent M, Andreu D, Nogues VM, Boix E (2011) Connecting peptide physicochemical and antimicrobial properties by a rational prediction model. *PLoS One* 6: e16968.
- Wang P, Hu L, Liu G, Jiang N, Chen X, et al. (2011) Prediction of antimicrobial peptides based on sequence alignment and feature selection methods. *PLoS One* 6: e18476.
- Fjell CD, Hiss JA, Hancock RE, Schneider G (2012) Designing antimicrobial peptides: form follows function. *Nat Rev Drug Discov* 11: 37–51.
- Ho TK (1998) The Random Subspace Method for Constructing Decision Forests. *IEEE Trans on Pattern Analysis and Machine Intelligence* 20: 832–844.
- Breiman L (1996) Bagging predictors. *Machine Learning* 24: 123–140.
- Caruana R, Karampatziakis N, Yessenaflina A (2008) An empirical evaluation of supervised learning in high dimensions.
- Breiman L (2001) Random Forests. *Machine Learning* 45: 5–32.
- Liaw A, Wiener M (2002) Classification and Regression by randomForest. *R News* 2: 18–22.
- Calle ML, Urrea V (2011) Letter to the editor: Stability of Random Forest importance measures. *Brief Bioinform* 12: 86–89.
- Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. *Nucleic Acids Res* 28: 374.
- Peng H, Long F, Ding C (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27: 1226–1238.
- Guruprasad K, Reddy BV, Pandit MW (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng* 4: 155–161.
- Ikai A (1980) Thermostability and aliphatic index of globular proteins. *J Biochem* 88: 1895–1898.

46. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157: 105–132.
47. Gasteiger EHC, Gattiker A, Duvaud S, Wilkins MR, Appel RD, et al. (2005) Protein Identification and Analysis Tools on the ExPASy Server: Humana Press. 571–607 p.
48. Conchillo-Sole O, de Groot NS, Aviles FX, Vendrell J, Daura X, et al. (2007) AGGRESCAN: a server for the prediction and evaluation of “hot spots” of aggregation in polypeptides. *BMC Bioinformatics* 8: 65.
49. Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* 9: 40.
50. PSSpred website. Available: <http://zhanglab.ccmb.med.umich.edu/PSSpred>. Accessed 2013 Jun 3.
51. Azzarito V, Long K, Murphy NS, Wilson AJ (2013) Inhibition of alpha-helix-mediated protein-protein interactions using designed molecules. *Nat Chem* 5: 161–173.